Acta Numerica (2025), pp. 579–804
doi:10.1017/S0962492924000084

Distributionally robust optimization

Daniel Kuhn
Risk Analytics and Optimization Chair,
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
E-mail: daniel.kuhn@epfl.ch

Soroosh Shafiee
School of Operations Research and Information Engineering,
Cornell University, Ithaca, NY, USA
E-mail: shafiee@cornell.edu

Wolfram Wiesemann
Imperial College Business School,
Imperial College London, London SW7 2AZ, UK
E-mail: ww@imperial.ac.uk

Distributionally robust optimization (DRO) studies decision problems under uncertainty where the probability distribution governing the uncertain problem parameters is itself uncertain. A key component of any DRO model is its ambiguity set, that is, a family of probability distributions consistent with any available structural or statistical information. DRO seeks decisions that perform best under the worst distribution in the ambiguity set. This worst case criterion is supported by findings in psychology and neuroscience, which indicate that many decision-makers have a low tolerance for distributional ambiguity. DRO is rooted in statistics, operations research and control theory, and recent research has uncovered its deep connections to regularization techniques and adversarial training in machine learning. This survey presents the key findings of the field in a unified and self-contained manner.

2020 Mathematics Subject Classification: Primary 90-02, 90C17
Secondary 90C15, 90C47

CONTENTS

1. Introduction

Traditionally, mathematical optimization studies problems of the form

$$\inf_{x \in \mathcal{X}} \ \ell(x),$$

where a decision x is sought from the set $\mathcal{X} \subseteq \mathbb{R}^n$ of feasible solutions that minimizes a loss function $\ell\colon \mathbb{R}^n \to \overline{\mathbb{R}}$. With its early roots in the development of calculus by Isaac Newton, Gottfried Wilhelm Leibniz, Pierre de Fermat and others in the late seventeenth century, mathematical optimization has a rich history that involves contributions from numerous mathematicians, economists, engineers and scientists. The birth of modern mathematical optimization is commonly credited to George Dantzig, whose simplex algorithm developed in 1947 solves linear optimization problems where ℓ is affine and $\mathcal{X}$ is a polyhedron (Dantzig 1956). Subsequent milestones include the development of the rich theory of convex analysis (Rockafellar 1970) as well as the discovery of polynomial-time solution methods for linear (Khachiyan 1979, Karmarkar 1984) and broad classes of nonlinear convex optimization problems (Nesterov and Nemirovskii 1994).

Classical optimization problems are *deterministic*, that is, all problem data are assumed to be known with certainty. However, most decision problems encountered in practice depend on parameters that are corrupted by measurement errors or that are revealed only *after* a decision must be determined and committed. A naïve approach to modelling uncertainty-affected decision problems as deterministic optimization problems would be to replace all uncertain parameters with their expected values or with appropriate point predictions. However, it has long been known and well documented that decision-makers who replace an uncertain parameter of an optimization problem with its mean value fall victim to the 'flaw of averages' (Savage, Scholtes and Zweidler 2006, Savage 2012). In order to account for uncertainty realizations that deviate from the mean value, Beale (1955) and

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Dantzig (1955) independently introduced *stochastic programs* of the form

$$\inf_{x\in\mathcal{X}} \ \mathbb{E}_{\mathbb{P}}[\ell(x,Z)], \tag{1.1}$$

which explicitly model the uncertain problem parameters Z as a random vector that is governed by a probability distribution $\mathbb{P}$, and where a decision is sought that performs best in expectation (or, subsequently, according to some risk measure). Since then, stochastic programming has grown into a mature field (Birge and Louveaux 2011, Shapiro, Dentcheva and Ruszczyński 2009), and it provides the theoretical underpinnings of the empirical risk minimization principle in machine learning (Bishop 2006, Hastie, Tibshirani and Friedman 2009).

Despite their success in theory and practice, stochastic programs suffer from at least two shortcomings. Firstly, the assumption that the probability distribution $\mathbb{P}$ is known precisely is unrealistic in many practical settings, and stochastic programs can be sensitive to mis-specifications of this distribution. This effect has been described by different communities as the optimizer's curse (Smith and Winkler 2006), the error-maximization effect of optimization (Michaud 1989, DeMiguel and Nogales 2009), the optimization bias (Shapiro 2003) or overfitting (Bishop 2006, Hastie *et al.* 2009). Secondly, evaluating the expected loss of a fixed decision requires computing a multi-dimensional integral, which is provably hard already for embarrassingly simple loss functions and distributions. Hence stochastic programs suffer from the curse of dimensionality, that is, their computational complexity generically displays an exponential dependence on the dimension of the random vector Z. To alleviate both shortcomings, Soyster (1973) proposed modelling uncertainty-affected decision problems as *robust optimization problems* of the form

$$\inf_{x\in\mathcal{X}} \ \sup_{z\in\mathcal{Z}} \ \ell(x,z).$$

Robust optimization replaces the probabilistic description of the uncertain problem parameters with a set-based description and seeks for decisions that perform best in view of the worst anticipated parameter realization z from within an *uncertainty set* $\mathcal{Z}$. After an extended period of neglect, the ideas of Soyster (1973) have been revisited and substantially extended in the late 1990s onwards by Kouvelis and Yu (1997), El Ghaoui, Oustry and Lebret (1998), El Ghaoui and Lebret (1998*a*,*b*), Ben-Tal and Nemirovski (1998, 1999*a*,*b*), Bertsimas and Sim (2004) and others. For reviews of the robust optimization literature, we refer to Ben-Tal, El Ghaoui and Nemirovski (2009), Rustem and Howe (2009) and Bertsimas and den Hertog (2022). We point out that similar ideas have been developed independently in the areas of robust stability (Horn and Johnson 1985, Doyle, Glover, Khargonekar and Francis 1989, Green and Limebeer 1995), which investigates whether a system remains stable in the face of parameter variations, and robust control (Zames 1966, Khalil 1996, Zhou, Doyle and Glover 1996), which designs systems that maintain a desirable performance in the presence of parameter variations. For textbook introductions to robust stability and control, we refer to Zhou and Doyle (1999) and

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Dullerud and Paganini (2001). Hansen and Sargent (2008) adapt robust control techniques to economic problems affected by model uncertainty, where they design policies that perform well across a range of possible model mis-specifications.

While robust optimization reduces the informational and computational burden that plagues stochastic programs, its equal treatment of all parameter realizations within the uncertainty set and its exclusive focus on worst-case scenarios can make it overly conservative for practical applications. These concerns prompted researchers to study *distributionally robust optimization problems* of the form

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(x,Z)], \tag{1.2}$$

which model the uncertain problem parameters Z as a random vector that is governed by some distribution $\mathbb{P}$ from within an *ambiguity set* $\mathcal{P}$, and where a decision is sought that performs best in view of its expected value under the worst distribution $\mathbb{P}\in\mathcal{P}$. Distributionally robust optimization (DRO) thus blends the distributional perspective of stochastic programming with the worst-case focus of robust optimization. Herbert E. Scarf is commonly credited with pioneering this approach in his study on newsvendor problems where the uncertain demand distribution is only characterized through its mean and variance (Scarf 1958). Subsequently, Dupačová (1966, 1987, 1994) and Shapiro and Kleywegt (2002) have studied DRO problems whose ambiguity sets specify the support, some lower-order moments, independence patterns or other structural properties of the unknown probability distribution. Ermoliev, Gaivoronski and Nedeva (1985) and Gaivoronski (1991) have developed early solution approaches for DRO problems over moment ambiguity sets. The advent of modern DRO is often attributed to the works of Bertsimas and Popescu (2002, 2005), who derive probability inequalities under partial distributional information and apply their techniques to option pricing problems, of El Ghaoui, Oks and Oustry (2003) and Calafiore and El Ghaoui (2006), who study DRO problems where a quantile of the objective function should be minimized, or a set of uncertainty-affected constraints should be satisfied with high probability, across all probability distributions with known moment bounds, and of Delage and Ye (2010), who study similar DRO problems with a worst-case expected value objective.

Early research on DRO has primarily focused on moment ambiguity sets, which contain all distributions on a prescribed support set $\mathcal{Z}$ that satisfy finitely many moment constraints. In contrast to stochastic programs, DRO problems with moment ambiguity sets sometimes exhibit favourable scaling with respect to the dimension of the random vector Z. However, strikingly different distributions can share identical moments. As a consequence, moment ambiguity sets always include a wide range of distributions, including some implausible ones that can safely be ruled out when ample historical data is available. This prompted Ben-Tal *et al.* (2013) and Wang, Glynn and Ye (2016) to introduce ambiguity sets that contain all distributions in some neighbourhood of a prescribed reference distribution (typically the empirical distribution that is formed from historical data). These neighbourhoods

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

can be defined with respect to a discrepancy function between probability distributions such as a ϕ-divergence (Csiszár 1963) or a Wasserstein distance (Villani 2008). Unlike moment ambiguity sets, discrepancy-based ambiguity sets have a tunable size parameter (e.g. a radius) and can thus be shrunk to a singleton that contains only the reference distribution. If the reference distribution converges to the unknown true distribution and the size parameter decays to 0 as more historical data becomes available, then the DRO problem eventually reduces to the classical stochastic program under the true distribution. Early work on discrepancy-based ambiguity sets relies on the assumption that Z is a *discrete* random vector with a finite support set $\mathcal{Z}$. Extensions to discrepancy-based DRO problems with generic (possibly *continuous*) random vectors are due to Mohajerin Esfahani and Kuhn (2018), Zhao and Guan (2018), Blanchet and Murthy (2019), Zhang, Yang and Gao (2024*b*) and Gao and Kleywegt (2023), who construct ambiguity sets using optimal transport discrepancies. We refer to Kuhn, Mohajerin Esfahani, Nguyen and Shafieezadeh-Abadeh (2019) and Rahimian and Mehrotra (2022) for prior surveys of the DRO literature.

Historically, the term 'distributional robustness' has its roots in robust statistics. The term was coined by Huber (1981) to describe methods aimed at making robust decisions in the presence of outlier data points. This idea expanded upon earlier works by Box (1953, 1979), who explores robustness in situations where the underlying distribution deviates from normality, a common assumption underlying many statistical models. To address the challenges posed by outliers, statisticians have developed several contamination models, each offering a unique approach to mitigating data irregularities. The Huber contamination model, introduced by Huber (1964, 1968) and further developed by Hampel (1968, 1971), assumes that the observed data is drawn from a mixture of the true distribution and an arbitrary contaminating distribution. Neighbourhood contamination models define deviations from the true distribution in terms of statistical distances such as the total variation (Donoho and Liu 1988) or Wasserstein distances (Zhu, Jiao and Steinhardt 2022*a*, Liu and Loh 2023). More recently, data-dependent adaptive contamination models allow for a fraction of the observed data points to be replaced with points drawn from an arbitrary distribution (Diakonikolas *et al.* 2019, Zhu *et al.* 2022*a*). Interestingly, the optimistic counterpart of a DRO model, which optimizes in view of the best (as opposed to the worst) distribution in the ambiguity set, recovers many estimators from robust statistics (Blanchet, Li, Lin and Zhang 2024*b*, Jiang and Xie 2024). For a survey of recent advances in algorithmic robust statistics we refer to Diakonikolas and Kane (2023).

Robust and distributionally robust optimization have found manifold applications in machine learning. For example, popular regularizers from the machine learning literature are known to admit a robustness interpretation, which offers theoretical insights into the strong empirical performance of regularization in practice (Xu, Caramanis and Mannor 2009, Shafieezadeh-Abadeh, Kuhn and Mohajerin Esfahani 2019, Li, Lin, Blanchet and Nguyen 2022, Gao, Chen and Kleywegt 2024*b*).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Likewise, optimistic counterparts of DRO models that optimize in view of the best (as opposed to the worst) distribution in the ambiguity set give rise to upper confidence bound algorithms that are ubiquitous in the bandit and reinforcement learning literature (Blanchet *et al.* 2024*b*, Jiang and Xie 2024). DRO is also related to adversarial training, which aims to improve the generalization performance of a machine learning model by training it in view of adversarial examples (Goodfellow, Shlens and Szegedy 2015). Adversarial examples are perturbations of existing data points that are designed to mislead a model into making incorrect predictions.

There are also deep connections between DRO and extensions of stochastic (dynamic) programming that replace the expected value with coherent risk measures. Similar to the expected value, a risk measure maps random variables to extended real numbers. In contrast to the expected value, which is risk-neutral since it weighs positive and negative outcomes equally, risk measures most commonly assign greater weights to negative outcomes and thus account for the risk aversion frequently observed among decision-makers. Artzner, Delbaen, Eber and Heath (1999) and Delbaen (2002) show that risk measures satisfying the axioms of coherence as well as a Fatou property can be equivalently represented as worst-case expectations over specific sets of distributions. In other words, there is a direct link between optimizing worst-case expectations (as done in DRO) and optimizing coherent risk measures. A similar representation theorem has been developed for a class of nonlinear expectations, the so-called G-expectations that are based on the solution of a backward stochastic differential equation, in the financial mathematics literature (Peng 1997, 2007*a*,*b*, 2019). Peng (2023) shows that sublinear G-expectations are equivalent to worst-case expectations over specific families of distributions, thus creating a bridge between the theory of G-expectations and DRO.

Philosophically, DRO is related to the principle of *ambiguity aversion*, under which individuals prefer known risks over unknown risks even when the unknown risks promise potentially higher rewards. In the economics literature, the distinction between risky outcomes whose probabilities are known and ambiguous outcomes whose probabilities are (partially) unknown goes back to at least Keynes (1921) and Knight (1921). The concept of ambiguity aversion has been widely popularized through the Ellsberg paradox (Ellsberg 1961), a thought experiment under which people are asked to choose between betting on an urn with a known distribution of coloured balls (e.g. 50 red and 50 blue) and an urn with an unknown distribution of the same coloured balls (i.e. the proportion of red to blue is unknown). Despite the potential for equal or better odds, many people prefer to bet on the urn with the known distribution, that is, they display ambiguity aversion. The Ellsberg paradox challenges classical expected utility theory, and it has led to extensions such as the maxmin expected utility theory (Gilboa and Schmeidler 1989) that serve as theoretical underpinnings of DRO. Ambiguity aversion has subsequently been identified in countless empirical economic studies across financial markets (Epstein and Miao 2003, Bossaerts, Ghirardato, Guarnaschelli and Zame 2010), insurance markets (Cabantous 2007), individual decision-making (Dimmock, Kouwenberg

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

and Wakker 2016), macroeconomic policy (Hansen and Sargent 2010), auctions (Salo and Weber 1995) and games of trust (Li, Turmunkh and Wakker 2019*b*).

There is also substantial medical and neuroscientific evidence that supports the presence of ambiguity aversion. Hsu *et al.* (2005) found that the amygdala, a key emotional processing centre in the brain, becomes more active when individuals are confronted with ambiguity compared to situations with known probabilities, indicating its role in driving ambiguity aversion. A meta-analysis by Krain *et al.* (2006) highlights the involvement of the prefrontal cortex, which is responsible for higher-order cognitive control, rational decision-making and emotional regulation, in processing ambiguity. In addition, a meta-analysis of Wu *et al.* (2021) shows that processing risk and ambiguity both rely on the anterior insula. Risk processing additionally activates the dorsomedial prefrontal cortex and ventral striatum, whereas ambiguity processing specifically engages the dorsolateral prefrontal cortex, inferior parietal lobe, and right anterior insula. This supports the notion that distinct neural mechanisms are engaged when individuals face ambiguous versus risky decisions. Genetic factors may influence an individual's tendency toward ambiguity aversion. He *et al.* (2010) link certain genetic polymorphisms to the performance of individuals in decision-making under risk and ambiguity. In a separate study, Buckert, Schwieren, Kudielka and Fiebach (2014) examine how hormonal changes, such as higher cortisol levels which are linked to stress and anxiety, affect decision-making under risk and ambiguity. These findings collectively suggest that perceptions of risk and ambiguity are not just a cognitive phenomenon but also influenced by brain structures and genetic and hormonal factors that shape individual differences in decision-making under ambiguity. Finally, we mention Hartley and Somerville (2015) and Blankenstein, Crone, van den Bos and van Duijvenvoorde (2016), who examine how ambiguity aversion differs between children, adolescents and adults, and Hayden, Heilbronner and Platt (2010), who observed that rhesus macaque monkeys also exhibit ambiguity aversion when offered the choice between risky and ambiguous games of large and small juice outcomes.

The remainder of this survey is structured as follows. A significant part of our analysis is dedicated to studying the worst-case expectation $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)]$, which constitutes the objective function of the DRO problem (1.2). Evaluating this expression typically requires the solution of a semi-infinite optimization problem over infinitely many variables that characterize the probability distribution $\mathbb{P}$, subject to finitely many constraints imposed by the ambiguity set $\mathcal{P}$. This problem, which we refer to as *nature's subproblem*, is the key feature that distinguishes the DRO problem (1.2) from deterministic, stochastic and robust optimization problems. Sections 2 and 3 review commonly studied ambiguity sets $\mathcal{P}$ and their topological properties, focusing especially on conditions under which nature's subproblem attains its optimal value. Sections 4 and 5 develop a duality theory for nature's subproblem that allows us to upper-bound or equivalently reformulate the worst-case expectation with a semi-infinite optimization problem over finitely many dual decision variables that are subjected to infinitely many constraints. This duality

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

framework lays the foundations for the analytical solution of nature's subproblem in Section 6, which relies on constructing primal and dual feasible solutions that yield the same objective value and thus enjoy strong duality. Sections 7 and 8 leverage the same duality theory to develop equivalent reformulations and conservative approximations of nature's subproblem as well as the overall DRO problem (1.2). Section 9 demonstrates how the duality theory gives rise to numerical solution techniques for nature's subproblem and the full DRO problem. Finally, Section 10 reviews the statistical guarantees enjoyed by different ambiguity sets.

Length restrictions dictated difficult trade-offs in the choice of topics covered by this survey. We decided to focus on the most commonly used ambiguity sets and to only briefly review other possible choices, such as marginal ambiguity sets, ambiguity sets with structural constraints (including, for example, symmetry and unimodality), Sinkhorn ambiguity sets or conditional relative entropy ambiguity sets. Likewise, we do not cover the important but somewhat more advanced topics of distributionally favourable optimization and decision randomization. Finally, we focus on single-stage problems where the uncertainty is fully resolved after the here-and-now decision $x \in \mathcal{X}$ is taken; two-stage and multi-stage DRO problems, where uncertainty unfolds over time and recourse decisions are possible, are reviewed by Delage and Iancu (2015) and Yanıkoğlu, Gorissen and den Hertog (2019).

1.1. Notation

All vector spaces considered in this paper are defined over the real numbers. For brevity, we simply refer to them as 'vector spaces' instead of 'real vector spaces'. We use $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ to denote the extended reals. The effective domain of a function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is defined as $\mathrm{dom}(f) = \{z \in \mathbb{R}^d \colon f(z) < \infty\}$, and the epigraph of f is defined as $\mathrm{epi}(f) = \{(z, \alpha) \in \mathbb{R}^d \times \mathbb{R} \colon f(z) \leq \alpha\}$. We say that f is proper if $\mathrm{dom}(f) \neq \emptyset$ and $f(z) > -\infty$ for all $z \in \mathbb{R}^d$. The convex conjugate of f is the function $f^*\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ defined by $f^*(y) = \sup_{z \in \mathbb{R}^d} y^\top z - f(z)$. A convex function f is called closed if it is proper and lower semicontinuous or if it is identically equal to $+\infty$ or to $-\infty$. One can show that f is closed if and only if it coincides with its bi-conjugate f^{**}, that is, with the conjugate of f^*. If f is proper, convex and lower semicontinuous, then its recession function $f^\infty\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is defined by $f^\infty(z) = \lim_{\alpha \to \infty} \alpha^{-1}(f(z_0 + \alpha z) - f(z_0))$, where z_0 is any point in $\mathrm{dom}(f)$ (Rockafellar 1970, Theorem 8.5). The perspective of f is the function $f^\pi\colon \mathbb{R}^d \times \mathbb{R} \to \overline{\mathbb{R}}$ defined by $f^\pi(z,t) = tf(z/t)$ if $t > 0$, $f^\pi(z,t) = f^\infty(z)$ if $t = 0$ and $f^\pi(z,t) = \infty$ if $t < 0$. One can show that f^π is proper, convex and lower semicontinuous (Rockafellar 1970, p. 67). When there is no risk of confusion, we occasionally use $tf(z/t)$ to denote $f^\pi(z,t)$ even if $t = 0$. The indicator function $\delta_{\mathcal{Z}}\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ of a set $\mathcal{Z} \subseteq \mathbb{R}^d$ is defined by $\delta_{\mathcal{Z}}(z) = 0$ if $z \in \mathcal{Z}$ and $\delta_{\mathcal{Z}}(z) = \infty$ if $z \notin \mathcal{Z}$. The conjugate $\delta^*_{\mathcal{Z}}$ of $\delta_{\mathcal{Z}}$ is called the support function of $\mathcal{Z}$. Thus it satisfies $\delta^*_{\mathcal{Z}}(y) = \sup_{z \in \mathcal{Z}} y^\top z$. Random objects are denoted by capital letters (e.g. Z) and their realizations are denoted by the corresponding lowercase letters (e.g. z). For

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

any closed set $\mathcal{Z} \subseteq \mathbb{R}^d$, we use $\mathcal{M}(\mathcal{Z})$ to denote the space of all finite signed Borel measures on $\mathcal{Z}$, while $\mathcal{M}_+(\mathcal{Z})$ stands for the convex cone of all (non-negative) Borel measures in $\mathcal{M}(\mathcal{Z})$, and $\mathcal{P}(\mathcal{Z})$ stands for the convex set of all probability distributions in $\mathcal{M}_+(\mathcal{Z})$. The expectation operator with respect to $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ is defined by $\mathbb{E}_{\mathbb{P}}[f(Z)] = \int_{\mathcal{Z}} f(z)\,\mathrm{d}\mathbb{P}(z)$ for any Borel function $f\colon \mathcal{Z} \to \overline{\mathbb{R}}$. If the integrals of the positive and the negative parts of f both evaluate to ∞, then we define $\mathbb{E}_{\mathbb{P}}[f(Z)]$ 'adversarially'. That is, we set $\mathbb{E}_{\mathbb{P}}[f(Z)] = \infty$ $(-\infty)$ if the integral appears in the objective function of a minimization (maximization) problem. The Dirac probability distribution that assigns unit probability to $z \in \mathcal{Z}$ is denoted as δ_z. The Dirac distribution δ_z should not be confused with the indicator function $\delta_{\{z\}}$ of the singleton $\{z\}$. For any $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ and any Borel-measurable transformation $f\colon \mathcal{Z} \to \mathcal{Z}'$ between Borel sets $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\mathcal{Z}' \subseteq \mathbb{R}^{d'}$, we let $\mathbb{P} \circ f^{-1}$ denote the pushforward distribution of $\mathbb{P}$ under f. Thus, if Z is a random vector on $\mathcal{Z}$ governed by $\mathbb{P}$, then $f(Z)$ is a random vector on $\mathcal{Z}'$ governed by $\mathbb{P} \circ f^{-1}$. The closure, the interior and the relative interior of a set $\mathcal{Z} \subseteq \mathbb{R}^d$ are denoted by $\mathrm{cl}(\mathcal{Z})$, $\mathrm{int}(\mathcal{Z})$ and $\mathrm{rint}(\mathcal{Z})$, respectively. We use $\mathbb{R}^d_+$ and $\mathbb{R}^d_{++}$ to denote the non-negative orthant in $\mathbb{R}^d$ and its interior. In addition, we use $\mathbb{S}^d$ to denote the space of all symmetric matrices in $\mathbb{R}^{d\times d}$. The cone of positive semidefinite matrices in $\mathbb{S}^d$ is denoted by $\mathbb{S}^d_+$, and $\mathbb{S}^d_{++}$ stands for its interior, that is, the set of all positive definite matrices in $\mathbb{S}^d$. The truth value $\mathbb{1}_{\mathcal{E}}$ of a logical statement evaluates to 1 if $\mathcal{E}$ is true and to 0 otherwise. The set of all natural numbers $\{1, 2, 3, \ldots\}$ is denoted by $\mathbb{N}$, and $[n] = \{1, \ldots, n\}$ stands for the set of all integers up to $n \in \mathbb{N}$.

2. Ambiguity sets

An ambiguity set $\mathcal{P}$ is a family of probability distributions on a common measurable space. Throughout this paper we assume that $\mathcal{P} \subseteq \mathcal{P}(\mathcal{Z})$, where $\mathcal{P}(\mathcal{Z})$ denotes the entirety of *all* Borel probability distributions on a closed set $\mathcal{Z} \subseteq \mathbb{R}^d$. This section reviews popular classes of ambiguity sets. For each class, we first give a formal definition and provide historical background information. Subsequently, we exemplify important instances of ambiguity sets and highlight how they are used.

2.1. *Moment ambiguity sets*

A moment ambiguity set is a family of probability distributions that satisfy finitely many (generalized) moment conditions. Formally, it can thus be represented as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathbb{E}_{\mathbb{P}}[f(Z)] \in \mathcal{F}\}, \tag{2.1}$$

where $f\colon \mathcal{Z} \to \mathbb{R}^m$ is a Borel-measurable moment function, and $\mathcal{F} \subseteq \mathbb{R}^m$ is an uncertainty set. By definition, the moment ambiguity set (2.1) thus contains all probability distributions $\mathbb{P}$ supported on $\mathcal{Z}$ whose generalized moments $\mathbb{E}_{\mathbb{P}}[f(Z)]$ are well-defined and belong to the uncertainty set $\mathcal{F}$. Ambiguity sets of the type (2.1) were first studied by Isii (1960, 1962) and Karlin and Studden (1966)

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

to establish the sharpness of generalized Chebyshev inequalities. The following subsections review popular instances of the moment ambiguity set.

2.1.1. *Support-only ambiguity sets*

The support-only ambiguity set contains all probability distributions supported on $\mathcal{Z} \subseteq \mathbb{R}^d$, that is, $\mathcal{P} = \mathcal{P}(\mathcal{Z})$. It can be viewed as an instance of (2.1) with $f(z) = 1$ and $\mathcal{F} = \{1\}$. Any DRO problem with ambiguity set $\mathcal{P}(\mathcal{Z})$ is ostensibly equivalent to a classical robust optimization problem with uncertainty set $\mathcal{Z}$, that is,

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] = \inf_{x\in\mathcal{X}} \sup_{z\in\mathcal{Z}} \ell(x, z).$$

For a comprehensive review of the theory and applications of robust optimization we refer to Ben-Tal and Nemirovski (1998, 1999*a*, 2000, 2002), Bertsimas and Sim (2004), Ben-Tal *et al.* (2009), Bertsimas, Brown and Caramanis (2011), Ben-Tal, den Hertog and Vial (2015*a*) and Bertsimas and den Hertog (2022).

If the uncertainty set $\mathcal{Z}$ covers a fraction of $1 - \varepsilon$ of the total probability mass of some distribution $\mathbb{P}$, then the worst-case loss $\sup_{z\in\mathcal{Z}} \ell(x, z)$ is guaranteed to exceed the $(1 - \varepsilon)$-quantile of $\ell(x, Z)$ under $\mathbb{P}$. This can be achieved by leveraging prior structural information or statistical data from $\mathbb{P}$. For example, $\mathbb{P}(Z \in \mathcal{Z}) \geq 1 - \varepsilon$ may hold (with certainty) if $\mathcal{Z}$ is an appropriately sized intersection of half-spaces and ellipsoids and if Z has independent, symmetric, unimodal and/or sub-Gaussian components under $\mathbb{P}$ (Bertsimas and Sim 2004, Janak, Lin and Floudas 2007, Ben-Tal *et al.* 2009, Li, Ding and Floudas 2011, Bertsimas, den Hertog and Pauphilet 2021). Alternatively, it may hold (with high confidence) if $\mathcal{Z}$ is constructed from independent samples from $\mathbb{P}$ by using statistical hypothesis tests (Postek, den Hertog and Melenberg 2016, Bertsimas, Gupta and Kallus 2018*b*,*a*), quantile estimation (Hong, Huang and Lam 2020) or learning-based methods (Han, Shang and Huang 2021, Goerigk and Kurtz 2023, Wang, Becker, Van Parys and Stellato 2023).

2.1.2. *Markov ambiguity sets*

Markov's inequality provides an upper bound on the probability that a non-negative univariate random variable Z with mean $\mu \geq 0$ exceeds a positive threshold $\tau > 0$. Formally, it states that $\mathbb{P}(Z \geq \tau) \leq \mu/\tau$ for every possible probability distribution of Z in the ambiguity set $\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}_+) \colon \mathbb{E}_{\mathbb{P}}[Z] = \mu\}$. If $\mu \leq \tau$, then Markov's inequality is sharp, that is, there exists a probability distribution $\mathbb{P}^\star \in \mathcal{P}$ for which the inequality holds as an equality. Indeed, the distribution $\mathbb{P}^\star = (1 - \mu/\tau)\delta_0 + \mu/\tau\delta_\tau$, where δ_z is the Dirac distribution that places point mass as $z \in \mathbb{R}$, is an element of $\mathcal{P}$ and satisfies $\mathbb{P}(Z \geq \tau) = \mu/\tau$. These insights imply that $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(Z \geq \tau) = \mu/\tau$ and that the supremum is attained by $\mathbb{P}^\star$ whenever $\mu \leq \tau$. Thus Markov's bound can be interpreted as the optimal value of a DRO problem. It is therefore common to refer to $\mathcal{P}$ as a Markov ambiguity set. More generally, we define the Markov ambiguity set corresponding to a closed support set $\mathcal{Z} \subseteq \mathbb{R}^d$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

and a mean vector $\mu \in \mathbb{R}^d$ as a family of multivariate distributions of the form

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{E}_{\mathbb{P}}[Z] = \mu\}. \tag{2.2}$$

Thus the Markov ambiguity set (2.2) contains all distributions supported on $\mathcal{Z}$ that share the same mean vector μ. However, these distributions may have dramatically different shapes and higher-order moments. Worst-case expectations over Markov ambiguity sets are sometimes used as efficiently computable upper bounds on the expected cost-to-go functions in stochastic programming. If the cost-to-go functions are concave in the uncertain problem parameters, then these worst-case expectations are closely related to Jensen's inequality (Jensen 1906); see also Section 6.1. If the cost-to-go functions are convex and $\mathcal{Z}$ is a polyhedron, on the other hand, then these worst-case expectations are related to the Edmundson–Madansky inequality (Edmundson 1956, Madansky 1959); see also Section 6.2.

2.1.3. Chebyshev ambiguity sets

Chebyshev's inequality provides an upper bound on the probability that a univariate random variable Z with finite mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ deviates from its mean by more than $k > 0$ standard deviations. Formally, it states that $\mathbb{P}(|Z - \mu| \geq k\sigma) \leq 1/k^2$ for every possible probability distribution of Z in the ambiguity set $\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}) \colon \mathbb{E}_{\mathbb{P}}[Z] = \mu,\ \mathbb{E}_{\mathbb{P}}[Z^2] = \sigma^2 + \mu^2\}$. Chebyshev's inequality is sharp if $k \geq 1$. Indeed, one readily verifies that the distribution

$$\mathbb{P}^\star = \frac{1}{2k^2}\delta_{\mu - k\sigma} + \left(1 - \frac{1}{k^2}\right)\delta_\mu + \frac{1}{2k^2}\delta_{\mu + k\sigma}$$

is an element of $\mathcal{P}$ and satisfies $\mathbb{P}(|Z - \mu| \geq k\sigma) = 1/k^2$. These insights imply that $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(|Z - \mu| \geq k\sigma) = 1/k^2$ and that the supremum is attained for $k \geq 1$. Thus Chebyshev's bound can be interpreted as the optimal value of a DRO problem. It is therefore common to refer to $\mathcal{P}$ as a Chebyshev ambiguity set. More generally, we define the Chebyshev ambiguity set corresponding to a closed support set $\mathcal{Z} \subseteq \mathbb{R}^d$, mean vector $\mu \in \mathbb{R}^d$ and second-order moment matrix $M \in \mathbb{S}_+^d$, $M \succeq \mu\mu^\top$, as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{E}_{\mathbb{P}}[Z] = \mu,\ \mathbb{E}_{\mathbb{P}}[ZZ^\top] = M\}. \tag{2.3}$$

Thus the Chebyshev ambiguity set (2.3) contains all distributions supported on $\mathcal{Z}$ that share the same mean vector μ and second-order moment matrix M (and thus also the same covariance matrix $\Sigma = M - \mu\mu^\top \in \mathbb{S}_+^d$). However, these distributions may have dramatically different shapes and higher-order moments.

The Chebyshev ambiguity set (2.3) captures the distributional information relevant for multivariate Chebyshev inequalities (Lal 1955, Marshall and Olkin 1960, Tong 1980, Rujeerapaiboon, Kuhn and Wiesemann 2018). In operations research, Chebyshev ambiguity sets are routinely used since the seminal work of Scarf (1958) on the distributionally robust newsvendor, which is widely perceived as the first paper on DRO. Since then a wealth of DRO models with Chebyshev ambiguity sets have emerged in the context of newsvendor and portfolio selection problems.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

These models involve a wide range of different decision criteria such as the expected value (Gallego and Moon 1993, Natarajan and Linyi 2007, Popescu 2007), the value-at-risk (El Ghaoui *et al.* 2003, Xu, Caramanis and Mannor 2012*b*, Zymler, Kuhn and Rustem 2013*a*,*b*, Rujeerapaiboon, Kuhn and Wiesemann 2016, Yang and Xu 2016, Zhang, Jiang and Shen 2018), the conditional value-at-risk (Natarajan, Sim and Uichanco 2010, Chen, He and Zhang 2011, Zymler *et al.* 2013*b*, Hanasusanto, Kuhn, Wallace and Zymler 2015*a*), spectral risk measures (Li 2018) and distortion risk measures (Cai, Li and Mao 2023, Pesenti, Wang and Wang 2024), as well as minimax regret criteria (Yue, Chen and Wang 2006, Perakis and Roels 2008). Besides this, Chebyshev ambiguity sets have found numerous applications in option and stock pricing (Bertsimas and Popescu 2002), statistics and machine learning (Lanckriet, El Ghaoui, Bhattacharyya and Jordan 2001, 2002, Strohmann and Grudic 2002, Huang *et al.* 2004, Bhattacharyya 2004, Farnia and Tse 2016, Nguyen *et al.* 2019, Rontsis, Osborne and Goulart 2020), stochastic programming (Birge and Wets 1986, Dulá and Murthy 1992, Dokov and Morton 2005, Bertsimas, Doan, Natarajan and Teo 2010, Natarajan, Teo and Zheng 2011), control (Van Parys, Kuhn, Goulart and Morari 2015, Yang 2018, Xin and Goldberg 2021, 2022), the operation of power systems (Xie and Ahmed 2017, Zhao and Jiang 2017), complex network analysis (Van Leeuwaarden and Stegehuis 2021, Brugman, Van Leeuwaarden and Stegehuis 2022), queuing systems (van Eekelen, Hanasusanto, Hasenbein and van Leeuwaarden 2025), healthcare (Mak, Rong and Zhang 2015, Shehadeh, Cohn and Jiang 2020) and extreme event analysis (Lam and Mottet 2017), among others.

2.1.4. Chebyshev ambiguity sets with uncertain moments

Working with Chebyshev ambiguity sets is appropriate when the first- and second-order moments of $\mathbb{P}$ are known, while all higher-order moments are unknown. In practice, however, even the first- and second-order moments are never known with absolute certainty. Instead, they must be estimated from statistical data and are thus subject to estimation errors. This prompted El Ghaoui *et al.* (2003) to introduce a Chebyshev ambiguity set with uncertain moments, which can be represented as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon (\mathbb{E}_{\mathbb{P}}[Z], \mathbb{E}_{\mathbb{P}}[ZZ^\top]) \in \mathcal{F}\}. \tag{2.4}$$

Here $\mathcal{F} \subseteq \mathbb{R}^d \times \mathbb{S}_+^d$ is a convex set that captures the moment uncertainty. Clearly, $\mathcal{P}$ can be expressed as a union of crisp Chebyshev ambiguity sets, that is, we have

$$\mathcal{P} = \bigcup_{(\mu,M)\in\mathcal{F}} \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathbb{E}_{\mathbb{P}}[Z] = \mu,\ \mathbb{E}_{\mathbb{P}}[ZZ^\top] = M\}.$$

Note that the Chebyshev ambiguity set with uncertain moments encapsulates the support-only ambiguity set, the Markov ambiguity set and the Chebyshev ambiguity set as special cases. They are recovered by setting $\mathcal{F} = \mathbb{R}^d \times \mathbb{S}_+^d$, $\mathcal{F} = \{\mu\} \times \mathbb{S}_+^d$, and $\mathcal{F} = \{\mu\} \times \{M\}$, respectively.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

El Ghaoui *et al.* (2003) capture the uncertainty in the moments using the box

$$\mathcal{F} = \{(\mu, M) \in \mathbb{R}^d \times \mathbb{S}_+^d : \underline{\mu} \leq \mu \leq \overline{\mu},\ \underline{M} \preceq M \preceq \overline{M}\}$$

parametrized by the moment bounds $\underline{\mu}, \overline{\mu} \in \mathbb{R}^d$ and $\underline{M}, \overline{M} \in \mathbb{S}_+^d$.

Given noisy estimates $\hat{\mu}$ and $\hat{\Sigma}$ for the unknown mean vector and covariance matrix of $\mathbb{P}$, respectively, Delage and Ye (2010) propose the ambiguity set

$$\mathcal{P} = \left\{ \mathbb{P} \in \mathcal{P}(\mathcal{Z}) : \begin{array}{l} (\mathbb{E}_{\mathbb{P}}[Z] - \hat{\mu})^\top \hat{\Sigma}^{-1} (\mathbb{E}_{\mathbb{P}}[Z] - \hat{\mu}) \leq \gamma_1 \\ \mathbb{E}_{\mathbb{P}}[(Z - \hat{\mu})(Z - \hat{\mu})^\top] \preceq \gamma_2 \hat{\Sigma} \end{array} \right\}.$$

By construction, $\mathcal{P}$ contains all distributions on $\mathcal{Z}$ whose first-order moments reside in an ellipsoid with centre $\hat{\mu}$ and whose second-order moments (relative to $\hat{\mu}$) reside in a semidefinite cone with apex $\gamma_2 \hat{\Sigma}$. An elementary calculation reveals that

$$\mathbb{E}_{\mathbb{P}}[(Z - \hat{\mu})(Z - \hat{\mu})^\top] = \mathbb{E}_{\mathbb{P}}[ZZ^\top] - \mathbb{E}_{\mathbb{P}}[Z]\hat{\mu}^\top - \hat{\mu}\mathbb{E}_{\mathbb{P}}[Z]^\top + \hat{\mu}\hat{\mu}^\top.$$

Thus $\mathcal{P}$ can be viewed as a Chebyshev ambiguity set with uncertain moments. Indeed, $\mathcal{P}$ is an instance of (2.4) if we define the moment uncertainty set as

$$\mathcal{F} = \left\{ (\mu, M) \in \mathbb{R}^d \times \mathbb{S}_+^d : \begin{array}{l} (\mu - \hat{\mu})^\top \hat{\Sigma} (\mu - \hat{\mu}) \leq \gamma_1 \\ M - \mu\hat{\mu}^\top - \hat{\mu}\mu^\top + \hat{\mu}\hat{\mu}^\top \preceq \gamma_2 \hat{\Sigma} \end{array} \right\}.$$

Delage and Ye (2010) show that if $\hat{\mu}$ and $\hat{\Sigma}$ are set to the sample mean and the sample covariance matrix constructed from a finite number of independent samples from $\mathbb{P}$, respectively, then one can tune the size parameters $\gamma_1 \geq 0$ and $\gamma_2 \geq 1$ to ensure that $\mathbb{P}$ belongs to $\mathcal{P}$ with any desired confidence.

Chebyshev as well as Markov ambiguity sets with uncertain moments have found various applications ranging from control (Nakao, Jiang and Shen 2021) to integer stochastic programming (Bertsimas, Natarajan and Teo 2004, Cheng, Delage and Lisser 2014), portfolio optimization (Natarajan *et al.* 2010), extreme event analysis (Bai, Lam and Zhang 2023*b*) and mechanism design and pricing (Bergemann and Schlag 2008, Bandi and Bertsimas 2014, Koçyiğit, Iyengar, Kuhn and Wiesemann 2020, Koçyiğit, Rujeerapaiboon and Kuhn 2022, Chen, Hu and Wang 2024*b*, Bayrak, Koçyiğit, Kuhn and Pınar 2025, Anunrojwong, Balseiro and Besbes 2024), among many others.

The uncertainty set $\mathcal{F}$ for the first- and second-order moments of $\mathbb{P}$ often corresponds to a neighbourhood of a nominal mean–covariance pair $(\hat{\mu}, \hat{\Sigma})$ with respect to some measure of discrepancy. For example, matrix norms such as the Frobenius norm, the spectral norm or the nuclear norm (Bernstein 2009, § 9) provide natural measures to quantify the dissimilarity of covariance matrices. The discrepancy between two mean–covariance pairs (μ, Σ) and $(\hat{\mu}, \hat{\Sigma})$ can also be defined as the discrepancy between the normal distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ with respect to a probability metric or an information-theoretic divergence such as the Kullback–Leibler divergence (Kullback 1959), the Fisher–Rao distance (Atkinson and Mitchell 1981) or other spectral divergences (Zorzi 2014).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

As we will discuss in more detail in Section 2.3, the 2-Wasserstein distance between two normal distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ coincides with the Gelbrich distance between the underlying mean–covariance pairs (μ, Σ) and $(\hat{\mu}, \hat{\Sigma})$. In the following, we first provide a formal definition of the Gelbrich distance and then exemplify how it can be used to define a moment uncertainty set $\mathcal{F}$.

Definition 2.1 (Gelbrich distance). The Gelbrich distance between two mean–covariance pairs (μ, Σ) and $(\hat{\mu}, \hat{\Sigma})$ in $\mathbb{R}^d \times \mathbb{S}_+^d$ is given by

$$\mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) = \sqrt{\|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}\big(\Sigma + \hat{\Sigma} - 2(\hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2})^{1/2}\big)}.$$

The Gelbrich distance is non-negative, symmetric and subadditive, and it vanishes if and only if $(\mu, \Sigma) = (\hat{\mu}, \hat{\Sigma})$. Thus it represents a metric on $\mathbb{R}^d \times \mathbb{S}_+^d$ (Givens and Shortt 1984, p. 239). When $\mu = \hat{\mu}$, then the Gelbrich distance collapses to the Bures distance between Σ and $\hat{\Sigma}$, which was conceived as a measure of dissimilarity between density matrices in quantum information theory. The Bures distance is known to induce a Riemannian metric on the space of positive semidefinite matrices (Bhatia, Jain and Lim 2018, 2019). When Σ and $\hat{\Sigma}$ are simultaneously diagonalizable, then their Bures distance coincides with the Hellinger distance between their spectra. The Hellinger distance is closely related to the Fisher–Rao metric ubiquitous in information theory (Liese and Vajda 1987). Even though the Gelbrich distance is non-convex, the *squared* Gelbrich distance is jointly convex in both of its arguments. This is an immediate consequence of the following proposition, discovered by Olkin and Pukelsheim (1982), Dowson and Landau (1982), Givens and Shortt (1984) and Panaretos and Zemel (2020).

Proposition 2.2 (SDP representation of the Gelbrich distance). For any mean–covariance pairs (μ, Σ) and $(\hat{\mu}, \hat{\Sigma})$ in $\mathbb{R}^d \times \mathbb{S}_+^d$, we have

$$\mathrm{G}^2((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) = \begin{cases} \min\limits_{C \in \mathbb{R}^{d\times d}} & \|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}(\Sigma + \hat{\Sigma} - 2C) \\ \text{s.t.} & \begin{bmatrix} \Sigma & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \succeq 0. \end{cases} \tag{2.5}$$

Proof. Throughout the proof we keep μ, $\hat{\mu}$ and Σ fixed and treat $\hat{\Sigma}$ as a parameter. We also use $f(\hat{\Sigma})$ as shorthand for the left-hand side of (2.5) and $g(\hat{\Sigma})$ as shorthand for the right-hand side of (2.5). Elementary manipulations show that

$$g(\hat{\Sigma}) = \|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}(\Sigma + \hat{\Sigma}) - \begin{cases} \max\limits_{C \in \mathbb{R}^{d\times d}} & \mathrm{Tr}(2C) \\ \text{s.t.} & \begin{bmatrix} \Sigma & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \succeq 0. \end{cases} \tag{2.6}$$

The maximization problem in (2.6) is dual to the following minimization problem:

$$\begin{array}{ll} \inf\limits_{A_{11}, A_{22} \in \mathbb{S}^d} & \mathrm{Tr}(A_{11}\Sigma) + \mathrm{Tr}(A_{22}\hat{\Sigma}) \\ \text{s.t.} & \begin{bmatrix} A_{11} & I_d \\ I_d & A_{22} \end{bmatrix} \succeq 0 \end{array}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Strong duality holds because $A_{11} = A_{22} = 2I_d$ constitutes a Slater point for the dual problem (Ben-Tal and Nemirovski 2001, Theorem 2.4.1). The existence of a Slater point further implies that the primal maximization problem in (2.6) as well as the minimization problem in (2.5) are solvable. By Bernstein (2009, Corollary 8.2.2), both A_{11} and A_{22} must be positive definite in order to be dual feasible. Thus they are invertible. We can therefore employ a Schur complement argument (Ben-Tal and Nemirovski 2001, Lemma 4.2.1) to simplify the dual problem to

$$\inf_{A_{11} \succeq A_{22}^{-1} \succ 0} \operatorname{Tr}(A_{11}\Sigma) + \operatorname{Tr}(A_{22}\hat{\Sigma}) = \inf_{A_{22} \succ 0} \operatorname{Tr}(A_{22}^{-1}\Sigma) + \operatorname{Tr}(A_{22}\hat{\Sigma}), \tag{2.7}$$

where the equality holds because $\Sigma \succeq 0$. The optimal value of the resulting minimization problem is concave and upper semicontinuous in $\hat{\Sigma}$ because it constitutes a pointwise infimum of affine functions of $\hat{\Sigma}$. Thus $g(\hat{\Sigma})$ is convex and lower semicontinuous. We now show that if $\hat{\Sigma} \succ 0$, then the convex minimization problem over A_{22} in (2.7) can be solved in closed form. To this end, we construct a positive definite matrix $A_{22}^\star$ that satisfies the problem's first-order optimality condition

$$\hat{\Sigma} - A_{22}^{-1}\Sigma A_{22}^{-1} = 0 \quad \Longleftrightarrow \quad A_{22}\hat{\Sigma}A_{22} - \Sigma = 0.$$

Indeed, multiplying the quadratic equation on the right from both sides with $\hat{\Sigma}^{1/2}$ yields the equivalent equation $(\hat{\Sigma}^{1/2}A_{22}\hat{\Sigma}^{1/2})^2 = \hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2}$. As $\hat{\Sigma} \succ 0$, this equation is uniquely solved by $A_{22}^\star = \hat{\Sigma}^{-1/2}(\hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2})^{1/2}\hat{\Sigma}^{-1/2}$. Substituting $A_{22}^\star$ into (2.7) reveals that the optimal value of the dual minimization problem is given by $2\operatorname{Tr}((\hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2})^{1/2})$. Substituting this value into (2.6) then shows that $g(\hat{\Sigma}) = f(\hat{\Sigma})$ whenever $\hat{\Sigma} \succ 0$.

It remains to be shown that $g(\hat{\Sigma}) = f(\hat{\Sigma})$ if $\hat{\Sigma}$ is singular. To this end, we recall from Nguyen, Shafieezadeh-Abadeh, Kuhn and Mohajerin Esfahani (2023*b*, Lemma A.2) that the matrix square root is continuous on $\mathbb{S}_+^d$, which implies that $f(\hat{\Sigma})$ is continuous on $\mathbb{S}_+^d$. For any singular $\hat{\Sigma} \succeq 0$, we thus have

$$f(\hat{\Sigma}) = \liminf_{\hat{\Sigma}' \to \hat{\Sigma},\, \hat{\Sigma}' \succ 0} f(\hat{\Sigma}') = \liminf_{\hat{\Sigma}' \to \hat{\Sigma},\, \hat{\Sigma}' \succ 0} g(\hat{\Sigma}') = g(\hat{\Sigma}).$$

Here the first equality exploits the continuity of f, and the second equality holds because $f(\hat{\Sigma}') = g(\hat{\Sigma}')$ for every $\hat{\Sigma}' \succ 0$. The third equality follows from the convexity and lower semicontinuity of g, which imply that the limit inferior can neither be larger nor smaller than $g(\hat{\Sigma})$, respectively. This completes the proof. □

Proposition 2.2 shows that the squared Gelbrich distance coincides with the optimal value of a tractable semidefinite program. This makes the Gelbrich distance attractive for computation. As a by-product, the proof of Proposition 2.2 reveals that the squared Gelbrich distance is convex as well as continuous on its domain.

Following Nguyen, Shafiee, Filipović and Kuhn (2021), we can now introduce the Gelbrich ambiguity set as an instance of the Chebyshev ambiguity set (2.4)

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

with uncertain moments. The corresponding moment uncertainty set is given by

$$\mathcal{F} = \left\{ (\mu, M) \in \mathbb{R}^d \times \mathbb{S}_+^d : \begin{array}{l} \exists \Sigma \in \mathbb{S}_+^d \text{ with } M = \Sigma + \mu\mu^\top, \\ \mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) \leq r \end{array} \right\}, \tag{2.8}$$

where $(\hat{\mu}, \hat{\Sigma})$ is a nominal mean–covariance pair, and the radius $r \geq 0$ serves as a tunable size parameter. Below we refer to $\mathcal{F}$ as the Gelbrich uncertainty set. The next proposition establishes basic topological and computational properties of $\mathcal{F}$.

Proposition 2.3 (Gelbrich uncertainty set). The uncertainty set $\mathcal{F}$ defined in (2.8) is convex and compact. In addition, it admits the semidefinite representation

$$\mathcal{F} = \left\{ (\mu, M) \in \mathbb{R}^d \times \mathbb{S}_+^d : \begin{array}{l} \exists C \in \mathbb{R}^{d\times d},\ U \in \mathbb{S}_+^d \text{ with} \\ \|\hat{\mu}\|_2^2 - 2\mu^\top \hat{\mu} + \mathrm{Tr}(M + \hat{\Sigma} - 2C) \leq r^2, \\ \begin{bmatrix} M - U & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \geq 0,\ \begin{bmatrix} U & \mu \\ \mu^\top & 1 \end{bmatrix} \geq 0 \end{array} \right\}.$$

Proof. The proof exploits the semidefinite representation of the squared Gelbrich distance established in Proposition 2.2. Note first that if $M = \Sigma + \mu\mu^\top$, then

$$\|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}(\Sigma + \hat{\Sigma} - 2C) = \|\hat{\mu}\|_2^2 - 2\mu^\top \hat{\mu} + \mathrm{Tr}(M + \hat{\Sigma} - 2C).$$

By Proposition 2.2, the Gelbrich uncertainty set $\mathcal{F}$ can thus be represented as

$$\mathcal{F} = \left\{ (\mu, M) \in \mathbb{R}^d \times \mathbb{S}_+^d : \begin{array}{l} \exists C \in \mathbb{R}^{d\times d} \text{ with} \\ \|\hat{\mu}\|_2^2 - 2\mu^\top \hat{\mu} + \mathrm{Tr}(M + \hat{\Sigma} - 2C) \leq r^2, \\ \begin{bmatrix} M - \mu\mu^\top & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \geq 0 \end{array} \right\}.$$

A standard Schur complement argument further reveals that

$$\begin{bmatrix} M - \mu\mu^\top & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \geq 0 \iff \exists U \in \mathbb{S}_+^d \text{ with } \begin{bmatrix} M - U & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \geq 0,\ \begin{bmatrix} U & \mu \\ \mu^\top & 1 \end{bmatrix} \geq 0.$$

Hence the Gelbrich uncertainty set $\mathcal{F}$ admits the semidefinite representation given in the proposition statement. Convexity is evident from this representation, which expresses $\mathcal{F}$ as the projection of a set defined by conic inequalities in a lifted space.

It remains to be shown that $\mathcal{F}$ is compact. To this end, we define

$$\mathcal{V} = \{(\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{S}_+^d : \mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) \leq r\}$$

as the ball of radius r around $(\hat{\mu}, \hat{\Sigma})$ with respect to the Gelbrich distance. Note that $\mathcal{F} = f(\mathcal{V})$, where the transformation $f : \mathbb{R}^d \times \mathbb{S}_+^d \to \mathbb{R}^d \times \mathbb{S}_+^d$ is defined by $f(\mu, \Sigma) = (\mu, \Sigma + \mu\mu^\top)$. We will now prove that $\mathcal{V}$ is compact. As f is continuous and as compactness is preserved under continuous transformations, this will readily imply that $\mathcal{F}$ is compact. Clearly, $\mathcal{V}$ is closed because the Gelbrich distance is continuous. To show that $\mathcal{V}$ is also bounded, fix any $(\mu, \Sigma) \in \mathcal{V}$. By the definition

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

of the Gelbrich distance, we have $\|\mu - \hat{\mu}\| \leq r^2$. In addition, we find

$$\begin{aligned}
&\mathrm{Tr}\big((\hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2})^{1/2}\big) \\
&\quad = \max_{C\in\mathbb{R}^{d\times d}} \left\{ \mathrm{Tr}(C) \colon \begin{bmatrix} \Sigma & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \succeq 0 \right\} \\
&\quad \leq \max_{C\in\mathbb{R}^{d\times d}} \left\{ \mathrm{Tr}(C) \colon C_{ij}^2 \leq \Sigma_{ii}\hat{\Sigma}_{jj} \ \ \forall i,j \in [d] \right\} \\
&\quad \leq \sqrt{\mathrm{Tr}(\Sigma)\,\mathrm{Tr}(\hat{\Sigma})},
\end{aligned}$$

where the equality has been established in the proof of Proposition 2.2. The two inequalities follow from a relaxation of the linear matrix inequality, which exploits the observation that all second principal minors of a positive semidefinite matrix are non-negative, and from the Cauchy–Schwarz inequality. Thus Σ satisfies

$$\big(\mathrm{Tr}(\Sigma)^{1/2} - \mathrm{Tr}(\hat{\Sigma})^{1/2}\big)^2 \leq \mathrm{Tr}\big(\Sigma + \hat{\Sigma} - 2(\hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2})^{1/2}\big) \leq r^2,$$

where the second inequality holds because $(\mu, \Sigma) \in \mathcal{V}$. We may therefore conclude that $\mathrm{Tr}(\Sigma) \leq (r + \mathrm{Tr}(\hat{\Sigma})^{1/2})^2$, which in turn implies that $0 \preceq \Sigma \preceq (r + (\mathrm{Tr}(\hat{\Sigma}))^{1/2})^2 I_d$. In summary, we have shown that both μ and Σ belong to bounded sets. As $(\mu, \Sigma) \in \mathcal{V}$ was chosen arbitrarily, this proves that $\mathcal{V}$ is indeed bounded and thus compact. □

Proposition 2.2 shows that the uncertainty set $\mathcal{F}$ is convex. This is surprising because $\mathcal{F} = f(\mathcal{V})$, where the Gelbrich ball $\mathcal{V}$ in the space of mean–covariance pairs is convex thanks to Proposition 2.2 and where f is a *quadratic* bijection. Indeed, convexity is usually only preserved under *affine* transformations.

Gelbrich ambiguity sets were introduced by Nguyen *et al.* (2021) in the context of robust portfolio optimization. They have also found use in machine learning (Bui, Nguyen and Nguyen 2022, Vu, Tran, Yue and Nguyen 2022, Nguyen, Bui and Nguyen 2023*a*), estimation (Nguyen *et al.* 2023*b*), filtering (Shafieezadeh-Abadeh, Nguyen, Kuhn and Mohajerin Esfahani 2018, Kargin, Hajar, Malik and Hassibi 2024*b*) and control (McAllister and Mohajerin Esfahani 2024, Al Taha, Yan and Bitar 2023, Hajar, Kargin and Hassibi 2023, Hakobyan and Yang 2024, Taşkesen, Iancu, Koçyiğit and Kuhn 2024, Kargin, Hajar, Malik and Hassibi 2024*a*,*c*,*d*).

2.1.5. *Mean-dispersion ambiguity sets*

If $\mathcal{K} \subseteq \mathbb{R}^k$ is a proper convex cone and $v_1, v_2 \in \mathbb{R}^k$, then the inequality $v_1 \preceq_{\mathcal{K}} v_2$ means that $v_2 - v_1 \in \mathcal{K}$. Also, a function $G\colon \mathbb{R}^m \to \mathbb{R}^k$ is called $\mathcal{K}$-convex if

$$G(\theta v_1 + (1-\theta)v_2) \preceq_{\mathcal{K}} \theta G(v_1) + (1-\theta)G(v_2) \quad \text{for all } v_1, v_2 \in \mathbb{R}^m,\ \theta \in [0,1].$$

The mean-dispersion ambiguity set corresponding to a convex closed support set $\mathcal{Z} \subseteq \mathbb{R}^d$, a mean vector $\mu \in \mathbb{R}^d$, a $\mathcal{K}$-convex dispersion function $G\colon \mathbb{R}^m \to \mathbb{R}^k$ and a dispersion bound $g \in \mathbb{R}^k$ is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{E}_{\mathbb{P}}[Z] = \mu,\ \mathbb{E}_{\mathbb{P}}[G(Z)] \preceq_{\mathcal{K}} g\}. \tag{2.9}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The mean-dispersion ambiguity set is highly expressive, that is, it can model various stylized features of the unknown probability distribution. For example, if $\|\cdot\|$ is a norm on $\mathbb{R}^d$, $G(z) = \|z - \mu\|$ is convex in the usual sense, and $g = \sigma \in \mathbb{R}_+$, then all distributions $\mathbb{P} \in \mathcal{P}$ have a mean absolute deviation from the mean that is bounded by σ. Alternatively, if $G(z) = (z-\mu)(z-\mu)^\top$ is $\mathbb{S}_+^d$-convex and $g = \Sigma \in \mathbb{S}_+^d$, then $\mathcal{P}$ reduces to a Chebyshev ambiguity set with moment uncertainty. Specifically, the covariance matrix of any $\mathbb{P} \in \mathcal{P}$ is bounded by Σ in Loewner order. Wiesemann, Kuhn and Sim (2014) show that the ambiguity set $\mathcal{P}$, which contains distributions of the d-dimensional random vector Z, is closely related to the lifted ambiguity set

$$\mathcal{Q} = \{\mathbb{Q} \in \mathcal{P}(\mathcal{C})\colon \mathbb{E}_\mathbb{Q}[Z] = \mu,\ \mathbb{E}_\mathbb{Q}[U] = g\}$$

with support set $\mathcal{C} = \{(z,u) \in \mathcal{Z} \times \mathbb{R}^k\colon G(z) \preceq_\mathcal{K} u\}$, which contains *joint* distributions of Z and an auxiliary m-dimensional random vector U. Indeed, one can prove that $\mathcal{P} = \{\mathbb{Q}_Z\colon \mathbb{Q} \in \mathcal{Q}\}$, where $\mathbb{Q}_Z$ denotes the marginal distribution of Z under $\mathbb{Q}$. As the loss function depends only on Z but not on U, this reasoning implies that the inner worst-case expectation problem in (1.2) satisfies

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_\mathbb{P}[\ell(x,Z)] = \sup_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}_\mathbb{Q}[\ell(x,Z)].$$

Hence one can replace the original ambiguity set $\mathcal{P}$ with the lifted ambiguity set $\mathcal{Q}$. This is useful because $\mathcal{Q}$ constitutes a simple Markov ambiguity set that specifies only the support set $\mathcal{C}$ and the mean (μ, g) of the joint random vector (Z, U). In addition, one can show that $\mathcal{Z}$ is convex because $\mathcal{Z}$ is convex and G is $\mathcal{K}$-convex. In summary, DRO problems with mean-dispersion ambiguity sets of the form (2.9) can systematically be reduced to DRO problems with Markov ambiguity sets.

A more general class of mean-dispersion ambiguity sets can be used to shape the moment generating function of Z under $\mathbb{P}$. Specifically, Chen, Sim and Xu (2019) introduce the *entropic dominance* ambiguity set

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathbb{E}_\mathbb{P}[Z] = \mu,\ \log(\mathbb{E}_\mathbb{P}[\exp(\theta^\top(Z-\mu))]) \le g(\theta)\ \ \forall \theta \in \mathbb{R}^d\},$$

where $g\colon \mathbb{R}^d \to \mathbb{R}$ is a convex and twice continuously differentiable function satisfying $g(0) = 0$ and $\nabla g(0) = 0$. The constraints parametrized by θ impose a continuum of upper bounds on the cumulant generating function (i.e. the logarithmic moment generating function) of the centred random variable $Z - \mu$ under $\mathbb{P}$. The choice of g determines the specific class of distributions included in the ambiguity set. For example, if $g(\theta) = \sigma^2\theta^\top\theta/2$ for some $\sigma > 0$, then the ambiguity set contains only sub-Gaussian distributions with variance proxy σ^2. Sub-Gaussian distributions are probability distributions whose tails are bounded by the tails of a Gaussian distribution. They play a significant role in probability theory and statistics, particularly in the study of concentration inequalities and high-dimensional phenomena (Vershynin 2018, Wainwright 2019).

The entropic dominance ambiguity set imposes infinitely many constraints on $\mathbb{P}$. Chen *et al.* (2019) show that worst-case expectation problems over this ambiguity

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

set can be reformulated as semi-infinite conic programs. They propose a cutting-plane algorithm to solve these conic programs efficiently. The entropic dominance ambiguity set has also found applications in the study of nonlinear and PDE-constrained DRO problems (Milz and Ulbrich 2020, 2022). Generalized entropic dominance ambiguity sets are considered by Chen *et al.* (2024*a*).

2.1.6. Higher-order moment ambiguity sets

Markov and Chebyshev ambiguity sets only impose conditions on the first- and/or second-order moments of $\mathbb{P}$. DRO problems with such ambiguity sets are often tractable. In this section we briefly comment on moment ambiguity sets that impose conditions on higher-order (polynomial) moments of $\mathbb{P}$, which generically lead to NP-hard DRO problems (Popescu 2005, Propositions 4.5 and 4.6).

Assume now that $\mathcal{Z}$ is a closed semi-algebraic set defined as the feasible set of finitely many polynomial inequalities. In addition, define the monomial of order $\alpha \in \mathbb{Z}_+^d$ in $z \in \mathbb{R}^d$ as the function $\prod_{i=1}^d z_i^{\alpha_i}$, which we denote more compactly as z^α. The higher-order moment ambiguity set induced by a finite index set $\mathcal{A} \subseteq \mathbb{Z}_+^d$ and the moment bounds $m_\alpha \in \mathbb{R}$, $\alpha \in \mathcal{A}$, is then given by

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{E}_{\mathbb{P}}[Z^\alpha] \leq m_\alpha \ \ \forall \alpha \in \mathcal{A}\}.$$

Evaluating the worst-case expectation of a polynomial function (or the characteristic function of a semi-algebraic set) over all distributions in $\mathcal{P}$ thus amounts to solving a generalized moment problem. This moment problem as well as its dual constitute semi-infinite linear programs, which can be recast as finite-dimensional conic optimization problems over certain moment cones and the corresponding dual cones of non-negative polynomials (Karlin and Studden 1966, Zuluaga and Pena 2005). Even though NP-hard in general, these conic problems can be approximated by increasingly tight sequences of tractable semidefinite programs by using tools from polynomial optimization (Parrilo 2000, 2003, Lasserre 2001, 2009). This general technique gives rise to worst-case expectation bounds and generalized Chebyshev inequalities with respect to the ambiguity set $\mathcal{P}$ (Bertsimas and Sethuraman 2000, Lasserre 2002, Popescu 2005, Lasserre 2008). In addition, it leads to tight bounds on worst-case risk measures (Natarajan, Pachamanova and Sim 2009*a*).

2.2. ϕ-divergence ambiguity sets

The dissimilarity between two probability distributions is often quantified in terms of a ϕ-divergence, which is uniquely determined by an entropy function ϕ.

Definition 2.4 (Entropy functions). An entropy function $\phi \colon \mathbb{R} \to \overline{\mathbb{R}}$ is a lower semicontinuous convex function with $\phi(1) = 0$ and $\phi(s) = +\infty$ for all $s < 0$.

Note that any entropy function ϕ is continuous relative to its domain. In fact, this is true for any *univariate* convex lower semicontinuous function. We emphasize, however, that *multivariate* convex lower semicontinuous functions can have points

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

of discontinuity within their domains (Rockafellar and Wets 2009, Example 2.38). The notion of a ϕ-divergence relies on the perspective ϕ^π of the entropy function ϕ.

Definition 2.5 (ϕ-divergences (Csiszár 1963, 1967, Ali and Silvey 1966)). The (generalized) ϕ-divergence of $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ with respect to $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is given by

$$\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) = \int_{\mathcal{Z}} \phi^\pi \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z), \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z) \right) \mathrm{d}\rho(z),$$

where ϕ is an entropy function and $\rho \in \mathcal{M}_+(\mathcal{Z})$ is any dominating measure. This means that $\mathbb{P}$ and $\hat{\mathbb{P}}$ are absolutely continuous with respect to ρ, that is, $\mathbb{P}, \hat{\mathbb{P}} \ll \rho$.

By the definition of ϕ^π and our convention that $0\phi(s/0)$ should be interpreted as the recession function $\phi^\infty(s)$, $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}})$ can be recast as

$$\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) = \int_{\mathcal{Z}} \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z) \cdot \phi \left(\frac{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)}{\frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)} \right) \mathrm{d}\rho(z).$$

A dominating measure ρ always exists, but it must depend on $\mathbb{P}$ and $\hat{\mathbb{P}}$. For example, one may set $\rho = \mathbb{P} + \hat{\mathbb{P}}$. The absolute continuity conditions $\mathbb{P} \ll \rho$ and $\hat{\mathbb{P}} \ll \rho$ ensure that the Radon–Nikodym derivatives $\mathrm{d}\mathbb{P}/\mathrm{d}\rho$ and $\mathrm{d}\hat{\mathbb{P}}/\mathrm{d}\rho$ are well-defined, respectively. The following proposition derives a dual representation of a generic ϕ-divergence, which reveals that $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}})$ is in fact independent of the choice of ρ.

Proposition 2.6 (Dual representation of ϕ-divergences). We have

$$\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \phi^*(f(z))\, \mathrm{d}\hat{\mathbb{P}}(z),$$

where $\mathcal{F}$ denotes the family of all bounded Borel functions $f : \mathcal{Z} \to \mathrm{dom}(\phi^*)$.

Proof. As the entropy function $\phi(s)$ is proper, convex and lower semicontinuous on $\mathbb{R}$ and as $0\phi(s/0)$ is interpreted as the recession function $\phi^\infty(s)$, the perspective function $\phi^\pi(s,t) = t\phi(s/t)$ is proper, convex and lower semicontinuous on $\mathbb{R} \times \mathbb{R}_+$. By Rockafellar (1970, Theorem 12.2), $\phi^\pi(s,t)$ can therefore be expressed as the conjugate of its conjugate. Note that the conjugate of $\phi^\pi(s,t)$ satisfies

$$\begin{aligned}
(\phi^\pi)^*(f,g) &= \sup_{s \in \mathbb{R},\, t \in \mathbb{R}_+} fs + gt - t\phi(s/t) \\
&= \sup_{t \in \mathbb{R}_+} gt + t\phi^*(f) \\
&= \begin{cases} 0 & \text{if } f \in \mathrm{dom}(\phi^*) \text{ and } g + \phi^*(f) \leq 0, \\ +\infty & \text{otherwise,} \end{cases}
\end{aligned}$$

for all $f, g \in \mathbb{R}$. The second equality in the above expression follows from Rockafellar (1970, Theorem 16.1). As $\phi^\pi(s,t) = \sup_{f,g \in \mathbb{R}} sf + tg - (\phi^\pi)^*(f,g)$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

by virtue of Rockafellar (1970, Theorem 12.2), the ϕ-divergence is thus given by

$$\begin{aligned}\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}}) &= \int_{\mathcal{Z}} \sup_{f,g\in\mathbb{R}} \left\{ \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)\cdot f + \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\cdot g - (\phi^\pi)^*(f,g) \right\} \mathrm{d}\rho(z) \\ &= \int_{\mathcal{Z}} \sup_{f\in\mathrm{dom}(\phi^*)} \left\{ \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)\cdot f - \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\cdot \phi^*(f) \right\} \mathrm{d}\rho(z) \\ &= \sup_{f\in\mathcal{F}} \int_{\mathcal{Z}} \left\{ \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)\cdot f(z) - \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\cdot \phi^*(f(z)) \right\} \mathrm{d}\rho(z),\end{aligned}$$

where the second equality exploits our explicit formula for $(\phi^\pi)^*$ derived above, while the third equality follows from Rockafellar and Wets (2009, Theorem 14.60). This theorem applies because the function $h\colon\ \mathrm{dom}(\phi^*)\times\mathcal{Z}\to\mathbb{R}$ defined by

$$h(f,z) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)\cdot f - \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\cdot\phi^*(f)$$

is continuous in f and Borel-measurable in z, thus constituting a Carathéodory integrand in the sense of Rockafellar and Wets (2009, Example 14.29). The claim then follows immediately from the definition of Radon–Nikodym derivatives. □

Proposition 2.6 reveals that $\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}})$ is jointly convex in $\mathbb{P}$ and $\hat{\mathbb{P}}$. If $\phi(s)$ grows superlinearly with s, that is, if the asymptotic growth rate $\phi^\infty(1)$ is infinite, then $\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}})$ is finite if and only if $\mathrm{d}\mathbb{P}/\mathrm{d}\rho(z) = 0$ for ρ-almost all $z\in\mathcal{Z}$ with $\mathrm{d}\hat{\mathbb{P}}/\mathrm{d}\rho(z) = 0$. Put differently, $\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}})$ is finite if and only if $\mathbb{P}\ll\hat{\mathbb{P}}$. In this special case, the chain rule for Radon–Nikodym derivatives implies that

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho} \Big/ \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho} = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\hat{\mathbb{P}}}.$$

If $\phi^\infty(1) = \infty$, the ϕ-divergence thus admits the more common (but less general) representation

$$\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}}) = \begin{cases} \int_{\mathcal{Z}} \phi\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\hat{\mathbb{P}}}(z)\right) \mathrm{d}\hat{\mathbb{P}}(z) & \text{if } \mathbb{P}\ll\hat{\mathbb{P}}, \\ +\infty & \text{otherwise.}\end{cases}$$

We are now ready to define the ϕ-divergence ambiguity set as

$$\mathcal{P} = \{\mathbb{P}\in\mathcal{P}(\mathcal{Z})\colon \mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}})\le r\}. \tag{2.10}$$

This set contains all probability distributions $\mathbb{P}$ supported on $\mathcal{Z}$ whose ϕ-divergence with respect to some prescribed reference distribution $\hat{\mathbb{P}}$ is at most $r\ge 0$.

Remark 2.7 (Csiszár duals). The family of generalized ϕ-divergences (which may adopt finite values even if $\mathbb{P}\not\ll\hat{\mathbb{P}}$) is invariant under permutations of $\mathbb{P}$ and $\hat{\mathbb{P}}$. Formally, we have $\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}}) = \mathrm{D}_\psi(\hat{\mathbb{P}},\mathbb{P})$, where ψ denotes the Csiszár dual of ϕ defined by $\psi(s) = \phi^\pi(1,s) = s\phi(1/s)$ (Ben-Tal, Ben-Israel and Teboulle 1991, Lemma 2.3). One readily verifies that if ϕ is a valid entropy function in the sense

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Table 2.1. Examples of entropy functions and their Csiszár duals.

Divergence	$\phi(s)$ $(s \geq 0)$	$\psi(s)$ $(s \geq 0)$	$\phi^\infty(1)$	$\psi^\infty(1)$
Kullback–Leibler	$s\log(s) - s + 1$	$-\log(s) + s - 1$	∞	1
Likelihood	$-\log(s) + s - 1$	$s\log(s) - s + 1$	1	∞
Total variation	$\frac{1}{2}\lvert s-1\rvert$	$\frac{1}{2}\lvert s-1\rvert$	$\frac{1}{2}$	$\frac{1}{2}$
Pearson χ^2	$(s-1)^2$	$\frac{1}{s}(s-1)^2$	∞	1
Neyman χ^2	$\frac{1}{s}(s-1)^2$	$(s-1)^2$	1	∞
Cressie–Read for $\beta \in (0,1)$	$\frac{s^\beta - \beta s + \beta - 1}{\beta(\beta-1)}$	$\frac{s^{1-\beta} - \beta + \beta s - s}{\beta(\beta-1)}$	$\frac{1}{1-\beta}$	$\frac{1}{\beta}$
Cressie–Read for $\beta > 1$	$\frac{s^\beta - \beta s + \beta - 1}{\beta(\beta-1)}$	$\frac{s^{1-\beta} - \beta + \beta s - s}{\beta(\beta-1)}$	∞	$\frac{1}{\beta-1}$

of Definition 2.4, then ψ is also a valid entropy function. This relationship shows that, even though ϕ-divergences are generically asymmetric, we do not sacrifice generality by focusing on divergence ambiguity sets of the form (2.10), with the nominal distribution $\hat{\mathbb{P}}$ being the second argument of the divergence. From the discussion after Proposition 2.6 it is clear that if $\phi^\infty(1) = \infty$, then all distributions $\mathbb{P}$ in the ϕ-divergence ambiguity set (2.10) satisfy $\mathbb{P} \ll \hat{\mathbb{P}}$. Similarly, if the Csiszár dual of ϕ satisfies $\psi^\infty(1) = \infty$, then all distributions $\mathbb{P}$ in the ϕ-divergence ambiguity set satisfy $\hat{\mathbb{P}} \ll \mathbb{P}$. Table 2.1 lists common entropy functions and their Csiszár duals. We emphasize that the family of Cressie–Read divergences includes the (scaled) Pearson χ^2-divergence for $\beta = 2$, the Kullback–Leibler divergence for $\beta \to 1$ and the likelihood divergence for $\beta \to 0$ as special cases.

The DRO literature often focuses on the *restricted* ϕ-divergence ambiguity set

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{P} \ll \hat{\mathbb{P}},\ \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq r\} \tag{2.11}$$

introduced by Ben-Tal *et al.* (2013). Unlike the standard ϕ-divergence ambiguity set (2.10), it contains only distributions that are absolutely continuous with respect to the reference distribution $\hat{\mathbb{P}}$ even if $\phi^\infty(1) < \infty$. Ben-Tal *et al.* (2013) study DRO problems over restricted ϕ-divergence ambiguity sets under the assumption that the reference distribution $\hat{\mathbb{P}}$ is discrete. In this case, the absolute continuity constraint $\mathbb{P} \ll \hat{\mathbb{P}}$ ensures that the ambiguity set contains only discrete distributions supported on the atoms of $\hat{\mathbb{P}}$, and thus nature's worst-case expectation problem reduces to a finite convex program. Ben-Tal *et al.* (2013) further develop a duality theory for this problem class. Shapiro (2017) extends this duality theory to general reference distributions $\hat{\mathbb{P}}$ that are not necessarily discrete. Hu, Hong and So (2013) and Jiang and Guan (2016) show that any distributionally robust individual chance

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

constraint with respect to a restricted ϕ-divergence ambiguity set is equivalent to a classical chance constraint under the reference distribution $\hat{\mathbb{P}}$ but with a rescaled confidence level. A classification of various ϕ-divergences and an analysis of the structural properties of the corresponding ϕ-divergence ambiguity sets is provided by Bayraksan and Love (2015) under the assumption that $\mathcal{Z}$ is finite. Below we review popular instances of the standard and restricted ϕ-divergence ambiguity sets.

2.2.1. Kullback–Leibler ambiguity sets

The Kullback–Leibler divergence is the ϕ-divergence corresponding to the entropy function that satisfies $\phi(s) = s\log(s) - s + 1$ for all $s \geq 0$; see also Table 2.1. As $\phi^\infty(1) = +\infty$, it thus admits the following equivalent definition.

Definition 2.8 (Kullback–Leibler divergence). The Kullback–Leibler divergence of $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ with respect to $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is given by

$$\mathrm{KL}(\mathbb{P}, \hat{\mathbb{P}}) = \begin{cases} \displaystyle\int_{\mathcal{Z}} \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\hat{\mathbb{P}}}(z)\right) \mathrm{d}\mathbb{P}(z) & \text{if } \mathbb{P} \ll \hat{\mathbb{P}}, \\ +\infty & \text{otherwise.} \end{cases}$$

We now review a famous variational formula for the Kullback–Leibler divergence.

Proposition 2.9 (Donsker and Varadhan 1983). The Kullback–Leibler divergence of $\mathbb{P}$ with respect to $\hat{\mathbb{P}}$ satisfies

$$\mathrm{KL}(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \log\left(\int_{\mathcal{Z}} \mathrm{e}^{f(z)}\, \mathrm{d}\hat{\mathbb{P}}(z)\right), \tag{2.12}$$

where $\mathcal{F}$ denotes the family of all bounded Borel functions $f\colon \mathcal{Z} \to \mathbb{R}^d$.

Proof. The convex conjugate of the entropy function ϕ inducing the Kullback–Leibler divergence satisfies $\phi^*(t) = \exp(t) - 1$ with $\mathrm{dom}(\phi^*) = \mathbb{R}$. Thus the dual representation of generic ϕ-divergences established in Proposition 2.6 implies that

$$\mathrm{KL}(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \left(\mathrm{e}^{f(z)} - 1\right) \mathrm{d}\hat{\mathbb{P}}(z),$$

where $\mathcal{F}$ denotes the family of all bounded Borel functions $f\colon \mathcal{Z} \to \mathbb{R}$. Note that $\mathcal{F}$ is invariant under constant shifts. That is, if $f(z)$ is a bounded Borel function, then so is $f(z) + c$ for any constant $c \in \mathbb{R}$. Without loss of generality, we may thus optimize over both $f \in \mathcal{F}$ and $c \in \mathbb{R}$ in the above maximization problem to obtain

$$\mathrm{KL}(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \sup_{c \in \mathbb{R}} \int_{\mathcal{Z}} (f(z) + c)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \left(\mathrm{e}^{f(z)+c} - 1\right) \mathrm{d}\hat{\mathbb{P}}(z).$$

For any fixed $f \in \mathcal{F}$, the inner maximization problem over c is uniquely solved by

$$c^\star = -\log\left(\int_{\mathcal{Z}} \mathrm{e}^{f(z)}\, \mathrm{d}\hat{\mathbb{P}}(z)\right).$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Substituting this expression back into the objective function yields (2.12). □

Proposition 2.9 establishes a link between the Kullback–Leibler divergence and the entropic risk measure. This connection will become useful in Section 4.3.

The Kullback–Leibler ambiguity set of radius $r \geq 0$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is given by

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{KL}(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}. \tag{2.13}$$

As $\phi^\infty(1) = +\infty$, all distributions $\mathbb{P} \in \mathcal{P}$ are absolutely continuous with respect to $\hat{\mathbb{P}}$. Thus $\mathcal{P}$ coincides with the *restricted* Kullback–Leibler ambiguity set. El Ghaoui *et al.* (2003) derive a closed-form expression for the worst-case value-at-risk of a linear loss function when $\hat{\mathbb{P}}$ is a *Gaussian* distribution. Hu and Hong (2013) use similar techniques to show that any distributionally robust individual chance constraint with respect to a Kullback–Leibler ambiguity set is equivalent to a classical chance constraint with a rescaled confidence level. Calafiore (2007) studies worst-case mean-risk portfolio selection problems when $\hat{\mathbb{P}}$ is a *discrete* distribution. The Kullback–Leibler ambiguity set has also found applications in least-squares estimation (Levy and Nikoukhah 2004), hypothesis testing (Levy 2008, Gül and Zoubir 2017), filtering (Levy and Nikoukhah 2012, Zorzi 2016, 2017*a*,*b*), the theory of risk measures (Ahmadi-Javid 2012, Postek *et al.* 2016) and extreme value analysis (Blanchet, He and Murthy 2020), among many others.

2.2.2. *Likelihood ambiguity sets*

As the Kullback–Leibler divergence fails to be symmetric, it gives rise to two strictly different ambiguity sets. The Kullback–Leibler ambiguity set from Section 2.2.1 is obtained by fixing the *second* argument of the Kullback–Leibler divergence to the reference distribution $\hat{\mathbb{P}}$ and considering all distributions $\mathbb{P}$ with $\mathrm{KL}(\mathbb{P}, \hat{\mathbb{P}}) \leq r$. An alternative ambiguity set is obtained by using $\hat{\mathbb{P}}$ as the *first* argument and setting

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{KL}(\hat{\mathbb{P}}, \mathbb{P}) \leq r\}. \tag{2.14}$$

We refer to $\mathcal{P}$ as the likelihood ambiguity set centred at $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$. Indeed, the likelihood or Burg-entropy divergence of $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ with respect to $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is usually defined as the reverse Kullback–Leibler divergence $\mathrm{KL}(\hat{\mathbb{P}}, \mathbb{P})$. This terminology is based on the following intuition. If $\mathcal{Z}$ is a discrete set and

$$\hat{\mathbb{P}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{z}_i}$$

is the empirical distribution corresponding to N independent samples $\{\hat{z}_i\}_{i=1}^N$ from an unknown distribution on $\mathcal{Z}$, then it is natural to construct the family of all distributions on $\mathcal{Z}$ that make the observed data achieve a prescribed level of likelihood. This distribution family corresponds to a superlevel set of the likelihood function $\mathcal{L}(\mathbb{P}) = \prod_{i=1}^N \mathbb{P}(Z = \hat{z}_i)$ over $\mathcal{P}(\mathcal{Z})$. One can show that any such *superlevel*

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

set coincides with a *sublevel* set of the likelihood divergence $\mathrm{KL}(\hat{\mathbb{P}}, \mathbb{P})$. Thus it constitutes a likelihood ambiguity set of the form (2.14). We emphasize that this correspondence does not easily carry over to situations where $\mathcal{Z}$ fails to be discrete.

Likelihood ambiguity sets were originally introduced by Wang *et al.* (2016) in the context of static DRO, and they were used by Wiesemann, Kuhn and Rustem (2013) in the context of robust Markov decision processes. Bertsimas *et al.* (2018*a*,*b*) show that the likelihood ambiguity set contains all distributions that pass a G-test of goodness-of-fit at a prescribed significance level.

Likelihood ambiguity sets display several statistical optimality properties even if $\mathcal{Z}$ is uncountable. To explain these properties, we consider the task of evaluating a $(1-\eta)$-upper confidence bound on the expected value of some loss function under an unknown distribution $\mathbb{P}$ when N independent samples from $\mathbb{P}$ are given. Leveraging the empirical likelihood theorem by Owen (1988), Lam (2019) shows a desirable property of the likelihood ambiguity set centred around the empirical distribution $\hat{\mathbb{P}}$: The associated worst-case expected loss provides the least conservative confidence bound for a constant significance level η asymptotically when the radius r decays at the rate $1/N$. Similar guarantees for a broader class of ϕ-divergences are reported by Duchi, Glynn and Namkoong (2021). In addition, Van Parys, Mohajerin Esfahani and Kuhn (2021) leverage Sanov's large deviation principle (Cover and Thomas 2006, Theorem 11.4.1) to prove that the worst-case expected loss with respect to a likelihood ambiguity set of constant radius r around $\hat{\mathbb{P}}$ provides the least conservative confidence bound for a decaying significance level $\eta \propto \mathrm{e}^{-rN}$ asymptotically for large N. Gupta (2019) further shows that a likelihood ambiguity set of radius $r \propto N^{-1/2}$ around $\hat{\mathbb{P}}$ represents the smallest convex ambiguity set that satisfies a Bayesian robustness guarantee.

2.2.3. Total variation ambiguity sets

The total variation distance of two distributions $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is the maximum absolute difference between the probabilities assigned to any event by $\mathbb{P}$ and $\hat{\mathbb{P}}$.

Definition 2.10 (Total variation distance). The total variation distance is the function $\mathrm{TV}\colon \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, 1]$ defined by

$$\mathrm{TV}(\mathbb{P}, \hat{\mathbb{P}}) = \sup\{|\mathbb{P}(\mathcal{B}) - \hat{\mathbb{P}}(\mathcal{B})| \colon \mathcal{B} \subseteq \mathcal{Z} \text{ is a Borel set}\}.$$

The total variation distance is ostensibly symmetric and satisfies the identity of indiscernible as well as the triangle inequality. Thus it constitutes a metric on $\mathcal{P}(\mathcal{Z})$. In addition, the total variation distance is an instance of a ϕ-divergence.

Proposition 2.11. The total variation distance coincides with the ϕ-divergence induced by the the entropy function with $\phi(s) = \frac{1}{2}|s - 1|$ for all $s \geq 0$.

Proof. The conjugate entropy function evaluates to $\phi^*(t) = \max\{t, -\frac{1}{2}\}$ if $t \leq \frac{1}{2}$ and to $\phi^*(t) = +\infty$ if $t > \frac{1}{2}$. By Proposition 2.6, the ϕ-divergence corresponding to

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

the given entropy function thus admits the dual representation

$$\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}}) = \sup_{f\in\mathcal{F}} \int_{\mathcal{Z}} f(z)\,\mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \max\left\{f(z), -\frac{1}{2}\right\}\mathrm{d}\hat{\mathbb{P}}(z), \tag{2.15}$$

where $\mathcal{F}$ denotes the family of all bounded Borel functions $f\colon \mathcal{Z}\to(-\infty,\frac{1}{2}]$. As clipping any $f\in\mathcal{F}$ from below at $-\frac{1}{2}$ creates a new function in $\mathcal{F}$ with a non-inferior objective value, we can in fact restrict attention to Borel functions $f\colon \mathcal{Z}\to[-\frac{1}{2},\frac{1}{2}]$. The objective function in (2.15) then simplifies to

$$\int_{\mathcal{Z}} f(z)\,\mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} f(z)\,\mathrm{d}\hat{\mathbb{P}}(z).$$

This simplified objective function remains unchanged when f is shifted by a constant. In summary, we may therefore conclude that (2.15) is equivalent to

$$\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}}) = \sup_{f\in\mathcal{F}'} \int_{\mathcal{Z}} f(z)\,\mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} f(z)\,\mathrm{d}\hat{\mathbb{P}}(z), \tag{2.16}$$

where $\mathcal{F}'$ denotes the family of all Borel functions $f\colon \mathcal{Z}\to[0,1]$. Moreover, as the objective function of the maximization problem in (2.16) is linear in f, we can further restrict $\mathcal{F}'$ to contain only binary Borel functions $f\colon \mathcal{Z}\to\{0,1\}$ without sacrificing optimality. As there is a one-to-one correspondence between Borel sets and their characteristic functions, we finally obtain the desired identity

$$\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}}) = \sup\{|\mathbb{P}(\mathcal{B}) - \hat{\mathbb{P}}(\mathcal{B})|\colon \mathcal{B}\subseteq\mathcal{Z} \text{ is a Borel set}\}.$$

Hence the claim follows. □

The total variation ambiguity set of radius $r\geq 0$ around $\hat{\mathbb{P}}\in\mathcal{P}(\mathcal{Z})$ is given by

$$\mathcal{P} = \{\mathbb{P}\in\mathcal{P}(\mathcal{Z})\colon \mathrm{TV}(\mathbb{P},\hat{\mathbb{P}})\leq r\}.$$

Most of the existing literature focuses on the *restricted* total variation ambiguity set, which contains all distributions $\mathbb{P}\in\mathcal{P}$ that satisfy $\mathbb{P}\ll\hat{\mathbb{P}}$. Jiang and Guan (2018, Theorem 1) and Shapiro (2017, Example 3.7) show that the worst-case expected loss with respect to a restricted total variation ambiguity set coincides with a combination of a conditional value-at-risk and the essential supremum of the loss with respect to $\hat{\mathbb{P}}$; see also Section 6.10. Rahimian, Bayraksan and Homem-de-Mello (2019*a*,*b*, 2022) study the worst-case distributions of DRO problems over unrestricted total variation ambiguity sets when $\mathcal{Z}$ is finite. The total variation ambiguity set is related to Huber's contamination model from robust statistics (Huber 1981), which assumes that a fraction $r\in(0,1)$ of all samples in a statistical dataset are drawn from an arbitrary contaminating distribution. Hence the total variation distance between the target distribution to be estimated and the contaminated data-generating distribution is at most r. It is thus natural to use a total variation ambiguity set of radius r around some estimated distribution as the search space for the target distribution (Nishimura and Ozaki 2004, 2006, Bose and Daripa 2009, Duchi, Hashimoto and Namkoong 2023, Tsang and Shehadeh 2024).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

2.2.4. χ^2-divergence ambiguity set

The χ^2-divergence is the ϕ-divergence corresponding to the entropy function that satisfies $\phi(s) = (s-1)^2$ for all $s \geq 0$; see also Table 2.1. As $\phi^\infty(1) = +\infty$, it thus admits the following equivalent definition.

Definition 2.12 (χ^2-divergence). The χ^2-divergence of $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ with respect to $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is given by

$$\chi^2(\mathbb{P}, \hat{\mathbb{P}}) = \begin{cases} \displaystyle\int_{\mathcal{Z}} \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\hat{\mathbb{P}}}(z) - 1 \right)^2 \mathrm{d}\hat{\mathbb{P}}(z) & \text{if } \mathbb{P} \ll \hat{\mathbb{P}}, \\ +\infty & \text{otherwise.} \end{cases}$$

The χ^2-divergence admits the following dual representation.

Proposition 2.13. The χ^2-divergence of $\mathbb{P}$ with respect to $\hat{\mathbb{P}}$ satisfies

$$\chi^2(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \frac{(\mathbb{E}_{\mathbb{P}}[f(Z)] - \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)])^2}{\mathbb{V}_{\hat{\mathbb{P}}}[f(Z)]},$$

where $\mathcal{F}$ is shorthand for the family of all bounded Borel functions $f\colon \mathcal{Z} \to \mathbb{R}$, and $\mathbb{V}_{\hat{\mathbb{P}}}[f(Z)]$ stands for the variance of $f(Z)$ under $\hat{\mathbb{P}}$. If $\mathbb{V}_{\hat{\mathbb{P}}}[f(Z)] = 0$, then the above fraction is interpreted as 0 if $\mathbb{E}_{\mathbb{P}}[f(Z)] = \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)]$ and as $+\infty$ otherwise.

Proof. The convex conjugate of the entropy function inducing the χ^2-divergence satisfies $\phi^*(t) = t^2/4 + t$ if $t \geq -2$ and $\phi^*(t) = -1$ if $t < -2$, and its domain is given by $\mathrm{dom}(\phi^*) = \mathbb{R}$. Consequently, Proposition 2.6 implies that

$$\chi^2(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \left(\frac{f(z)^2}{4} + f(z) \right) \mathrm{d}\hat{\mathbb{P}}(z),$$

where $\mathcal{F}$ denotes the family of all bounded Borel functions $f\colon \mathcal{Z} \to \mathbb{R}$. Note that we have replaced $\phi^*(f(z))$ with $f(z)^2/4 + f(z)$ in the second integral. This may be done without loss of generality. Indeed, if the function $f(z)$ adopts values below -2, then it is (weakly) dominated by the function $f'(z) = \max\{f(z), -2\}$. Note also that $\mathcal{F}$ is invariant under constant shifts. That is, if $f(z)$ is a bounded Borel function, then so is $f(z) + c$ for any constant $c \in \mathbb{R}$. An elementary calculation reveals that, for any fixed $f \in \mathcal{F}$, the optimal shift is $c^\star = -\mathbb{E}_{\hat{\mathbb{P}}}[f(Z)]$. Hence we may replace $f(z)$ with $f(z) - \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)]$ in the above expression, which yields

$$\chi^2(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[f(Z)] - \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)] - \frac{\mathbb{V}_{\hat{\mathbb{P}}}[f(Z)]}{4}.$$

Note that the set $\mathcal{F}$ is also invariant under scaling. That is, if $f(z)$ is a bounded Borel function, then so is $cf(z)$ for any constant $c \in \mathbb{R}$. We may thus optimize

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

separately over $f \in \mathcal{F}$ and $c \in \mathbb{R}$ in the above maximization problem to obtain

$$\chi^2(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \sup_{c \in \mathbb{R}} (\mathbb{E}_{\mathbb{P}}[f(Z)] - \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)])c - \frac{\mathbb{V}_{\hat{\mathbb{P}}}[f(Z)]}{4} c^2$$
$$= \sup_{f \in \mathcal{F}} \frac{(\mathbb{E}_{\mathbb{P}}[f(Z)] - \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)])^2}{\mathbb{V}_{\hat{\mathbb{P}}}[f(Z)]}.$$

Note that the inner maximization problem over c simply evaluates the conjugate of the convex quadratic function $\mathbb{V}_{\hat{\mathbb{P}}}[f(Z)]c^2/4$ at $\mathbb{E}_{\mathbb{P}}[f(Z)] - \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)]$, which is available in closed form. Thus the claim follows. □

As the χ^2-divergence fails to be symmetric, it gives rise to two complementary ambiguity sets, which differ according to whether the reference distribution $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is used as the first or the second argument of the χ^2-divergence. Lam (2018) defines the *Pearson* χ^2-ambiguity set of radius $r \geq 0$ around $\hat{\mathbb{P}}$ as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \chi^2(\mathbb{P}, \hat{\mathbb{P}}) \leq r\} \tag{2.17}$$

in order to analyse operations and service systems with dependent data. Philpott, de Matos and Kapelevich (2018) develop a stochastic dual dynamic programming algorithm for solving distributionally robust multistage stochastic programs with a Pearson ambiguity set. In the context of static DRO, Duchi and Namkoong (2019) show that robustification with respect to a Pearson ambiguity set is closely related to variance regularization. Note that as $\phi^\infty(1) = +\infty$, the Pearson ambiguity set coincides with its *restricted* version, which contains only distributions $\mathbb{P} \ll \hat{\mathbb{P}}$.

Klabjan, Simchi-Levi and Song (2013) define the *Neyman* χ^2-ambiguity set as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \chi^2(\hat{\mathbb{P}}, \mathbb{P}) \leq r\}$$

in order to formulate robust lot-sizing problems. Hanasusanto and Kuhn (2013) use a Neyman ambiguity set with finite $\mathcal{Z}$ in the context of robust data-driven dynamic programming. Finally, Hanasusanto *et al.* (2015*a*) use the same ambiguity set to model the uncertainty in the mixture weights of multimodal demand distributions.

2.3. Optimal transport ambiguity sets

Optimal transport theory offers a natural way to quantify the difference between probability distributions and gives rise to a rich family of ambiguity sets. To explain this, we first introduce the notion of a transportation cost function.

Definition 2.14 (Transportation cost function). A lower semicontinuous function $c \colon \mathcal{Z} \times \mathcal{Z} \to [0, +\infty]$ with $c(z, z) = 0$ for all $z \in \mathcal{Z}$ is a transportation cost function.

Every transportation cost function induces an optimal transport discrepancy.

Definition 2.15 (Optimal transport discrepancy). The optimal transport discrepancy $\mathrm{OT}_c \colon \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, +\infty]$ associated with any given transportation

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

cost function c is defined by

$$\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) = \inf_{\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})} \mathbb{E}_\gamma[c(Z, \hat{Z})], \tag{2.18}$$

where $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ represents the set of all couplings γ of $\mathbb{P}$ and $\hat{\mathbb{P}}$, that is, all joint probability distributions of Z and $\hat{Z}$ with marginals $\mathbb{P}$ and $\hat{\mathbb{P}}$, respectively.

By definition, we have $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$ if and only if $\gamma((Z, \hat{Z}) \in \mathcal{B} \times \mathcal{Z}) = \mathbb{P}(Z \in \mathcal{B})$ and $\gamma((Z, \hat{Z}) \in \mathcal{Z} \times \hat{\mathcal{B}}) = \hat{\mathbb{P}}(\hat{Z} \in \hat{\mathcal{B}})$ for all Borel sets $\mathcal{B}, \hat{\mathcal{B}} \subseteq \mathcal{Z}$. If the probability distributions $\mathbb{P}$ and $\hat{\mathbb{P}}$ are visualized as two piles of sand, then any coupling $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$ can be interpreted as a transportation plan, that is, an instruction for morphing $\hat{\mathbb{P}}$ into the shape of $\mathbb{P}$ by moving sand between various origin–destination pairs in $\mathcal{Z}$. Indeed, for any fixed origin $\hat{z} \in \mathcal{Z}$, the conditional probability $\gamma(z \leq Z \leq z + \mathrm{d}z \mid \hat{Z} = \hat{z})$ determines the proportion of the sand located at $\hat{z}$ that should be moved to (an infinitesimally small rectangle at) the destination z. If the cost of moving one unit of probability mass from $\hat{z}$ to z amounts to $c(z, \hat{z})$, then $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}})$ is the minimal amount of money that is needed to morph $\hat{\mathbb{P}}$ into $\mathbb{P}$. We now provide a dual representation for generic optimal transport discrepancies.

Proposition 2.16 (Kantorovich duality I). We have

$$\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) = \begin{cases} \sup\limits_{f \in \mathcal{L}^1(\mathbb{P}),\, g \in \mathcal{L}^1(\hat{\mathbb{P}})} & \displaystyle\int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} g(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z}) \\ \text{s.t.} & f(z) - g(\hat{z}) \leq c(z, \hat{z}) \;\; \forall z, \hat{z} \in \mathcal{Z}, \end{cases} \tag{2.19}$$

where $\mathcal{L}^1(\mathbb{P})$ and $\mathcal{L}^1(\hat{\mathbb{P}})$ denote the sets of all Borel functions from $\mathcal{Z}$ to $\mathbb{R}$ that are integrable with respect to $\mathbb{P}$ and $\hat{\mathbb{P}}$, respectively.

The dual problem (2.19) represents the profit maximization problem of a third party that redistributes the sand from $\hat{\mathbb{P}}$ to $\mathbb{P}$ on behalf of the problem owner by buying sand at the origin $\hat{z}$ at unit price $g(\hat{z})$ and selling sand at the destination z at unit price $f(z)$. The constraints ensure that it is cheaper for the problem owner to use the services of the third party instead of moving the sand without external help at the transportation cost $c(z, \hat{z})$ for every origin–destination pair $(\hat{z}, z)$. The optimal price functions $f^\star$ and $g^\star$, if they exist, are termed Kantorovich potentials.

Proof of Proposition 2.16. For a general proof we refer to Villani (2008, Theorem 5.10 (i)). We prove the claim under the simplifying assumption that $\mathcal{Z}$ is compact. In this case, the family $\mathcal{C}(\mathcal{Z} \times \mathcal{Z})$ of all continuous (and thus bounded) functions $f\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ equipped with the supremum norm constitutes a Banach space. Its topological dual is the space $\mathcal{M}(\mathcal{Z} \times \mathcal{Z})$ of all finite signed Borel measures on $\mathcal{Z} \times \mathcal{Z}$ equipped with the total variation norm (Folland 1999, Corollary 7.18). This means that for every continuous linear functional $\varphi\colon \mathcal{C}(\mathcal{Z} \times \mathcal{Z}) \to \mathbb{R}$ there exists $\gamma \in \mathcal{M}(\mathcal{Z} \times \mathcal{Z})$ such that

$$\varphi(f) = \int_{\mathcal{Z} \times \mathcal{Z}} f(z, \hat{z})\, \mathrm{d}\gamma(z, \hat{z}) \quad \text{for all } f \in \mathcal{C}(\mathcal{Z} \times \mathcal{Z}).$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

We first use the Fenchel–Rockafellar duality theorem to show that

$$\mathrm{OT}_c(\mathbb{P},\hat{\mathbb{P}}) = \begin{cases} \sup\limits_{f,g\in\mathcal{C}(\mathcal{Z})} & \displaystyle\int_{\mathcal{Z}} f(z)\,\mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} g(\hat{z})\,\mathrm{d}\hat{\mathbb{P}}(\hat{z}) \\ \text{s.t.} & f(z) - g(\hat{z}) \le c(z,\hat{z}) \;\; \forall z,\hat{z}\in\mathcal{Z}, \end{cases} \tag{2.20}$$

that is, we prove that strong duality holds if the price functions f and g in the dual problem are restricted to the space $\mathcal{C}(\mathcal{Z})$ of continuous functions from $\mathcal{Z}$ to $\mathbb{R}$. To this end, we re-express the maximization problem in (2.20) more compactly as

$$\sup_{h\in\mathcal{C}(\mathcal{Z}\times\mathcal{Z})} -\phi(h) - \psi(h), \tag{2.21}$$

where the convex functions $\phi,\psi : \mathcal{C}(\mathcal{Z}\times\mathcal{Z}) \to (-\infty,+\infty]$ are defined by

$$\phi(h) = \begin{cases} 0 & \text{if } -h(z,\hat{z}) \le c(z,\hat{z}) \;\; \forall z,\hat{z}\in\mathcal{Z}, \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$\psi(h) = \begin{cases} \displaystyle\int_{\mathcal{Z}}\int_{\mathcal{Z}} h(z,\hat{z})\,\mathrm{d}\mathbb{P}(z)\,\mathrm{d}\hat{\mathbb{P}}(\hat{z}) & \begin{cases}\text{if } \exists f,g\in\mathcal{C}(\mathcal{Z}) \text{ with} \\ h(z,\hat{z}) = g(\hat{z}) - f(z) \;\; \forall z,\hat{z}\in\mathcal{Z}, \end{cases} \\ +\infty & \text{otherwise.} \end{cases}$$

Note that (2.21) can be viewed as the conjugate of $\phi+\psi$ with respect to the pairing of $\mathcal{C}(\mathcal{Z}\times\mathcal{Z})$ and $\mathcal{M}(\mathcal{Z}\times\mathcal{Z})$ evaluated at the zero measure. Note also that ϕ is continuous at the constant function $h_0 \equiv 1$ because the transportation cost function c is non-negative. In addition, h_0 belongs to the domain of ψ. The Fenchel–Rockafellar duality theorem (Brezis 2011, Theorem 1.12) thus ensures that the conjugate of the sum of the proper convex functions ϕ and ψ coincides with the infimal convolution of their conjugates ϕ^* and ψ^*. Hence (2.21) equals

$$(\phi+\psi)^*(0) = \inf_{\gamma\in\mathcal{M}(\mathcal{Z}\times\mathcal{Z})} \phi^*(-\gamma) + \psi^*(\gamma). \tag{2.22}$$

It remains to evaluate the conjugates of ϕ and ψ. For any $\gamma\in\mathcal{M}(\mathcal{Z}\times\mathcal{Z})$ we have

$$\begin{aligned} \phi^*(-\gamma) &= \sup_{h\in\mathcal{C}(\mathcal{Z}\times\mathcal{Z})} \left\{ -\int_{\mathcal{Z}\times\mathcal{Z}} h(z,\hat{z})\,\mathrm{d}\gamma(z,\hat{z}) \colon -h(z,\hat{z}) \le c(z,\hat{z}) \;\; \forall z,\hat{z}\in\mathcal{Z} \right\} \\ &= \begin{cases} \displaystyle\int_{\mathcal{Z}\times\mathcal{Z}} c(z,\hat{z})\,\mathrm{d}\gamma(z,\hat{z}) & \text{if } \gamma\in\mathcal{M}_+(\mathcal{Z}\times\mathcal{Z}), \\ +\infty & \text{otherwise,} \end{cases} \end{aligned}$$

where $\mathcal{M}_+(\mathcal{Z}\times\mathcal{Z})$ stands for the cone of finite Borel measures on $\mathcal{Z}\times\mathcal{Z}$. Indeed, if $\gamma\in\mathcal{M}_+(\mathcal{Z}\times\mathcal{Z})$, then the second equality follows from the monotone convergence theorem, which applies because c is lower semicontinuous and can thus be written as the pointwise limit of a non-decreasing sequence of continuous functions (see also Lemma 3.1 below). On the other hand, if $\gamma\notin\mathcal{M}_+(\mathcal{Z}\times\mathcal{Z})$, then the second equality holds because every $\gamma\in\mathcal{M}(\mathcal{Z}\times\mathcal{Z})$ is a Radon measure, which ensures that

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

the measure of any Borel set can be approximated with the integral of a continuous function. Similarly, for any $\gamma \in \mathcal{M}(\mathcal{Z} \times \mathcal{Z})$ one readily verifies that

$$\psi^*(\gamma) = \begin{cases} 0 & \text{if } \gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}}), \\ +\infty & \text{otherwise.} \end{cases}$$

Substituting the above formulas for ϕ^* and ψ^* into (2.22) yields (2.20).

Relaxing the requirement $f, g \in \mathcal{C}(\mathcal{Z})$ to $f \in \mathcal{L}^1(\mathbb{P})$ and $g \in \mathcal{L}^1(\hat{\mathbb{P}})$ on the right-hand side of (2.20) immediately leads to the upper bound

$$\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \leq \begin{cases} \sup\limits_{f \in \mathcal{L}^1(\mathbb{P}),\, g \in \mathcal{L}^1(\hat{\mathbb{P}})} & \displaystyle\int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} g(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z}) \\ \text{s.t.} & f(z) - g(\hat{z}) \leq c(z, \hat{z}) \;\; \forall z, \hat{z} \in \mathcal{Z}. \end{cases} \tag{2.23}$$

On the other hand, it is clear that

$$\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) = \inf_{\gamma \in \mathcal{M}_+(\mathcal{Z} \times \mathcal{Z})} \sup_{f \in \mathcal{L}^1(\mathbb{P}), g \in \mathcal{L}^1(\hat{\mathbb{P}})} \int_{\mathcal{Z} \times \mathcal{Z}} (c(z, \hat{z}) - f(z) + g(\hat{z}))\, \mathrm{d}\gamma(z, \hat{z}) + \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} g(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z}).$$

Interchanging the order of minimization and maximization in the above expression and then evaluating the inner infimum in closed form yields

$$\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \geq \begin{cases} \sup\limits_{f \in \mathcal{L}^1(\mathbb{P}),\, g \in \mathcal{L}^1(\hat{\mathbb{P}})} & \displaystyle\int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} g(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z}) \\ \text{s.t.} & f(z) - g(\hat{z}) \leq c(z, \hat{z}) \;\; \forall z, \hat{z} \in \mathcal{Z}. \end{cases} \tag{2.24}$$

Combining (2.23) with (2.24) proves (2.19), and thus the claim follows. □

The dual optimal transport problem (2.19) constitutes a linear program over the price functions $f \in \mathcal{L}^1(\mathbb{P})$ and $g \in \mathcal{L}^1(\hat{\mathbb{P}})$, and its objective function is linear in $\mathbb{P}$ and $\hat{\mathbb{P}}$. As pointwise suprema of linear functions are convex, $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}})$ is thus jointly convex in $\mathbb{P}$ and $\hat{\mathbb{P}}$. Problem (2.19) can be further simplified by invoking the c-transform $f^c : \mathcal{Z} \to (-\infty, +\infty]$ of the price function f, which is defined by

$$f^c(\hat{z}) = \sup_{z \in \mathcal{Z}} f(z) - c(z, \hat{z}). \tag{2.25}$$

The constraints of the dual problem (2.19) can now be re-expressed as

$$g(\hat{z}) \geq f(z) - c(z, \hat{z}) \quad \text{for all } z, \hat{z} \in \mathcal{Z} \qquad \Longleftrightarrow \qquad g(\hat{z}) \geq f^c(\hat{z}) \quad \text{for all } \hat{z} \in \mathcal{Z}.$$

Note that problem (2.19) seeks a price function g that is as *small* as possible. As g is lower-bounded by f^c, this suggests that $g = f^c$ at optimality. Conversely, defining the c-transform $g^c : \mathcal{Z} \to [-\infty, +\infty)$ of the price function g through

$$g^c(z) = \inf_{\hat{z} \in \mathcal{Z}} g(\hat{z}) + c(z, \hat{z}), \tag{2.26}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

the constraint of problem (2.19) can be re-expressed as

$$f(z) \leq g(\hat{z}) + c(z, \hat{z}) \quad \text{for all } z, \hat{z} \in \mathcal{Z} \qquad \Longleftrightarrow \qquad f(z) \leq g^c(z) \quad \text{for all } z \in \mathcal{Z}.$$

This suggests that $f = g^c$ at optimality. Note that f^c and g^c may fail to be integrable with respect to $\hat{\mathbb{P}}$ and $\mathbb{P}$, respectively. If $f \in \mathcal{L}^1(\mathbb{P})$ and $g \in \mathcal{L}^1(\hat{\mathbb{P}})$, however, then one can verify that the integrals $\int_{\mathcal{Z}} f^c(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z}) < +\infty$ and $\int_{\mathcal{Z}} g^c(z)\, \mathrm{d}\mathbb{P}(z) > -\infty$ exist as extended real numbers. The above insights culminate in the following corollary, which we state without proof. For details see Villani (2008, Theorem 5.10 (i)).

Corollary 2.17 (Kantorovich duality II). We have

$$\begin{aligned} \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) &= \sup_{f \in \mathcal{L}^1(\mathbb{P})} \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} f^c(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z}) \\ &= \sup_{g \in \mathcal{L}^1(\hat{\mathbb{P}})} \int_{\mathcal{Z}} g^c(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} g(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z}), \end{aligned}$$

where the c-transforms f^c and g^c are defined in (2.25) and (2.26), respectively. In addition, the first (second) supremum does not change if we require that $f = g^c$ ($g = f^c$) for some function $g\colon \mathcal{Z} \to (-\infty, +\infty]$ ($f\colon \mathcal{Z} \to [-\infty, +\infty)$).

Given any transportation cost function c, reference distribution $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ and transportation budget $r \geq 0$, the optimal transport ambiguity set is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}. \tag{2.27}$$

By construction, $\mathcal{P}$ contains all probability distributions $\mathbb{P}$ that can be obtained by reshaping the reference distribution $\hat{\mathbb{P}}$ at a finite cost of at most $r \geq 0$. The optimal transport ambiguity set was first studied by Pflug and Wozabal (2007), who propose a successive linear programming algorithm to solve robust mean-risk portfolio selection problems when $\mathcal{Z}$ is *finite*. Postek *et al.* (2016) leverage tools from conjugate duality theory to develop an exact solution method for the same problem class. Wozabal (2012) and Pflug and Pichler (2014, § 7.1) reformulate DRO problems with optimal transport ambiguity sets over *uncountable* support sets $\mathcal{Z} \subseteq \mathbb{R}^d$ as finite-dimensional *non*-convex programs and address them with methods from global optimization. Mohajerin Esfahani and Kuhn (2018) and Zhao and Guan (2018) use specialized duality results to show that these DRO problems are in fact equivalent to generalized moment problems that admit exact reformulations as finite-dimensional convex programs. Blanchet and Murthy (2019), Gao and Kleywegt (2023) and Zhang *et al.* (2024*b*) show that the underlying duality results remain valid even when $\mathcal{Z}$ is a Polish space. For recent surveys of the theory and applications of DRO with optimal transport ambiguity sets we refer to Kuhn *et al.* (2019) and Blanchet, Murthy and Nguyen (2021).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

2.3.1. p-Wasserstein ambiguity sets

It is common to set the transportation cost function c in Definition 2.15 to the pth power of some metric on $\mathcal{Z}$. In this case, the pth root of the optimal transport discrepancy is termed the p-Wasserstein distance.

Definition 2.18 (p-Wasserstein distance). Assume that $d(\cdot,\cdot)$ is a metric on $\mathcal{Z}$ and $p \in [1,+\infty)$ is a prescribed exponent. Then the p-Wasserstein distance $\mathrm{W}_p : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0,+\infty]$ corresponding to d and p is defined via

$$\mathrm{W}_p(\mathbb{P},\hat{\mathbb{P}}) = \inf_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} (\mathbb{E}_\gamma[d(Z,\hat{Z})^p])^{1/p}.$$

Definition 2.18 implies that if $c(z,\hat{z}) = d(z,\hat{z})^p$, then $W_p^p(\mathbb{P},\hat{\mathbb{P}}) = \mathrm{OT}_c(\mathbb{P},\hat{\mathbb{P}})$. In the following we use $\mathcal{P}_p(\mathcal{Z}) = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{E}_\mathbb{P}[d(Z,\hat{z}_0)^p] < \infty\}$ to denote the family of all distributions on $\mathcal{Z}$ with finite pth moment. As d is a metric, $\mathcal{P}_p(\mathcal{Z})$ is independent of the choice of the reference point $\hat{z}_0 \in \mathcal{Z}$. The p-Wasserstein distance constitutes a metric on $\mathcal{P}_p(\mathcal{Z})$. Indeed, it is evident that $W_p(\mathbb{P},\hat{\mathbb{P}})$ is symmetric and vanishes if and only if $\mathbb{P} = \hat{\mathbb{P}}$. The proof that $W_p(\mathbb{P},\hat{\mathbb{P}})$ obeys the triangle inequality requires a glueing lemma for transportation plans and is therefore more intricate; see e.g. Villani (2008, § 1). The p-Wasserstein distance further metrizes the weak convergence of distributions *and* the convergence of their pth moments. This means that $W_p(\mathbb{P},\hat{\mathbb{P}}_N)$ converges to 0 if and only if $\hat{\mathbb{P}}_N$ converges weakly to $\mathbb{P}$ *and* $\mathbb{E}_{\hat{\mathbb{P}}_N}[d(Z,\hat{z}_0)^p]$ converges to $\mathbb{E}_\mathbb{P}[d(Z,\hat{z}_0)^p]$ as N grows (Villani 2008, Theorem 6.9). Furthermore, the p-Wasserstein distance enjoys attractive measure concentration properties. Specifically, if $\hat{\mathbb{P}}_N$ represents the empirical distribution obtained from N independent samples from $\mathbb{P}$, then the rate at which $\hat{\mathbb{P}}_N$ converges to $\mathbb{P}$ in p-Wasserstein distance admits sharp asymptotic and finite-sample bounds (Fournier and Guillin 2015, Weed and Bach 2019).

As the p-Wasserstein distance constitutes the pth root of an optimal transport discrepancy, Proposition 2.16 and Corollary 2.17 readily imply that it admits a dual representation. For $p = 1$ this dual representation becomes particularly simple. Indeed, one can show that the 1-Wasserstein distance coincides with the integral probability metric generated by all test functions that are Lipschitz-continuous with respect to the metric d and have Lipschitz modulus at most 1.

Corollary 2.19 (Kantorovich–Rubinstein duality). We have

$$\mathrm{W}_1(\mathbb{P},\hat{\mathbb{P}}) = \sup_{f\in\mathcal{L}^1(\mathbb{P}),\,\mathrm{lip}(f)\le 1} \int_\mathcal{Z} f(z)\,\mathrm{d}\mathbb{P}(z) - \int_\mathcal{Z} f(\hat{z})\,\mathrm{d}\hat{\mathbb{P}}(\hat{z}).$$

Proof. Corollary 2.17 implies that

$$\mathrm{W}_1(\mathbb{P},\hat{\mathbb{P}}) = \sup_{f\in\mathcal{L}^1(\mathbb{P})} \int_\mathcal{Z} f(z)\,\mathrm{d}\mathbb{P}(z) - \int_\mathcal{Z} f^c(\hat{z})\,\mathrm{d}\hat{\mathbb{P}}(\hat{z}).$$

In addition, it ensures that the supremum does not change if we restrict the search space to functions that are representable as $f = g^c$ for some $g \colon \mathcal{Z} \to (-\infty,+\infty]$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

By (2.26), we thus have $f(z) = \inf_{\hat{z}\in\mathcal{Z}} g(\hat{z}) + d(z,\hat{z})$. For any fixed $\hat{z} \in \mathcal{Z}$, the auxiliary function $f_{\hat{z}}(z) = g(\hat{z}) + d(z,\hat{z})$ is ostensibly 1-Lipschitz with respect to the metric d. As infima of 1-Lipschitz functions remain 1-Lipschitz, we thus find $\mathrm{lip}(f) \leq 1$. In summary, we have shown that restricting attention to 1-Lipschitz functions does not reduce the supremum of the dual optimal transport problem. Next, we prove that $\mathrm{lip}(f) \leq 1$ implies that $f^c = f$. Indeed, for any $\hat{z} \in \mathcal{Z}$ we have

$$f(\hat{z}) \leq \sup_{z\in\mathcal{Z}} f(z) - d(z,\hat{z}) \leq \sup_{z\in\mathcal{Z}} f(\hat{z}) + d(z,\hat{z}) - d(z,\hat{z}) = f(\hat{z}),$$

where the two inequalities hold because $d(\hat{z},\hat{z}) = 0$ and $\mathrm{lip}(f) \leq 1$, respectively. This implies via (2.25) that $f(\hat{z}) = \sup_{z\in\mathcal{Z}} f(z) - d(z,\hat{z}) = f^c(\hat{z})$ for all $\hat{z} \in \mathcal{Z}$. Hence f^c coincides with f whenever $\mathrm{lip}(f) \leq 1$, and thus the claim follows. □

The p-Wasserstein ambiguity set of radius $r \geq 0$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{W}_p(\mathbb{P},\hat{\mathbb{P}}) \leq r\}. \tag{2.28}$$

Pflug, Pichler and Wozabal (2012) study robust portfolio selection problems, where the uncertainty about the asset return distribution is captured by a p-Wasserstein ball. They prove that – as r approaches infinity – it becomes optimal to distribute one's capital equally among all available assets. Hence this result reveals that the popular $1/N$-investment strategy widely used in practice (DeMiguel, Garlappi and Uppal 2009) is optimal under extreme ambiguity. Pflug *et al.* (2012), Pichler (2013) and Wozabal (2014) further show that, for a broad range of convex risk measures, the worst-case portfolio risk across all distributions in a p-Wasserstein ball equals the nominal risk under $\hat{\mathbb{P}}$ plus a regularization term that scales with the Wasserstein radius r; see also Section 8.3.

The Wasserstein ambiguity set corresponding to $p = 1$ enjoys particular prominence in DRO. The Kantorovich–Rubinstein duality can be used to construct a simple upper bound on the worst-case expectation of a Lipschitz-continuous loss function across all distributions in a 1-Wasserstein ball. This upper bound is given by the sum of the expected loss under the nominal distribution $\hat{\mathbb{P}}$ plus a regularization term that consists of the Lipschitz modulus of the loss function weighted by the radius r of the ambiguity set. Shafieezadeh-Abadeh, Mohajerin Esfahani and Kuhn (2015) demonstrate that this upper bound is exact for distributionally robust logistic regression problems. However, this exactness result extends in fact to many linear prediction models with convex (Chen and Paschalidis 2018, 2019, Blanchet, Kang and Murthy 2019*b*, Shafieezadeh-Abadeh *et al.* 2019, Wu, Li and Mao 2022) and even non-convex loss functions (Gao *et al.* 2024*b*, Ho-Nguyen and Wright 2023). More generally, 1-Wasserstein ambiguity sets have found numerous applications in diverse areas such as two-stage and multi-stage stochastic programming (Zhao and Guan 2018, Hanasusanto and Kuhn 2018, Duque and Morton 2020, Bertsimas, Shtern and Sturt 2023), chance-constrained programming (Chen, Kuhn and Wiesemann 2024*c*, Xie 2021, Ho-Nguyen, Kılınç-Karzan, Küçükyavuz and Lee 2022, Shen and Jiang 2023), inverse optimization (Mohajerin Esfahani,

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Shafieezadeh-Abadeh, Hanasusanto and Kuhn 2018), statistical learning (Blanchet, Glynn, Yan and Zhou 2019*a*, Zhu *et al.* 2022*b*), hypothesis testing (Gao, Xie, Xie and Xu 2018), contextual stochastic optimization (Zhang, Yang and Gao 2024*a*), transportation (Sun, Xie and Witten 2023), control (Cherukuri and Cortés 2019, Yang 2020, Boskos, Cortés and Martínez 2020, Li and Martínez 2020, Coulson, Lygeros and Dörfler 2021, Aolaritei, Lanzetti, Chen and Dörfler 2022*a*, Terpin *et al.* 2022, Terpin, Lanzetti and Dörfler 2024) and power systems analysis (Wang *et al.* 2018, Ordoudis, Nguyen, Kuhn and Pinson 2021), among others.

The Wasserstein ambiguity set corresponding to $p = 2$ also enjoys wide popularity. Before reviewing its various uses, we highlight an interesting connection between the 2-Wasserstein distance and the Gelbrich distance introduced in Section 2.1.4 (see Definition 2.1). As pointed out by Gelbrich (1990, Theorem 2.1), the 2-Wasserstein distance between two probability distributions provides an upper bound on the Gelbrich distance between their mean–covariance pairs.

Theorem 2.20 (Gelbrich bound). Assume that $\mathcal{Z}$ is equipped with the Euclidean metric $d(z, \hat{z}) = \|z - \hat{z}\|_2$. For any distributions $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ with finite mean vectors $\mu, \hat{\mu} \in \mathbb{R}^d$ and covariance matrices $\Sigma, \hat{\Sigma} \in \mathbb{S}_+^d$, respectively, we have

$$\mathrm{W}_2(\mathbb{P}, \hat{\mathbb{P}}) \geq \mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})).$$

Proof. By definition, the squared 2-Wasserstein distance satisfies

$$\begin{aligned}
\mathrm{W}_2^2(\mathbb{P}, \hat{\mathbb{P}}) &= \inf_{\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})} \int_{\mathcal{Z} \times \mathcal{Z}} \|z - \hat{z}\|_2^2 \, \mathrm{d}\gamma(z, \hat{z}) \\
&= \begin{cases} \inf & \|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}[\Sigma + \hat{\Sigma} - 2C] \\ \text{s.t.} & \gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}}), \ C \in \mathbb{R}^{d \times d} \\ & \displaystyle\int_{\mathcal{Z} \times \mathcal{Z}} \begin{bmatrix} z - \mu \\ \hat{z} - \hat{\mu} \end{bmatrix} \begin{bmatrix} z - \mu \\ \hat{z} - \hat{\mu} \end{bmatrix}^\top \mathrm{d}\gamma(z, \hat{z}) = \begin{bmatrix} \Sigma & C \\ C^\top & \hat{\Sigma} \end{bmatrix}, \ \begin{bmatrix} \Sigma & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \succeq 0. \end{cases}
\end{aligned}$$

Note that the new decision variable C is uniquely determined by the transportation plan γ, that is, it represents the cross-covariance matrix of Z and $\hat{Z}$ under γ. Thus its presence does not enlarge the feasible set. Note also that the linear matrix inequality in the last expression is redundant because the second-order moment matrix of γ is necessarily positive semidefinite. Thus its presence does not reduce the feasible set. Finally, note that the integral of the quadratic function

$$\begin{aligned}
\|z - \hat{z}\|_2^2 = \|\mu - \hat{\mu}\|_2^2 + \|z - \mu\|_2^2 + \|\hat{z} - \hat{\mu}\|_2^2 - 2(z - \mu)^\top (\hat{z} - \hat{\mu}) \\
+ 2(\mu - \hat{\mu})^\top (z - \mu) - 2(\mu - \hat{\mu})^\top (\hat{z} - \hat{\mu})
\end{aligned}$$

with respect to γ is uniquely determined by the first- and second-order moments of γ and evaluates to $\|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}[\Sigma + \hat{\Sigma} - 2C]$. Relaxing the last optimization

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

problem by removing all constraints that involve γ then yields

$$\mathrm{W}_2^2(\mathbb{P},\hat{\mathbb{P}}) \geq \begin{cases} \min\limits_{C\in\mathbb{R}^{d\times d}} & \|\mu-\hat{\mu}\|_2^2 + \mathrm{Tr}[\Sigma+\hat{\Sigma}-2C] \\ \text{s.t.} & \begin{bmatrix} \Sigma & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \succeq 0. \end{cases}$$

By Proposition 2.2, the optimal value of the resulting semidefinite program amounts to $\mathrm{G}^2((\mu,\Sigma),(\hat{\mu},\hat{\Sigma}))$. The claim follows by taking square roots on both sides. □

The proof of Theorem 2.20 reveals that the squared Gelbrich distance coincides with the minimum of a relaxed optimal transport problem, which only requires the marginals of the transportation plan γ to have the same first- and second-order moments as $\mathbb{P}$ and $\hat{\mathbb{P}}$, respectively. Gelbrich's inequality may be useful when the exact 2-Wasserstein distance is inaccessible. Indeed, computing the 2-Wasserstein distance between a discrete and a continuous distribution is #P-hard already when the discrete distribution has only two atoms (Taşkesen, Shafieezadeh-Abadeh and Kuhn 2023*a*). Computing the 2-Wasserstein distance may even be #P-hard when both distributions are discrete (Taşkesen, Shafieezadeh-Abadeh, Kuhn and Natarajan 2023*b*). If both $\mathbb{P}$ and $\hat{\mathbb{P}}$ are Gaussian, then Gelbrich's inequality collapses to an equality. Thus the 2-Wasserstein distance between two Gaussian distributions matches the Gelbrich distance between their mean vectors and covariance matrices (Givens and Shortt 1984, Proposition 7). This classical result, which actually predates Gelbrich's inequality, is now recognized as an immediate consequence of a celebrated optimality condition for optimal transport problems by Brenier (1991). Using Brenier's optimality condition, one can prove more generally that if $\hat{\mathbb{P}}$ is a positive semidefinite affine pushforward of $\mathbb{P}$, that is, if there exists an affine function $f(z) = Az + b$ with $A \in \mathbb{S}_+^d$ and $b \in \mathbb{R}^d$ such that $\hat{\mathbb{P}} = \mathbb{P} \circ f^{-1}$, then the 2-Wasserstein distance between $\mathbb{P}$ and $\hat{\mathbb{P}}$ again matches the Gelbrich distance between their mean vectors and covariance matrices (Nguyen *et al.* 2021, Theorem 2).

The 2-Wasserstein ambiguity set has found applications in machine learning (Sinha, Namkoong and Duchi 2018, Blanchet *et al.* 2019*b*, Blanchet, Murthy and Si 2022*b*, Blanchet, Murthy and Zhang 2022*c*), inverse optimization (Mohajerin Esfahani *et al.* 2018), two-stage stochastic programming (Hanasusanto and Kuhn 2018), estimation and filtering (Shafieezadeh-Abadeh *et al.* 2018, Nguyen *et al.* 2023*b*, Kargin *et al.* 2024*b*), portfolio optimization (Blanchet, Chen and Zhou 2022*a*, Nguyen *et al.* 2021) and control theory (Al Taha *et al.* 2023, Hajar *et al.* 2023, Hakobyan and Yang 2024, Taşkesen *et al.* 2024, Kargin *et al.* 2024*a*,*c*,*d*).

2.3.2. *Lévy–Prokhorov ambiguity sets*

The Lévy–Prokhorov distance is one of the most widely used probability metrics because it metrizes the topology of weak convergence on $\mathcal{P}(\mathcal{Z})$. We assume below that $d(\cdot,\cdot)$ is a continuous metric on $\mathcal{Z}$. For any set $\mathcal{B} \subseteq \mathcal{Z}$ and $r \geq 0$, we use

$$\mathcal{B}_r = \{z \in \mathcal{Z} : \exists z' \in \mathcal{B} \text{ with } d(z,z') \leq r\} \tag{2.29}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

to denote the r-neighbourhood of $\mathcal{B}$. The dependence of $\mathcal{B}_r$ on the metric d is notationally suppressed because d is usually obvious from the context. With these preparations, we are now ready to define the Lévy–Prokhorov distance.

Definition 2.21 (Lévy–Prokhorov distance). For any metric $d(\cdot,\cdot)$ on $\mathcal{Z}$, the Lévy–Prokhorov distance $\mathrm{LP}\colon \mathcal{P}(\mathcal{Z})\times\mathcal{P}(\mathcal{Z})\to[0,1]$ induced by d is defined via

$$\mathrm{LP}(\mathbb{P},\hat{\mathbb{P}}) = \inf\{r\geq 0\colon \mathbb{P}(\mathcal{B})\leq \hat{\mathbb{P}}(\mathcal{B}_r)+r \text{ for all Borel sets } \mathcal{B}\subseteq\mathcal{Z}\},$$

where $\mathcal{B}_r$ is defined in (2.29).

The Lévy–Prokhorov distance is bounded by 1 and vanishes if and only if its arguments match. In addition, one can easily show that it satisfies the triangle inequality. However, it appears to be asymmetric. The next proposition reveals that the Lévy–Prokhorov distance is closely linked to the theory of optimal transport.

Proposition 2.22 (Strassen 1965). If the transportation cost function c_r corresponding to $r\geq 0$ is defined by $c_r(z,\hat{z}) = \mathbb{1}_{d(z,\hat{z})>r}$ for all $z,\hat{z}\in\mathcal{Z}$, then

$$\mathrm{LP}(\mathbb{P},\hat{\mathbb{P}}) = \inf\{r\geq 0\colon \mathrm{OT}_{c_r}(\mathbb{P},\hat{\mathbb{P}})\leq r\}.$$

Proof. Note that c_r is lower semicontinuous because the metric d is continuous by assumption. By Proposition 2.16, $\mathrm{OT}_{c_r}(\mathbb{P},\hat{\mathbb{P}})$ thus admits the dual representation

$$\begin{array}{ll}\sup\limits_{f\in\mathcal{L}^1(\mathbb{P}),\, g\in\mathcal{L}^1(\hat{\mathbb{P}})} & \displaystyle\int_{\mathcal{Z}} f(z)\,\mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} g(\hat{z})\,\mathrm{d}\hat{\mathbb{P}}(\hat{z}) \\ \text{s.t.} & f(z)-g(\hat{z})\leq \mathbb{1}_{d(z,\hat{z})>r}\quad \forall z,\hat{z}\in\mathcal{Z}.\end{array} \tag{2.30}$$

Here, for any fixed g, it is optimal to push f up such that for all $z\in\mathcal{Z}$ we have

$$f(z) = \inf_{\hat{z}\in\mathcal{Z}} g(\hat{z}) + \mathbb{1}_{d(z,\hat{z})>r} \implies \inf_{\hat{z}\in\mathcal{Z}} g(\hat{z}) \leq f(z) \leq 1 + \inf_{\hat{z}\in\mathcal{Z}} g(\hat{z}). \tag{2.31a}$$

Also, for any fixed f, it is optimal to push g down such that for all $\hat{z}\in\mathcal{Z}$ we have

$$g(\hat{z}) = \sup_{z\in\mathcal{Z}} f(z) - \mathbb{1}_{d(z,\hat{z})>r} \implies \sup_{z\in\mathcal{Z}} f(z) - 1 \leq g(\hat{z}) \leq \sup_{z\in\mathcal{Z}} f(z). \tag{2.31b}$$

Combining the upper bound on $g(\hat{z})$ in (2.31b) with the upper bound on $f(z)$ in (2.31a) further implies that $g(\hat{z})\leq \sup_{z\in\mathcal{Z}} f(z)\leq 1+\inf_{z'\in\mathcal{Z}} g(z')$. At optimality, (2.31a) and (2.31b) must hold simultaneously, and thus we have

$$\inf_{z'\in\mathcal{Z}} g(z') \leq f(z) \leq 1 + \inf_{z'\in\mathcal{Z}} g(z') \quad\text{and}\quad \inf_{z'\in\mathcal{Z}} g(z') \leq g(\hat{z}) \leq 1 + \inf_{z'\in\mathcal{Z}} g(z')$$

for all $z,\hat{z}\in\mathcal{Z}$. Note that, as both $\mathbb{P}$ and $\hat{\mathbb{P}}$ are probability distributions, the objective function of the dual optimal transport problem (2.30) remains invariant under the substitutions $f(z)\leftarrow f(z)-\inf_{z'\in\mathcal{Z}} g(z')$ and $g(\hat{z})\leftarrow g(\hat{z})-\inf_{z'\in\mathcal{Z}} g(z')$. In the following, we may thus assume without loss of generality that $0\leq f(z)\leq 1$ for all $z\in\mathcal{Z}$ and that $0\leq g(\hat{z})\leq 1$ for all $\hat{z}\in\mathcal{Z}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

As f and g are now normalized to $[0,1]$, they admit the integral representations

$$f(z) = \int_0^1 \mathbb{1}_{f(z)\geq\tau}\,\mathrm{d}\tau \quad \text{for all } z \in \mathcal{Z}, \quad g(\hat{z}) = \int_0^1 \mathbb{1}_{g(\hat{z})\geq\tau}\,\mathrm{d}\tau \quad \text{for all } \hat{z} \in \mathcal{Z}.$$

Next, one can show that f and g satisfy the constraints in (2.30) if and only if

$$\mathbb{1}_{f(z)\geq\tau} - \mathbb{1}_{g(\hat{z})\geq\tau} \leq \mathbb{1}_{d(z,\hat{z})>r} \quad \text{for all } z, \hat{z} \in \mathcal{Z},\ \tau \in [0,1]. \tag{2.32}$$

Note first that (2.32) is trivially satisfied unless its left-hand side evaluates to 1 and its right-hand side evaluates to 0. This happens if and only if $f(z) \geq \tau$ and $g(\hat{z}) < \tau$ for some $\tau \in [0,1]$ and $z, \hat{z} \in \mathcal{Z}$ with $d(z,\hat{z}) \leq r$. This is impossible, however, because it implies that $f(z) - g(\hat{z}) > 0$ for some $z, \hat{z}$ with $d(z,\hat{z}) \leq r$, thus contradicting the constraints in (2.30). Hence the constraints in (2.30) imply (2.32). The converse implication follows immediately from the integral representations of f and g.

Finally, note that $\mathbb{1}_{f(z)\geq\tau}$ and $\mathbb{1}_{g(\hat{z})\geq\tau}$ are the characteristic functions of the Borel sets $\mathcal{B} = \{z \in \mathcal{Z} \colon f(z) \geq \tau\}$ and $\mathcal{C} = \{\hat{z} \in \mathcal{Z} \colon g(\hat{z}) \geq \tau\}$, respectively. Note also that (2.32) holds if and only if $\mathcal{C} \supseteq \mathcal{B}_r$. Recalling their integral representations, we may thus conclude that the functions f and g are feasible in (2.30) if and only if they represent convex combinations of (infinitely many) characteristic functions of the form $\mathbb{1}_{z\in\mathcal{B}}$ and $\mathbb{1}_{\hat{z}\in\mathcal{C}}$ for some Borel sets $\mathcal{B}$ and $\mathcal{C}$ with $\mathcal{C} \supseteq \mathcal{B}_r$. As the objective function of (2.30) is linear in f and g, its supremum does not change if we restrict the feasible set to such characteristic functions. Hence (2.30) reduces to

$$\mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) = \sup\{\mathbb{P}(\mathcal{B}) - \hat{\mathbb{P}}(\mathcal{C}) \colon \mathcal{B}, \mathcal{C} \subseteq \mathcal{Z} \text{ are Borel sets with } \mathcal{C} \supseteq \mathcal{B}_r\}.$$

Clearly, it is always optimal to set $\mathcal{C} = \mathcal{B}_r$, and thus the claim follows. □

While Proposition 2.22 follows from Strassen (1965, Theorem 11), the proof shown here parallels that of Villani (2003, Theorem 1.27). As a by-product, Proposition 2.22 reveals that the Lévy–Prokhorov distance is symmetric, which is not evident from its definition. Thus it indeed constitutes a metric.

The Lévy–Prokhorov ambiguity set of radius $r \geq 0$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{LP}(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

For our purposes, the most important implication of Proposition 2.22 is that $\mathcal{P}$ can be viewed as special instance of an optimal transport ambiguity set, that is, we have

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}$$

for any radius $r \geq 0$. Lévy–Prokhorov ambiguity sets were first introduced in the context of chance-constrained programming (Erdoğan and Iyengar 2006). They also naturally emerge in data-driven decision-making and the training of robust machine learning models (Pydi and Jog 2021, Bennouna and Van Parys 2023, Bennouna, Lucas and Van Parys 2023). We close this section with a useful corollary, which follows immediately from the last part of the proof of Proposition 2.22.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Corollary 2.23. If the transportation cost function c_r corresponding to $r \geq 0$ is defined by $c_r(z, \hat{z}) = \mathbb{1}_{d(z,\hat{z})>r}$ for all $z, \hat{z} \in \mathcal{Z}$, then we have

$$\mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) = \sup\{\mathbb{P}(\mathcal{B}) - \hat{\mathbb{P}}(\mathcal{B}_r) \colon \mathcal{B} \subseteq \mathcal{Z} \text{ is a Borel set}\},$$

where the r-neighbourhood $\mathcal{B}_r$ is defined in (2.29).

2.3.3. Total variation ambiguity sets revisited

In Section 2.2.3 we showed that the total variation distance constitutes an instance of a ϕ-divergence; see Proposition 2.11. We can now demonstrate that the total variation distance is also an instance of an optimal transport discrepancy.

Proposition 2.24. If $c(z, \hat{z}) = \mathbb{1}_{z \neq \hat{z}}$ for all $z, \hat{z} \in \mathcal{Z}$, then we have

$$\mathrm{TV}(\mathbb{P}, \hat{\mathbb{P}}) = \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) = \inf_{\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})} \gamma(Z \neq \hat{Z}).$$

Proof. By Definition 2.10, the total variation distance satisfies

$$\begin{aligned} \mathrm{TV}(\mathbb{P}, \hat{\mathbb{P}}) &= \sup\{|\mathbb{P}(\mathcal{B}) - \hat{\mathbb{P}}(\mathcal{B})| \colon \mathcal{B} \subseteq \mathcal{Z} \text{ is a Borel set}\} \\ &= \sup\{\mathbb{P}(\mathcal{B}) - \hat{\mathbb{P}}(\mathcal{B}) \colon \mathcal{B} \subseteq \mathcal{Z} \text{ is a Borel set}\} \\ &= \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}), \end{aligned}$$

where the second equality holds because the complement of any Borel set is again a Borel set. The third equality follows from Corollary 2.23 for $r = 0$, which applies because $c(z, \hat{z}) = \mathbb{1}_{d(z,\hat{z})>0}$ for any (continuous) metric d on $\mathcal{Z}$. Since $c(z, \hat{z}) = \mathbb{1}_{z \neq \hat{z}}$, we also have

$$\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) = \inf_{\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})} \mathbb{E}_\gamma\left[\mathbb{1}_{Z \neq \hat{Z}}\right] = \inf_{\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})} \gamma(Z \neq \hat{Z}).$$

This observation completes the proof. □

Proposition 2.24 readily implies that any total variation ambiguity set can also be viewed as a special instance of an optimal transport ambiguity set.

2.3.4. ∞-Wasserstein ambiguity sets

Section 2.3.1 focuses exclusively on p-Wasserstein distances corresponding to finite exponents $p \in [1, \infty)$. The ∞-Wasserstein distance requires special treatment.

Definition 2.25 (∞-Wasserstein distance). The ∞-Wasserstein distance

$$\mathrm{W}_\infty \colon \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty]$$

corresponding to a continuous metric $d(\cdot, \cdot)$ on $\mathcal{Z}$ is

$$\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) = \inf_{\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})} \operatorname{ess\,sup}_\gamma[d(Z, \hat{Z})], \tag{2.33}$$

where the essential supremum of $d(Z, \hat{Z})$ under γ is given by

$$\operatorname{ess\,sup}_\gamma[d(Z, \hat{Z})] = \inf_{\tau \in \mathbb{R}} \{\tau \colon \gamma(d(Z, \hat{Z}) > \tau) = 0\}.$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Definition 2.25 makes sense because the ∞-Wasserstein distance can be obtained from the p-Wasserstein distance in the limit when p tends to infinity.

Proposition 2.26 (Givens and Shortt 1984). For any $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ we have

$$\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) = \lim_{p\to\infty} \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{p\geq 1} \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}).$$

Proof. If $p \geq q \geq 1$, then $f(t) = t^{q/p}$ is concave on $\mathbb{R}_+$. This implies that

$$\begin{aligned}\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) &= \inf_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} (\mathbb{E}_\gamma[d(Z, \hat{Z})^p]^{q/p})^{1/q} \\ &\geq \inf_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} (\mathbb{E}_\gamma[d(Z, \hat{Z})^q])^{1/q} \\ &= \mathrm{W}_q(\mathbb{P}, \hat{\mathbb{P}})\end{aligned}$$

thanks to Jensen's inequality. Hence $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})$ is non-decreasing in the exponent p as long as $p \in [1, \infty)$. In addition, for any transportation plan $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$ and exponent $p \in [1, \infty)$, the definition of the essential supremum readily implies that

$$(\mathbb{E}_\gamma[d(Z, \hat{Z})^p])^{1/p} \leq \operatorname{ess\,sup}_\gamma[d(Z, \hat{Z})^p]^{1/p} = \operatorname{ess\,sup}_\gamma[d(Z, \hat{Z})].$$

Minimizing both sides of this inequality across all $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$ further implies that $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) \leq \mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}})$ for all $p \in [1, \infty)$. In summary, we may thus conclude that

$$\lim_{p\to\infty} \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{p\geq 1} \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) \leq \mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}).$$

It remains to be shown that the last inequality in fact holds as an equality. To see this, fix some tolerance $\varepsilon > 0$. For any $p \in \mathbb{N}$, let $\gamma_p \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$ be a coupling with $\mathbb{E}_{\gamma_p}[d(Z, \hat{Z})^p]^{1/p} = \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})$. Note that γ_p exists because, as we will see in Corollary 3.16 and Proposition 3.3 below, $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ is weakly compact and $\mathbb{E}_\gamma[d(Z, \hat{Z})^p]$ is weakly lower semicontinuous in γ. Next, let $\{\gamma_{p(j)}\}_{j\in\mathbb{N}}$ be a subsequence that converges weakly to some coupling $\gamma_\infty \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$, which exists again because $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ is weakly compact. We proceed by case distinction.

Case 1. If $\operatorname{ess\,sup}_{\gamma_\infty}[d(Z, \hat{Z})]$ is finite, define the open set

$$\mathcal{B} = \{(z, \hat{z}) \in \mathcal{Z} \times \mathcal{Z} \colon d(z, \hat{z}) > \operatorname{ess\,sup}_{\gamma_\infty}[d(Z, \hat{Z})] - \varepsilon\},$$

and note that $\gamma_\infty(\mathcal{B}) > 0$ by the definition of the essential supremum. We then find

$$\begin{aligned}\mathrm{W}_{p(j)}(\mathbb{P}, \hat{\mathbb{P}}) &\geq \left(\int_{\mathcal{B}} d(z, \hat{z})^{p(j)}\, \mathrm{d}\gamma_{p(j)}(z, \hat{z})\right)^{1/(p(j))} \\ &\geq \gamma_{p(j)}(\mathcal{B})^{1/(p(j))}(\operatorname{ess\,sup}_{\gamma_\infty}[d(Z, \hat{Z})] - \varepsilon) \\ &\geq \gamma_{p(j)}(\mathcal{B})^{1/(p(j))}(\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) - \varepsilon).\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Since $\mathcal{B}$ is open and $\gamma_{p(j)}$ converges weakly to γ_∞ as j grows, the Portmanteau theorem (Billingsley 2013, Theorem 2.1 (iiv)) implies that $\liminf_{j\to\infty} \gamma_{p(j)}(\mathcal{B}) \geq \gamma_\infty(\mathcal{B}) > 0$. Thus $\gamma_{p(j)}(\mathcal{B})^{1/p(j)}$ converges to 1 as j grows, and we obtain

$$\lim_{p\to\infty} \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) \geq \mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) - \varepsilon.$$

As this inequality holds for any tolerance $\varepsilon > 0$, the above reasoning finally implies that $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})$ indeed converges to $\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}})$ for large p.

Case 2. If $\operatorname{ess\,sup}_{\gamma_\infty}[d(Z, \hat{Z})] = \infty$, then we replace $\operatorname{ess\,sup}_{\gamma_\infty}[d(Z, \hat{Z})]$ in the definition of the open set $\mathcal{B}$ with an arbitrarily large constant. Proceeding as in Case 1 eventually reveals that $\lim_{p\to\infty} \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) = \mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) = \infty$. □

To develop some intuition for Proposition 2.26, consider the optimal transport problem in the definition of $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})$. If $p > 1$, then the cost $c(z, \hat{z}) = d(z, \hat{z})^p$ of transporting one unit of probability mass from $\hat{z}$ to z grows superlinearly with the distance $d(z, \hat{z})$. Hence, parts of the distribution $\hat{\mathbb{P}}$ that are transported further under an optimal transportation plan contribute more to $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})$. In addition, as p tends to infinity, eventually only the portion of the distribution $\hat{\mathbb{P}}$ that is transported the furthest has an impact on $\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}})$. Even more, only the largest transportation *distance* matters, whereas the *amount* of probability mass transported is irrelevant.

Despite Proposition 2.26, the optimal transport problems in the definitions of the Wasserstein distances of order $p < \infty$ and of order $p = \infty$ are fundamentally different. Indeed, if $p < \infty$, then the objective function $\mathbb{E}_\gamma[d(Z, \hat{Z})^p]$ of the optimal transport problem is linear in the transportation plan γ. If $p = \infty$, on the other hand, then the objective function $\operatorname{ess\,sup}_\gamma[d(Z, \hat{Z})]$ is not even convex, but rather quasi-convex, in γ (Jylhä 2015, Lemma 2.2); see also Champion, De Pascale and Juutinen (2008). Thus ∞-Wasserstein distances require a more subtle treatment.

The next proposition relates the ∞-Wasserstein distance to a standard optimal transport problem. Therefore it has computational relevance.

Proposition 2.27. If the transportation cost function c_r corresponding to $r \geq 0$ is defined by $c_r(z, \hat{z}) = \mathbb{1}_{d(z,\hat{z})>r}$ for all $z, \hat{z} \in \mathcal{Z}$, then we have

$$\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) = \inf\{r \geq 0 \colon \mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) \leq 0\}.$$

Proof. Recall that $\mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) = \inf\{\mathbb{E}_\gamma[c_r(Z, \hat{Z})] \colon \gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})\}$. Note that the underlying optimal transport problem is solvable because $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ is weakly compact and because $\mathbb{E}_\gamma[d(Z, \hat{Z})^p]$ is weakly lower semicontinuous in γ thanks to Corollary 3.16 and Proposition 3.3 below, respectively. Therefore we have

$$\begin{aligned}
&\inf\{r \geq 0 \colon \mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) \leq 0\} \\
&\quad = \inf\{r \geq 0 \colon \exists \gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}}) \text{ with } \mathbb{E}_\gamma[c_r(Z, \hat{Z})] = 0\} \\
&\quad = \inf_{\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}}),\, r \in \mathbb{R}_+} \{r \colon \gamma[d(Z, \hat{Z}) > r] = 0\} \\
&\quad = \mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}),
\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where the first equality holds because $\mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}})$ is non-negative and because the underlying optimal transport problem is solvable. The second equality follows from the definitions of c_r and the ∞-Wasserstein distance. □

Combining Proposition 2.27 with Corollary 2.23 immediately yields the following equivalent characterization of the ∞-Wasserstein distance.

Corollary 2.28 (Givens and Shortt 1984). The ∞-Wasserstein distance satisfies

$$\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) = \inf\{r \geq 0 \colon \mathbb{P}(\mathcal{B}) \leq \hat{\mathbb{P}}(\mathcal{B}_r) \text{ for all Borel sets } \mathcal{B} \subseteq \mathcal{Z}\},$$

where the r-neighbourhood $\mathcal{B}_r$ is defined in (2.29).

The ∞-Wasserstein ambiguity set of radius $r \geq 0$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}. \tag{2.34}$$

Proposition 2.27 implies that $\mathcal{P}$ coincides with an optimal transport ambiguity set with transportation cost function $c_r(z, \hat{z}) = \mathbb{1}_{d(z,\hat{z})>r}$, that is, we have

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) \leq 0\}.$$

DRO with ∞-Wasserstein ambiguity sets has strong connections to adversarial machine learning (Gao, Chen and Kleywegt 2017, García Trillos and García Trillos 2022, García Trillos and Murray 2022, García Trillos and Jacobs 2023, Bungert, García Trillos and Murray 2023, Bungert, Laux and Stinson 2024, Gao *et al.* 2024*b*, Pydi and Jog 2024, Frank and Niles-Weed 2024*a*,*b*) and kernel density estimation (Xu, Caramanis and Mannor 2012*a*). In addition, ∞-Wasserstein ambiguity sets are used in two- and multi-stage stochastic programming (Xie 2020, Bertsimas, Shtern and Sturt 2022, Bertsimas *et al.* 2023), portfolio optimization (Nguyen *et al.* 2024) and robust learning (Nguyen *et al.* 2020, Wang, Nguyen and Hanasusanto 2024*d*).

2.4. Other ambiguity sets

There exist several ambiguity sets that cannot be classified as moment, ϕ-divergence or optimal transport ambiguity sets. In the following we offer a brief overview of these ambiguity sets without providing extensive mathematical details.

2.4.1. Marginal ambiguity sets

Marginal ambiguity sets specify the marginal distributions of multiple subvectors of Z without detailing their joint distribution. The simplest example of a marginal ambiguity set is the Fréchet ambiguity set, which specifies the marginal distributions of all individual components of Z but provides no information about their copula. Thus the Fréchet ambiguity set is parametrized by d marginal cumulative distribution functions $F_i \colon \mathbb{R} \to [0, 1]$, $i \in [d]$, and can be represented as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d) \colon \mathbb{P}(Z_i \leq z_i) = F_i(z_i) \ \ \forall z_i \in \mathbb{R}, \ \forall i \in [d]\}. \tag{2.35}$$

Here F_i is an arbitrary cumulative distribution function, that is, a right-continuous, non-decreasing function with $\lim_{z_i \to -\infty} F_i(z_i) = 0$ and $\lim_{z_i \to +\infty} F_i(z_i) = 1$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Fréchet ambiguity sets are relevant for probabilistic logic. Imagine that each Z_i represents a binary variable that evaluates to 1 if a certain event occurs and to 0 otherwise, and assume that the probability of each event is known, whereas the joint distribution of all events is unknown. In this setting, Boole (1854) was interested in computing bounds on the probability of a composite event encoded by a Boolean function of the variables $Z_i, i \in [d]$. Almost a century later, Fréchet (1935) derived explicit inequalities for the probabilities of such composite events, which are now called Fréchet inequalities. Note that these Fréchet inequalities can be obtained by minimizing or maximizing the probability of the composite event over all distributions in a Fréchet ambiguity set with Bernoulli marginals. More recently, there has been growing interest in generalized Fréchet inequalities, which bound the risk of general (not necessarily Boolean) functions of Z with respect to all distributions in a Fréchet ambiguity set with general (not necessarily Bernoulli) marginals. For example, a wealth of Fréchet inequalities for the risk of a sum of random variables have emerged in finance and risk management (Rüschendorf 1983, 1991, Embrechts and Puccetti 2006, Wang and Wang 2011, Wang, Peng and Yang 2013, Puccetti and Rüschendorf 2013, Van Parys, Goulart and Embrechts 2016*a*, Blanchet, Lam, Liu and Wang 2024*a*). In addition, Natarajan, Song and Teo (2009*b*) derive sharp bounds for the worst-case expectation of a piecewise affine functions over a Fréchet ambiguity set. We highlight that Fréchet ambiguity sets are also relevant because they coincide with the feasible sets of multi-marginal optimal transport problems, which can sometimes be solved in polynomial time (Pass 2015, Altschuler and Boix-Adsera 2023, Natarajan, Padmanabhan and Ramachandra 2023).

General marginal ambiguity sets specify the marginal distributions of several (possibly overlapping) subsets of the set $\{Z_i : i \in [d]\}$ of random variables. However, checking whether such an ambiguity set is non-empty is NP-complete even if each Z_i is a Bernoulli random variable and each subset accommodates merely two elements (Honeyman, Ladner and Yannakakis 1980, Georgakopoulos, Kavvadias and Papadimitriou 1988). Computing worst-case expectations over marginal ambiguity sets is thus intractable unless the subsets of random variables with known marginals are disjoint (Doan and Natarajan 2012) or if the corresponding overlap graph displays a running intersection property (Doan, Li and Natarajan 2015).

Marginal ambiguity sets are attractive because, given limited statistical data, it is far easier to estimate low-dimensional marginals than their global dependence structure. However, even univariate marginals cannot be estimated exactly. For this reason, several researchers study marginal ambiguity sets that provide only limited information about the marginals such as bounds on marginal moments or marginal dispersion measures (Bertsimas *et al.* 2004, Bertsimas, Natarajan and Teo 2006*a*,*b*, Chen, Sim, Sun and Teo 2010, Mishra, Natarajan, Tao and Teo 2012, Natarajan, Sim and Uichanco 2018).

A related stream of literature focuses on ambiguity sets under which the random variables $Z_i, i \in [d]$, are *independent* and governed by ambiguous marginal distributions. For example, the Hoeffding ambiguity set contains all joint distri-

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

butions on a box with independent (and completely unknown) marginals, whereas the Bernstein ambiguity set contains all distributions from within the Hoeffding ambiguity set subject to marginal moment bounds (Nemirovski and Shapiro 2007, Hanasusanto *et al.* 2015*a*). Bernstein ambiguity sets that constrain the mean as well as the mean-absolute deviation of each marginal are used to derive safe tractable approximations for distributionally robust chance-constrained programs (Postek, Ben-Tal, den Hertog and Melenberg 2018), two-stage integer programs (Postek *et al.* 2018, Postek, Romeijnders, den Hertog and van der Vlerk 2019) and queueing systems (Wang, Prasad, Hanasusanto and Hasenbein 2024*e*).

DRO with marginal ambiguity sets has close connections to submodularity and to the theory of comonotonicity in risk management (Tchen 1980, Rüschendorf 2013, Bach 2013, 2019, Natarajan *et al.* 2023, Long, Qi and Zhang 2024). It has a broad range of diverse applications ranging from discrete choice modelling (Natarajan *et al.* 2009*b*, Mishra *et al.* 2014, Chen *et al.* 2022, Ruan, Li, Murthy and Natarajan 2023) to queuing theory (van Eekelen, den Hertog and van Leeuwaarden 2022), transportation (Wang, Chen and Liu 2020, Shehadeh 2023), chance-constrained programming (Xie, Ahmed and Jiang 2022), scheduling (Mak *et al.* 2015), inventory management (Liu, Chen, Wang and Wang 2024*a*), the analysis of complex networks (Chen, Padmanabhan, Lim and Natarajan 2020, Van Leeuwaarden and Stegehuis 2021, Brugman *et al.* 2022) and mechanism design (Carroll 2017, Gravin and Lu 2018, Chen *et al.* 2024*b*, Wang, Liu and Zhang 2024*c*, Wang 2024). For further details we refer to the comprehensive monograph by Natarajan (2021).

2.4.2. Mixture ambiguity sets and structural ambiguity sets

Let $\Theta \subseteq \mathbb{R}^m$ be a Borel set and $\mathbb{P}_\theta \in \mathcal{P}(\mathcal{Z})$ a parametric distribution that is uniquely determined by $\theta \in \Theta$. Assume that $\mathbb{P}_\theta(Z \in \mathcal{B})$ is a Borel-measurable function of θ for every fixed Borel set $\mathcal{B} \subseteq \mathcal{Z}$. The parametric distribution family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ can then be used as a mixture family, which induces the mixture ambiguity set

$$\mathcal{P} = \left\{ \int_\Theta \mathbb{P}_\theta \, \mathrm{d}\mathbb{Q}(\theta) : \mathbb{Q} \in \mathcal{P}(\Theta) \right\}. \tag{2.36}$$

Thus $\mathcal{P}$ contains all distributions that can be represented as mixtures of the distributions $\mathbb{P}_\theta$, $\theta \in \Theta$. Put differently, for every $\mathbb{P} \in \mathcal{P}$ there exists a mixture distribution $\mathbb{Q} \in \mathcal{P}(\Theta)$ with $\mathbb{P}(Z \in \mathcal{B}) = \int_\Theta \mathbb{P}_\theta(Z \in \mathcal{B}) \, \mathrm{d}\mathbb{Q}(\theta)$ for all Borel sets $\mathcal{B} \subseteq \mathcal{Z}$. This construction ensures that $\mathcal{P} \subseteq \mathcal{P}(\mathcal{Z})$ is convex. For example, if $\mathbb{P}_\theta$ is a Gaussian distribution whose mean and covariance matrix are encoded by θ, then $\mathcal{P}$ contains (possibly continuous) mixtures of Gaussians. Mixture ambiguity sets corresponding to compact parameter sets Θ are studied by Lasserre and Weisser (2021), who develop a semidefinite programming-based hierarchy of increasingly tight inner approximations for the feasible set of a distributionally robust chance constraint.

Note that $\mathcal{P}$ can be viewed as the convex hull of the parametric distribution family $\{\mathbb{P}_\theta : \theta \in \Theta\}$. A classical result in convex analysis due to Minkowski asserts that

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

any compact convex subset of a Euclidean vector space coincides with the convex hull of its extreme points. Choquet theory (Phelps 1965) seeks similar extreme point representations for convex compact subsets of topological vector spaces. For example, if $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is the set of all extreme distributions of a weakly compact convex ambiguity set $\mathcal{P}$, then (2.36) constitutes a Choquet representation of $\mathcal{P}$.

Families of distributions that share certain structural properties sometimes admit a Choquet representation of the form (2.36). For example, let $\mathcal{P}$ be the family of all distributions $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ that are point symmetric about the origin. This means that $\mathbb{P}(Z \in \mathcal{B}) = \mathbb{P}(-Z \in \mathcal{B})$ for every Borel set $\mathcal{B} \subseteq \mathbb{R}^d$. One can then show that all extreme distributions of $\mathcal{P}$ are representable as $\mathbb{P}_\theta = \frac{1}{2}\delta_{+\theta} + \frac{1}{2}\delta_{-\theta}$ for some $\theta \in \mathbb{R}^d$. Thus $\mathcal{P}$ admits a Choquet representation of the form (2.36). As another example, let $\mathcal{P}$ be the family of all distributions $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ that are α-unimodal about the origin for some $\alpha > 0$. This means that $t^\alpha \mathbb{P}(Z \in \mathcal{B}/t)$ is non-decreasing in $t > 0$ for every Borel set $\mathcal{B} \subseteq \mathbb{R}^d$. One can then show that every extreme distribution of $\mathcal{P}$ is a distribution $\mathbb{P}_\theta$ supported on the line segment from 0 to $\theta \in \mathbb{R}^d$ with the property that $\mathbb{P}_\theta(\|Z\|_2 \leq t\|\theta\|_2) = t^\alpha$ for all $t \in [0, 1]$. Thus $\mathcal{P}$ again admits a Choquet representation of the form (2.36). We remark that d-unimodal distributions on $\mathbb{R}^d$ are also called star-unimodal. One readily verifies that a distribution with a continuous probability density function is star-unimodal if and only if the density function is non-increasing along each ray emanating from the origin. In addition, one can show that the family of all α-unimodal distributions converges – in a precise sense – to the family of *all* possible distributions on $\mathbb{R}^d$ as α tends to infinity. For more information on structural distribution families and their Choquet representations, we refer to Dharmadhikari and Joag-Dev (1988).

The moment ambiguity sets of Section 2.1 are known to contain discrete distributions with only very few atoms; see Section 7. However, uncertainties encountered in real physical, technical or economic systems are unlikely to follow such discrete distributions. Instead, they are often expected to be unimodal. Hence an effective means to eliminate the pathological discrete distributions from a moment ambiguity set is to intersect it with the structural ambiguity set of all α-unimodal distributions for some $\alpha > 0$. Popescu (2005) combines ideas from Choquet theory and sums-of-squares polynomial optimization to approximate worst-case expectations over the resulting intersection ambiguity sets by a hierarchy of increasingly accurate bounds, each of which is computed by solving a tractable semidefinite program. Van Parys, Goulart and Kuhn (2016*b*) and Van Parys, Goulart and Morari (2019) extend this approach and establish *exact* semidefinite programming reformulations for the worst-case probability of a polyhedron and the worst-case conditional value-at-risk of a piecewise linear convex loss function across all α-unimodal distributions in a Chebyshev ambiguity set; see also Hanasusanto, Roitch, Kuhn and Wiesemann (2015*b*). Li, Jiang and Mathieu (2019*a*) demonstrate that these semidefinite programming reformulations can sometimes be simplified to highly tractable second-order cone programs. Complementing moment information with structural information generally leads to less conservative DRO models

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

as Li, Jiang and Mathieu (2016) demonstrate in the context of a power system application. Lam, Liu and Zhang (2021) consider another basic notion of distributional shape known as ortho-unimodality and build a corresponding Choquet representation to address multivariate extreme event estimation. More recently, Lam, Liu and Singham (2024) combine Choquet theory with importance sampling and likelihood ratio techniques for modelling distribution shapes.

2.4.3. Non-standard ϕ-divergence and optimal transport ambiguity sets
A wealth of non-standard ϕ-divergences and optimal transport discrepancies have been proposed to measure the dissimilarity between probability distributions. They offer great flexibility in designing ambiguity sets with complementary computational and statistical properties. Non-standard distance measures notably include smoothed ϕ-divergences (Zeitouni and Gutman 1991, Yang and Chen 2018, Liu, Van Parys and Lam 2023) as well as combinations of ϕ-divergences and optimal transport discrepancies (Reid and Williamson 2011, Dupuis and Mao 2022, Van Parys 2024). In addition, they include coherent Wasserstein distances (Li and Mao 2022) and Sinkhorn divergences (Wang, Gao and Xie 2021), as well as divergences based on causal optimal transport (Analui and Pflug 2014, Pflug and Pichler 2014, Yang *et al.* 2022, Gao, Arora and Huang 2024*a*, Jiang and Obloj 2024), outlier-robust optimal transport (Nietert, Goldfeld and Shafiee 2024*a*,*b*), mixed-feature optimal transport (Selvi, Belbasi, Haugh and Wiesemann 2022, Belbasi, Selvi and Wiesemann 2023), cluster-based optimal transport (Wang, Becker, Van Parys and Stellato 2024*a*), partial optimal transport (Esteban-Pérez and Morales 2022), sliced optimal transport (Olea, Rush, Velez and Wiesel 2022), multimarginal optimal transport (Lau and Liu 2022, García Trillos, Jacobs and Kim 2023, Rychener, Esteban-Pérez, Morales and Kuhn 2024) and constrained conditional moment optimal transport (Li *et al.* 2022, Blanchet, Kuhn, Li and Taşkesen 2023, Sauldubois and Touzi 2024).

2.4.4. Ambiguity sets based on integral probability metrics
Let $\mathcal{F}$ be a family of Borel-measurable test functions $f\colon \mathcal{Z} \to \mathbb{R}$ such that $f \in \mathcal{F}$ if and only if $-f \in \mathcal{F}$. The integral probability metric generated by $\mathcal{F}$ is defined via

$$\mathrm{D}_{\mathcal{F}}(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} f(\hat{z})\, \mathrm{d}\hat{\mathbb{P}}(\hat{z})$$

for all distributions $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ under which all test functions $f \in \mathcal{F}$ are integrable. The underlying maximization problem probes how well the test functions can distinguish $\mathbb{P}$ from $\hat{\mathbb{P}}$. By construction, $\mathrm{D}_{\mathcal{F}}$ constitutes a pseudo-metric, that is, it is non-negative and symmetric (because $\mathcal{F} = -\mathcal{F}$), vanishes if its arguments match, and satisfies the triangle inequality. In addition, $\mathrm{D}_{\mathcal{F}}$ becomes a proper metric if $\mathcal{F}$ separates distributions, in which case $\mathrm{D}_{\mathcal{F}}(\mathbb{P}, \hat{\mathbb{P}})$ vanishes only if $\mathbb{P} = \hat{\mathbb{P}}$. The ambiguity set of radius $r \geq 0$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ with respect to $\mathrm{D}_{\mathcal{F}}$ is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathrm{D}_{\mathcal{F}}(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The proof of Proposition 2.11 reveals that the total variation distance is the integral probability metric generated by all Borel functions $f\colon \mathcal{Z} \to [-1/2, 1/2]$; see (2.16). The Kantorovich–Rubinstein duality established in Corollary 2.19 further shows that the 1-Wasserstein distance is the integral probability metric generated by all Lipschitz-continuous functions $f\colon \mathcal{Z} \to \mathbb{R}$ with $\mathrm{lip}(f) \leq 1$. In addition, if $\mathcal{H}$ is a reproducing kernel Hilbert space of Borel functions $f\colon \mathcal{Z} \to \mathbb{R}$ with Hilbert norm $\|\cdot\|_{\mathcal{H}}$, then the maximum mean discrepancy distance corresponding to $\mathcal{H}$ is the integral probability metric generated by the standard unit ball $\mathcal{F} = \{f \in \mathcal{H}\colon \|f\|_{\mathcal{H}} \leq 1\}$ in $\mathcal{H}$. Maximum mean discrepancy ambiguity sets are studied by Staib and Jegelka (2019), Zhu, Jitkrittum, Diehl and Schölkopf (2020, 2021), Zeng and Lam (2022) and Iyengar, Lam and Wang (2023). Husain (2020) uncovers a deep connection between DRO problems and regularized empirical risk minimization problems, which holds whenever the ambiguity set is defined via an integral probability metric.

3. Topological properties of ambiguity sets

A fundamental question of theoretical as well as practical interest is whether nature's subproblem in (1.2) is solvable or, in other words, whether the inner supremum in (1.2) is attained. In this section we will investigate under what conditions the Weierstrass extreme value theorem applies to nature's subproblem. That is, we will develop easily checkable conditions under which the ambiguity set $\mathcal{P}$ is weakly compact and the expected loss $\mathbb{E}_{\mathbb{P}}[\ell(x, Z)]$ is weakly upper semicontinuous in $\mathbb{P}$. Throughout this discussion, we assume that $\mathcal{Z}$ is a closed subset of $\mathbb{R}^d$.

A classical result by Baire asserts that a function on the real line is lower semicontinuous if and only if it can be represented as the pointwise supremum of a non-decreasing sequence of continuous functions (Baire 1905). Below we will use the following multivariate generalization of this result.

Lemma 3.1 (Stromberg 2015, p. 132). A function $f\colon \mathcal{Z} \to (-\infty, +\infty]$ is lower semicontinuous if and only if there is a non-decreasing sequence of continuous functions $f_i\colon \mathcal{Z} \to \mathbb{R}$, $i \in \mathbb{N}$, with $f(z) = \sup_{i\in\mathbb{N}} f_i(z)$ for all $z \in \mathcal{Z}$.

If f is bounded from below, then the continuous functions f_i can be assumed to be uniformly bounded. Indeed, if $f(z) \geq 0$, say, then the continuous function $f_i(z)$ can be replaced with the bounded continuous function $f_i'(z) = \min\{\max\{f_i(z), 0\}, i\}$. The sequence f_i', $i \in \mathbb{N}$, is still non-decreasing and converges pointwise to f.

Definition 3.2 (Weak convergence of probability distributions). A sequence of probability distributions $\mathbb{P}_j \in \mathcal{P}(\mathcal{Z})$, $j \in \mathbb{N}$, converges weakly to $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ if, for every bounded and continuous function $f\colon \mathcal{Z} \to \mathbb{R}$, we have

$$\lim_{j\in\mathbb{N}} \mathbb{E}_{\mathbb{P}_j}[f(Z)] = \mathbb{E}_{\mathbb{P}}[f(Z)].$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

There is a close link between the continuity properties of the expected value of $f(Z)$ with respect to the distribution $\mathbb{P}$ and the continuity properties of f. Recall that a function $F\colon \mathcal{P}(\mathcal{Z}) \to \overline{\mathbb{R}}$ is weakly continuous if $\lim_{i\to\infty} F(\mathbb{P}_i) = F(\mathbb{P})$ for every sequence $\mathbb{P}_i \in \mathcal{P}(\mathcal{Z})$, $i \in \mathbb{N}$, that converges weakly to $\mathbb{P}$. Weak lower and upper semicontinuity are defined analogously in the obvious way.

Proposition 3.3 (Continuity of expected values). If $f\colon \mathcal{Z} \to [-\infty, +\infty]$ is lower semicontinuous and bounded from below, then $\mathbb{E}_{\mathbb{P}}[f(Z)]$ is weakly lower semicontinuous in $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$. Conversely, if f is upper semicontinuous and bounded from above, then $\mathbb{E}_{\mathbb{P}}[f(Z)]$ is weakly upper semicontinuous in $\mathbb{P} \in \mathbb{P}(\mathcal{Z})$. Finally, if f is continuous and bounded, then $\mathbb{E}_{\mathbb{P}}[f(Z)]$ is weakly continuous in $\mathbb{P} \in \mathbb{P}(\mathcal{Z})$.

Proof. Assume first that f is lower semicontinuous and bounded from below. In the following, we assume without loss of generality that f is in fact non-negative. Then, by Lemma 3.1, there is a non-decreasing sequence of bounded, continuous and non-negative functions f_i, $i \in \mathbb{N}$, with $f(z) = \sup_{i\in\mathbb{N}} f_i(z)$. If $\mathbb{P}_j \in \mathcal{P}(\mathcal{Z})$, $j \in \mathbb{N}$, is any sequence of distributions that converges weakly to $\mathbb{P}$, then we find

$$\begin{aligned}
\liminf_{j\in\mathbb{N}} \mathbb{E}_{\mathbb{P}_j}[f(Z)] &= \sup_{k\in\mathbb{N}} \inf_{j\geq k} \mathbb{E}_{\mathbb{P}_j}\left[\sup_{i\in\mathbb{N}} f_i(Z)\right] \\
&= \sup_{k\in\mathbb{N}} \inf_{j\geq k} \sup_{i\in\mathbb{N}} \mathbb{E}_{\mathbb{P}_j}[f_i(Z)] \\
&\geq \sup_{i\in\mathbb{N}} \sup_{k\in\mathbb{N}} \inf_{j\geq k} \mathbb{E}_{\mathbb{P}_j}[f_i(Z)] \\
&= \sup_{i\in\mathbb{N}} \mathbb{E}_{\mathbb{P}}[f_i(\xi)] \\
&= \mathbb{E}_{\mathbb{P}}[f(Z)].
\end{aligned}$$

Here, both the second and the last equality follow from the monotone convergence theorem, which applies because each f_i is bounded and thus integrable with respect to any probability distribution and because the f_i, $i \in \mathbb{N}$, form a non-decreasing sequence of non-negative functions. The inequality follows from the interchange of the supremum over i and the infimum over j, and the third equality holds because $\mathbb{P}_j$ converges weakly to $\mathbb{P}$ and because f_i is continuous and bounded. This shows that $\mathbb{E}_{\mathbb{P}}[f(Z)]$ is weakly lower semicontinuous in $\mathbb{P}$.

The proofs of the assertions regarding weak upper semicontinuity and weak continuity are analogous and therefore omitted for brevity. □

In the following we equip the family $\mathcal{P}(\mathcal{Z})$ of all probability distributions on $\mathcal{Z}$ with the weak topology, which is generated by the open sets

$$U_{f,\delta} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon |\mathbb{E}_{\mathbb{P}}[f(Z)]| < \delta\}$$

encoded by any continuous bounded function $f\colon \mathcal{Z} \to \mathbb{R}$ and tolerance $\delta > 0$. The weak topology on $\mathcal{P}(\mathcal{Z})$ is metrized by the Prokhorov metric (Billingsley 2013, Theorem 6.8), and therefore the notions of sequential compactness and compactness are equivalent on $\mathcal{P}(\mathcal{Z})$; see e.g. Munkres (2000, Theorem 28.2).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Definition 3.4 (Tightness). A family $\mathcal{P} \subseteq \mathcal{P}(\mathcal{Z})$ of distributions is tight if, for any tolerance $\varepsilon > 0$, there is a compact set $\mathcal{C} \subseteq \mathcal{Z}$ with $\mathbb{P}(Z \notin \mathcal{C}) \leq \varepsilon$ for all $\mathbb{P} \in \mathcal{P}$.

A classical result by Prokhorov asserts that a distribution family is weakly compact if and only if it is tight and weakly closed. Prokhorov's theorem is the key tool to show that an ambiguity set is weakly compact. We state it without proof.

Theorem 3.5 (Billingsley 2013, Theorem 5.1). A family $\mathcal{P} \subseteq \mathcal{P}(\mathcal{Z})$ of distributions is weakly compact if and only if it is tight as well as weakly closed.

In the following we revisit the ambiguity sets of Section 2 one by one and determine under what conditions they are tight, weakly closed and weakly compact.

3.1. Moment ambiguity sets

The support-only ambiguity sets arguably form the simplest class of moment ambiguity sets because they impose no moment conditions at all. In fact, *all* other ambiguity sets considered in this paper are subsets of a support-only ambiguity set.

Proposition 3.6 (Support-only ambiguity sets). The set $\mathcal{P}(\mathcal{Z})$ of all distributions supported on $\mathcal{Z} \subseteq \mathbb{R}^d$ is weakly compact if and only if $\mathcal{Z}$ is compact.

Proof. Note first that $\mathcal{P}(\mathcal{Z})$ is tight if and only if $\mathcal{Z}$ is bounded. Indeed, if $\mathcal{Z}$ is bounded, then it is compact because $\mathcal{Z}$ is closed thanks to our blanket assumption. Given any $\varepsilon > 0$, we may thus set $\mathcal{C} = \mathcal{Z}$, which ensures that $\mathbb{P}(Z \notin \mathcal{C}) = 0 \leq \varepsilon$ for all $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$. Hence $\mathcal{P}(\mathcal{Z})$ is tight. If $\mathcal{Z}$ is unbounded, on the other hand, then $\mathcal{P}(\mathcal{Z})$ trivially fails to be tight. Indeed, for any compact set $\mathcal{C} \subseteq \mathcal{Z}$, the complement $\mathcal{Z} \backslash \mathcal{C}$ is non-empty because $\mathcal{C}$ is bounded and $\mathcal{Z}$ is not. Hence there exists a probability distribution $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ supported on $\mathcal{Z} \backslash \mathcal{C}$ such that $\mathbb{P}(Z \notin \mathcal{C}) = 1$.

Next, note that $\mathcal{P}(\mathcal{Z})$ is weakly closed if and only if $\mathcal{Z}$ is closed. To see this, assume first that $\mathcal{Z}$ is closed. Recall that the indicator function $\delta_{\mathcal{Z}}$ is defined by $\delta_{\mathcal{Z}}(z) = 0$ if $z \in \mathcal{Z}$ and $\delta_{\mathcal{Z}}(z) = +\infty$ if $z \notin \mathcal{Z}$. Thus, it is lower semicontinuous and bounded below. By Proposition 3.3, $\mathbb{E}_{\mathbb{P}}[\delta_{\mathcal{Z}}(Z)]$ is therefore weakly lower semicontinuous in $\mathbb{P}$. If $\mathbb{P}_j \in \mathcal{P}(\mathcal{Z})$, $j \in \mathbb{N}$, converges weakly to $\mathbb{P}$, we then have

$$0 = \liminf_{j \in \mathbb{N}} \mathbb{E}_{\mathbb{P}_j}[\delta_{\mathcal{Z}}(Z)] \geq \mathbb{E}_{\mathbb{P}}[\delta_{\mathcal{Z}}(Z)] \geq 0,$$

where the equality holds because $\mathbb{P}_j$ is supported on $\mathcal{Z}$ for every $j \in \mathbb{N}$, and the first inequality follows from weak lower semicontinuity. This implies that $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$, and thus $\mathcal{P}(\mathcal{Z})$ is weakly closed. Conversely, assume that $\mathcal{P}(\mathcal{Z})$ is weakly closed, and consider a sequence $z_j \in \mathcal{Z}$, $j \in \mathbb{N}$, converging to z. Then the sequence of Dirac distributions δ_{z_j}, $j \in \mathbb{N}$, converges weakly to δ_z, and thus we find

$$0 = \liminf_{j \in \mathbb{N}} \mathbb{E}_{\delta_{z_j}}[\delta_{\mathcal{Z}}(Z)] \geq \mathbb{E}_{\delta_z}[\delta_{\mathcal{Z}}(Z)] \geq 0.$$

Here the first inequality again holds because $\mathbb{E}_{\mathbb{P}}[\delta_{\mathcal{Z}}(Z)]$ is weakly lower semicontinuous in $\mathbb{P}$. This implies that $\mathbb{E}_{\delta_z}[\delta_{\mathcal{Z}}(Z)] = 0$, which holds if and only if $z \in \mathcal{Z}$. Thus $\mathcal{Z}$ is closed. Given these insights, the claim follows from Theorem 3.5. □

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

By using Proposition 3.6, we can now show that a moment ambiguity set of the form (2.1) is weakly compact whenever the underlying support set $\mathcal{Z}$ is compact, the moment function f is continuous and the uncertainty set $\mathcal{F}$ is closed.

Proposition 3.7 (Moment ambiguity sets). If $\mathcal{Z} \subseteq \mathbb{R}^d$ is a compact support set, $f\colon \mathcal{Z} \to \mathbb{R}^m$ is a continuous moment function and $\mathcal{F} \subseteq \mathbb{R}^m$ is a closed uncertainty set, then the moment ambiguity set $\mathcal{P}$ defined in (2.1) is weakly compact.

Proof. As $\mathcal{Z}$ is compact, the support-only ambiguity set $\mathcal{P}(\mathcal{Z})$ is weakly compact by virtue of Proposition 3.6. Consequently, $\mathcal{P}(\mathcal{Z})$ is tight and weakly closed. This readily implies that $\mathcal{P}$ is tight as a subset of a tight set remains tight. Proposition 3.3 further implies that $\mathbb{E}_{\mathbb{P}}[f(Z)]$ is weakly continuous in $\mathbb{P}$. As $\mathcal{F}$ is closed and as the pre-image of any closed set under a continuous transformation is closed, we may conclude that $\mathcal{P}_f = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)\colon \mathbb{E}_{\mathbb{P}}[f(Z)] \in \mathcal{F}\}$ is weakly closed. Hence $\mathcal{P} = \mathcal{P}(\mathcal{Z}) \cap \mathcal{P}_f$ is weakly closed as the intersection of two weakly closed sets. Given these insights, the claim follows readily from Theorem 3.5. □

The conditions of Proposition 3.7 are only sufficient but not necessary for weak compactness. The next examples show that moment ambiguity sets can be tight or weakly compact even if the support set $\mathcal{Z}$ or the moment function f are unbounded.

Example 3.8 (Markov ambiguity sets). The Markov ambiguity set (2.2) fails to be tight if $\mathcal{Z} = \mathbb{R}^d$. For example, if $\mathcal{Z} = \mathbb{R}$ and $\mu = 0$, then for every compact set $\mathcal{C} \subseteq \mathbb{R}$ there is a constant $R > 0$ such that the two-point distribution $\mathbb{P} = \frac{1}{2}\delta_{-R} + \frac{1}{2}\delta_R$ is fully supported on the complement of $\mathcal{C}$. However, the Markov ambiguity set $\mathcal{P}$ becomes tight if $\mathcal{Z} = \mathbb{R}_+$ and $\mu = 1$. Indeed, in this case Markov's inequality implies that $\mathbb{P}(Z \notin \mathcal{C}) \leq \varepsilon$ for every $\mathbb{P} \in \mathcal{P}$ and $\varepsilon > 0$ if we define $\mathcal{C}$ as the compact interval $[0, 1/\varepsilon]$. Even in this case, however, $\mathcal{P}$ fails to be weakly closed. Indeed, the distributions

$$\mathbb{P}_i = \frac{i}{i+1}\delta_0 + \frac{1}{i+1}\delta_{i+1}$$

belong to $\mathcal{P}$ for all $i \in \mathbb{N}$, but their weak limit $\mathbb{P} = \delta_0$ is no member of $\mathcal{P}$. If $\mathcal{Z}$ is convex, one can extend this reasoning in the obvious way to show that $\mathcal{P}$ is weakly compact if and only if $\mathcal{Z}$ is compact.

The next example shows that Chebyshev ambiguity sets are tight irrespective of $\mathcal{Z}$. Nevertheless, they are not always weakly compact.

Example 3.9 (Chebyshev ambiguity sets). The Chebyshev ambiguity set $\mathcal{P}$ defined in (2.3) is always tight. To see this, assume without loss of generality that $\mu = 0$ and $M = I_d$, which can always be enforced by applying an affine coordinate transformation. Given any $\varepsilon > 0$, we can define a compact set $\mathcal{C} = \{z \in \mathcal{Z}\colon \|z\|_2 \leq \sqrt{d/\varepsilon}\}$. It is then easy to see that any distribution $\mathbb{P} \in \mathcal{P}$ satisfies

$$\mathbb{P}(Z \notin \mathcal{C}) = \mathbb{P}\big(\|Z\|_2 > \sqrt{d/\varepsilon}\big) \leq \mathbb{E}_{\mathbb{P}}\big[\|Z\|_2^2 \cdot \varepsilon/d\big] = \varepsilon,$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where the inequality holds because the quadratic function $q(z) = \|z\|_2^2 \cdot \varepsilon/d$ majorizes the characteristic function of $\mathcal{Z}\backslash\mathcal{C}$. Hence $\mathcal{P}$ is indeed tight. However, $\mathcal{P}$ is not necessarily weakly closed. To see this, suppose that $d = 1$ and that $\mathcal{Z} = \mathbb{R}$. In this case the distributions

$$\mathbb{P}_i = \frac{1}{2i^2}\delta_{-i} + \frac{i^2-1}{i^2}\delta_0 + \frac{1}{2i^2}\delta_i$$

have zero mean and unit variance for all $i \in \mathbb{N}$. That is, they all belong to $\mathcal{P}$. However, they converge weakly to $\mathbb{P} = \delta_0$, which is not an element of $\mathcal{P}$. Thus $\mathcal{P}$ fails to be weakly compact.

The family of all distributions on $\mathbb{R}^d$ with bounded pth-order moments is always weakly compact even though ambiguity sets that fix the pth-order moments to prescribed values (e.g. the Chebyshev ambiguity set) may *not* be weakly compact.

Example 3.10 (pth-order moment ambiguity sets). The ambiguity set

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{E}_{\mathbb{P}}[\|Z\|^p] \leq R\}$$

induced by any norm $\|\cdot\|$ on $\mathbb{R}^d$ and two parameters $p, R > 0$ is weakly compact. Using reasoning similar to Example 3.9, one can show that for any $\varepsilon > 0$ there exists a compact set, namely $\mathcal{C} = \{z \in \mathcal{Z} \colon \|z\| \leq (R/\varepsilon)^{1/p}\}$, which satisfies $\mathbb{P}(Z \notin \mathcal{C}) \leq \varepsilon$. Thus $\mathcal{P}$ is tight. To see that $\mathcal{P}$ is also weakly closed, note that $f(z) = \|z\|^p$ is continuous and bounded below. By Proposition 3.3, the expected value $\mathbb{E}_{\mathbb{P}}[\|Z\|^p]$ is therefore weakly lower semicontinuous in $\mathbb{P}$ and has weakly closed sublevel sets. Therefore $\mathcal{P}$ is weakly compact by virtue of Theorem 3.5.

3.2. ϕ-divergence ambiguity sets

In this section we show that ϕ-divergence ambiguity sets of the form (2.10) are weakly compact whenever the entropy function ϕ grows superlinearly. Otherwise, if ϕ grows at most linearly, then the corresponding ϕ-divergence ambiguity sets generically fail to be weakly compact. Recall that an entropy function ϕ in the sense of Definition 2.4 grows superlinearly if and only if $\phi^\infty(1) = \infty$; see also Table 2.1.

Lemma 3.11 (Worst-case probability maps). Let $\mathcal{P}$ be the ϕ-divergence ambiguity set of radius $r > 0$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ defined in (2.10), and assume that ϕ is continuous at 1 and that $\phi^\infty(1) = \infty$. Then there is a continuous, concave and surjective function $p \colon [0,1] \to [0,1]$ that depends only on ϕ and r such that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(Z \in \mathcal{B}) = p(\hat{\mathbb{P}}(Z \in \mathcal{B}))$$

for every Borel set $\mathcal{B} \subseteq \mathcal{Z}$.

Proof. The proof is constructive. That is, we define the function p through

$$p(t) = \inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \lambda_0 + \lambda r + t \cdot (\phi^*)^\pi(1-\lambda_0, \lambda) + (1-t) \cdot (\phi^*)^\pi(-\lambda_0, \lambda)$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

for all $t \in [0,1]$. In the remainder we show that p satisfies all desired properties. By construction, p depends only on ϕ and r and coincides with the lower envelope of infinitely many linear functions in t. Hence p is concave as well as upper semicontinuous. By the definition of $\mathcal{P}$ and by Theorem 4.15 below, we also have

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(Z \in \mathcal{B}) &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\mathbb{1}_{\mathcal{B}}(Z)] : \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq r\} \\ &= \inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \lambda_0 + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\mathbb{1}_{\mathcal{B}}(Z) - \lambda_0, \lambda)] \\ &= p(\hat{\mathbb{P}}(Z \in \mathcal{B})), \end{aligned} \tag{3.1}$$

for any Borel set $\mathcal{B}$, where the last equality follows from the definition of p. As the worst-case probability on the left-hand side of (3.1) falls within $[0,1]$ and as $\hat{\mathbb{P}}(Z \in \mathcal{B})$ can adopt any value in $[0,1]$, it is clear that the range of p is a subset of $[0,1]$. Next, we show that p is continuous. To this end, note that the concavity and finiteness of p on $[0,1]$ imply via Rockafellar (1970, Theorem 10.1) that p is continuous on $(0,1)$. In addition, its upper semicontinuity prevents p from jumping at 0 or at 1. Thus p is indeed continuous throughout $[0,1]$. Finally, setting $\mathcal{B} = \emptyset$ or $\mathcal{B} = \mathcal{Z}$ in (3.1) shows that $p(0) = 0$ and $p(1) = 1$, respectively. Consequently, we may conclude that p is surjective. This observation completes the proof. □

As $\hat{\mathbb{P}} \in \mathcal{P}$, the worst-case probability map p from Lemma 3.11 satisfies $p(t) \geq t$ for all $t \in [0,1]$, that is, the worst-case probability is never smaller than the nominal probability. We remark that the map p also emerges in the study of distributionally robust chance constraints over ϕ-divergence ambiguity sets with $\phi^\infty(1) = \infty$. Indeed, any such distributionally robust chance constraint with violation probability $\varepsilon \in (0,1)$ is equivalent to a classical chance constraint under the reference distribution $\hat{\mathbb{P}}$ with (smaller) violation probability $p^{-1}(\varepsilon)$; see El Ghaoui *et al.* (2003), Jiang and Guan (2016) and Shapiro (2017). We can now show that divergence ambiguity sets corresponding to superlinear entropy functions are weakly compact.

Proposition 3.12 (ϕ-divergence ambiguity sets). If ϕ is an entropy function with $\phi^\infty(1) = \infty$, then the corresponding ϕ-divergence ambiguity set $\mathcal{P}$ defined in (2.10) is weakly compact for any closed set $\mathcal{Z} \subseteq \mathbb{R}^d$, distribution $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ and $r \geq 0$.

Proof. We first show that $\mathcal{P}$ is tight. To this end, select any $\varepsilon \in (0,1)$, and define $p^{-1}(\varepsilon)$ as the unique $t \in (0,1]$ satisfying $p(t) = \varepsilon$, where p represents the worst-case probability map from Lemma 3.11. Note that $p^{-1}(\varepsilon)$ is well-defined because p is concave and surjective and because $p(0) = 0$ and $p(1) = 1$. Note also that $p^{-1}(\varepsilon) \leq \varepsilon$ because $p(t) \geq t$. Next, select a sufficiently large $R > 0$ such that $\hat{\mathbb{P}}(\|Z\|_2 > R) \leq p^{-1}(\varepsilon)$, and define a compact set $\mathcal{C} = \{z \in \mathcal{Z} : \|z\|_2 \leq R\}$. Lemma 3.11 applied to $\mathcal{B} = \mathcal{Z}\backslash\mathcal{C}$ then allows us to conclude that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(Z \notin \mathcal{C}) = p(\hat{\mathbb{P}}(Z \notin \mathcal{C})) \leq p(p^{-1}(\varepsilon)) = \varepsilon,$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where the inequality follows from the monotonicity of p and choice of R. We have thus shown that $\mathbb{P}(Z \notin \mathcal{C}) \leq \varepsilon$ for all $\mathbb{P} \in \mathcal{P}$, and thus $\mathcal{P}$ is tight.

It remains to be shown that $\mathcal{P}$ is weakly closed. To this end, recall first that $\mathcal{P}(\mathcal{Z})$ is weakly closed because $\mathcal{Z}$ is closed; see Proposition 3.6. Next, recall from Proposition 2.6 that any ϕ-divergence admits a dual representation of the form

$$\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \phi^*(f(z))\, \mathrm{d}\hat{\mathbb{P}}(z), \tag{3.2}$$

where $\mathcal{F}$ denotes the family of all bounded Borel functions $f\colon \mathcal{Z} \to \mathrm{dom}(\phi^*)$. In fact, $\mathcal{F}$ can be restricted to the space $\mathcal{F}^c$ of all *continuous* bounded functions without reducing the supremum in (3.2). This is a direct consequence of Lusin's theorem, which ensures that for any $\delta > 0$ and $f \in \mathcal{F}$ there exists a compact set $\mathcal{A} \subseteq \mathcal{Z}$ with $\hat{\mathbb{P}}(Z \notin \mathcal{A}) \leq \delta$ and a bounded continuous function $f_\delta \in \mathcal{F}^c$ that coincides with f on $\mathcal{A}$ and satisfies $\sup_{z \in \mathcal{Z}} |f_\delta(z)| \leq \sup_{z \in \mathcal{Z}} |f(z)| = \|f\|_\infty$. As the convex lower semicontinuous function ϕ^* is continuous on its domain, both

$$\phi_l^* = \inf_{s \in \mathrm{dom}(\phi^*)} \{\phi^*(s)\colon |s| \leq \|f\|_\infty\} \quad \text{and} \quad \phi_u^* = \sup_{s \in \mathrm{dom}(\phi^*)} \{\phi^*(s)\colon |s| \leq \|f\|_\infty\}$$

are finite. Therefore we have

$$\begin{aligned}
&\int_{\mathcal{Z}} f_\delta(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \phi^*(f_\delta(z))\, \mathrm{d}\hat{\mathbb{P}}(z) \\
&\geq \int_{\mathcal{Z}} f(z)\, \mathrm{d}\mathbb{P}(z) - \int_{\mathcal{Z}} \phi^*(f(z))\, \mathrm{d}\hat{\mathbb{P}}(z) - 2\|f\|_\infty\, \mathbb{P}(Z \notin \mathcal{A}) - (\phi_u^* - \phi_l^*)\, \hat{\mathbb{P}}(Z \notin \mathcal{A}).
\end{aligned}$$

As $\phi^\infty(1) = \infty$ implies $\mathbb{P} \ll \hat{\mathbb{P}}$ and as $\hat{\mathbb{P}}(Z \notin \mathcal{Z}) \leq \delta$, both $\mathbb{P}(Z \notin \mathcal{A})$ and $\hat{\mathbb{P}}(Z \notin \mathcal{A})$ decay to 0 as δ is reduced. Thus the objective function value of f_δ in problem (3.2) is asymptotically non-inferior to that of f. This confirms that restricting $\mathcal{F}$ to $\mathcal{F}^c$ has no impact on the supremum in (3.2). Recall now from Proposition 3.3 that, for any bounded continuous function $f \in \mathcal{F}^c$, the first integral in (3.2) is weakly continuous in $\mathbb{P}$. Thus $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}})$ is weakly lower semicontinuous in $\mathbb{P}$ as a pointwise supremum of weakly continuous functions. This implies that any sublevel set of the function $f(\mathbb{P}) = \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}})$ is weakly closed. We thus conclude that the divergence ambiguity set is weakly closed. The claim then follows from Theorem 3.5. □

The proof of Proposition 3.12 critically relies on the assumption that $\phi^\infty(1) = \infty$, which ensures that the divergence ambiguity set contains only distributions that are absolutely continuous with respect to $\hat{\mathbb{P}}$. Below we show that if the entropy function ϕ grows at most linearly (i.e. if $\phi^\infty(1) < \infty$) and $\mathcal{Z}$ is unbounded, then the corresponding divergence ambiguity set fails to be weakly compact. As a preparation, we first establish an upper bound on any ϕ-divergence on $\mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z})$.

Lemma 3.13 (Upper bounds on ϕ-divergences). If ϕ is an entropy function and $\mathcal{Z} \subseteq \mathbb{R}^d$ a closed set, then we have $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq \phi(0) + \phi^\infty(1)$ for all $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$. This upper bound is attained if $\mathbb{P}$ and $\hat{\mathbb{P}}$ are mutually singular, that is, if $\mathbb{P} \perp \hat{\mathbb{P}}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Proof. In the first part of the proof we derive the desired upper bound. To this end, assume that $\phi(0) < \infty$ and $\phi^\infty(1) < \infty$, for otherwise the upper bound is trivially satisfied. As the entropy function is convex, we then have

$$\phi(s) \leq \frac{\Delta}{s+\Delta}\phi(0) + \frac{s}{s+\Delta}\phi(s+\Delta) \iff \phi(s) \leq \phi(0) + s\,\frac{\phi(s+\Delta)-\phi(0)}{s+\Delta}$$

for every $s, \Delta \geq 0$. Letting Δ tend to infinity, this implies that $\phi(s) \leq \phi(0) + s\,\phi^\infty(1)$ for all $s \geq 0$. The ϕ-divergence between any $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ thus satisfies

$$\begin{aligned}
\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) &= \int_{\mathcal{Z}} \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\, \phi\left(\frac{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)}{\frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)}\right) \mathrm{d}\rho(z) \\
&\leq \int_{\mathcal{Z}} \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\, \phi(0)\, \mathrm{d}\rho(z) + \int_{\mathcal{Z}} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)\, \phi^\infty(1)\, \mathrm{d}\rho(z) \\
&= \phi(0) + \phi^\infty(1),
\end{aligned}$$

where we may assume without loss of generality that the dominating measure $\rho \in \mathcal{M}_+(\mathcal{Z})$ is given by $\rho = \mathbb{P} + \hat{\mathbb{P}}$. This establishes the desired upper bound. It remains to be shown that this bound is attained even if $\phi(0)$ or $\phi^\infty(1)$ evaluate to infinity. To this end, suppose that $\mathbb{P}$ and $\hat{\mathbb{P}}$ are mutually singular. This means that there exist disjoint Borel sets $\mathcal{B}, \hat{\mathcal{B}} \subseteq \mathcal{Z}$ with $\mathbb{P}(Z \in \mathcal{B}) = 1$ and $\hat{\mathbb{P}}(Z \in \hat{\mathcal{B}}) = 1$. We thus have

$$\begin{aligned}
\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) &= \int_{\hat{\mathcal{B}}} \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\, \phi(0)\, \mathrm{d}\rho(z) + \int_{\mathcal{B}} 0\, \phi\left(\frac{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)}{0}\right) \mathrm{d}\rho(z) \\
&= \phi(0) + \int_{\mathcal{B}} \phi^\infty\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z)\right) \mathrm{d}\rho(z) \\
&= \phi(0) + \phi^\infty(1).
\end{aligned}$$

The first equality holds because $\mathrm{d}\mathbb{P}/\mathrm{d}\rho(z) = 0$ for ρ-almost all $z \in \hat{\mathcal{B}}$ and $\mathrm{d}\hat{\mathbb{P}}/\mathrm{d}\rho(z) = 0$ for ρ-almost all $z \in \mathcal{B}$. The second equality follows from the definition of the perspective function and exploits that the restriction of ρ to $\hat{\mathcal{B}}$ coincides with $\hat{\mathbb{P}}$. The third equality, finally, holds because the restriction of ρ to $\mathcal{B}$ coincides with $\mathbb{P}$. Note that the upper bound is attained even if $\phi(0) = \infty$ or $\phi^\infty(1) = \infty$. □

The following example reveals that ϕ-divergence ambiguity sets fail to be weakly compact if $\phi^\infty(1) < \infty$ and if the set $\mathcal{Z}$ without the atoms of $\hat{\mathbb{P}}$ is unbounded.

Example 3.14 (ϕ-divergence ambiguity sets). Consider an entropy function ϕ with $\phi^\infty(1) < \infty$. By Lemma 3.13, $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}})$ is bounded above by $\overline{r} = \phi(0) + \phi^\infty(1)$ for all $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}$. In addition, let $\mathcal{P}$ be the ϕ-divergence ambiguity set with centre $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ and radius $r \in (0, \overline{r})$ defined in (2.10). Assume that for every $R > 0$ there exists $z_0 \in \mathcal{Z}$ with $\|z_0\|_2 \geq R$ and $\hat{\mathbb{P}}(Z = z_0) = 0$. This assumption holds, for example, whenever $\mathcal{Z}$ is unbounded and convex, and it implies that $\mathcal{P}$ fails to be tight. To see this, fix an arbitrary compact set $\mathcal{C} \subseteq \mathcal{Z}$, and select any point

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$z_0 \in \mathcal{Z}\backslash\mathcal{C}$ with $\hat{\mathbb{P}}(Z = z_0) = 0$. Such a point exists by assumption. Next, consider the distributions $\mathbb{P}_\theta = (1-\theta)\hat{\mathbb{P}} + \theta\,\delta_{z_0}$ parametrized by $\theta \in [0,1]$. Note that $\hat{\mathbb{P}}$ and δ_{z_0} are mutually singular and that $f(\theta) = \mathrm{D}_\phi(\mathbb{P}_\theta, \hat{\mathbb{P}})$ is a convex continuous bijective function from $[0,1]$ to $[0,\overline{r}]$. Set now $\varepsilon = \frac{1}{2}f^{-1}(r)$. For $\theta = f^{-1}(r)$, the distribution $\mathbb{P}_\theta$ satisfies $\mathrm{D}_\phi(\mathbb{P}_\theta, \hat{\mathbb{P}}) = f(f^{-1}(r)) = r$ and thus belongs to $\mathcal{P}$. In addition, $\mathbb{P}_\theta(Z \notin \mathcal{C}) \geq f^{-1}(r) > \varepsilon$ because $z_0 \notin \mathcal{C}$. Note that ε is independent of $\mathcal{C}$ and z_0 as long as $\hat{\mathbb{P}}(Z = z_0) = 0$. As the compact set $\mathcal{C}$ was chosen arbitrarily, this implies that $\mathcal{P}$ fails to be tight and weakly compact.

3.3. Marginal ambiguity sets

As a preparation towards exploring the topological properties of optimal transport ambiguity sets, we first study marginal ambiguity sets. The following proposition shows that Fréchet ambiguity sets, which prescribe the marginal distributions of all d individual components of Z, are always weakly compact.

Proposition 3.15 (Fréchet ambiguity sets). The Fréchet ambiguity set $\mathcal{P}$ defined in (2.35) is weakly compact for any cumulative distribution functions F_i, $i \in [d]$.

Proof. We first show that the Fréchet ambiguity set is tight. For any $\varepsilon > 0$ and $i \in [d]$, we can set $\underline{z}_i$ and $\overline{z}_i$ to the $\varepsilon/(2d)$-quantile and the $(1-\varepsilon/(2d))$-quantile of the distribution function F_i, respectively. Setting $\mathcal{C} = \times_{i\in[d]}[\underline{z}_i, \overline{z}_i]$ yields

$$\mathbb{P}(Z \notin \mathcal{C}) \leq \sum_{i\in[d]} \mathbb{P}\big(Z_i \notin [\underline{z}_i, \overline{z}_i]\big) = \sum_{i\in[d]} \varepsilon/d = \varepsilon,$$

where the inequality follows from the union bound. Thus $\mathcal{P}$ is tight. It remains to be shown that $\mathcal{P}$ is weakly closed. Note that the distribution function of Z_i under $\mathbb{P}$ matches F_i if and only if, for every bounded continuous function f, we have

$$\mathbb{E}_{\mathbb{P}}[f(Z_i)] = \int_{-\infty}^{+\infty} f(z_i)\,\mathrm{d}F_i(z_i).$$

This is true because every Borel distribution on $\mathbb{R}$ constitutes a Radon measure. The set of all $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ satisfying the above equality for any fixed bounded and continuous function f and any fixed index $i \in [d]$ is weakly closed by Proposition 3.3. Hence $\mathcal{P}$ is weakly closed because closedness is preserved by intersection. □

It is straightforward to generalize Proposition 3.15 from Fréchet ambiguity sets to generic marginal ambiguity sets as discussed in Section 2.4.1, which prescribe *multivariate* marginal distributions. Details are omitted for brevity.

3.4. Optimal transport ambiguity sets

Recall that $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ denotes the family of all transportation plans linking the probability distributions $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$. Thus $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ contains all joint distributions

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

γ of Z and $\hat{Z}$ with marginals $\mathbb{P}$ and $\hat{\mathbb{P}}$, respectively. The set $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ appears in the definition of the optimal transport discrepancy $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}})$; see Definition 2.15. The reasoning in Section 3.3 immediately implies that $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ is weakly compact because it constitutes a marginal ambiguity set. This insight is formalized in the following simple corollary of Proposition 3.15. Its proof is omitted for brevity.

Corollary 3.16 (Transportation plans). The set of all transportation plans $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ with marginal distributions $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is weakly compact.

Corollary 3.16 enables us to show that the optimal transport problem in (2.18) is solvable as the transportation cost function is assumed to be lower semicontinuous.

Lemma 3.17 (Solvability of optimal transport problems). The infimum in (2.18) is attained.

Proof. By Corollary 3.16, the set $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ is weakly compact. In addition, the transportation cost function $c(z, \hat{z})$ is lower semicontinuous and bounded below. By Proposition 3.3, the expected value $\mathbb{E}_\gamma[c(Z, \hat{Z})]$ is therefore weakly lower semicontinuous in γ. Thus the optimal transport problem in (2.18) is solvable thanks to Weierstrass's theorem, and its infimum is attained. □

Lemma 3.17 allows us to prove that the optimal transport discrepancy $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}})$ constitutes a weakly lower semicontinuous function of its inputs $\mathbb{P}$ and $\hat{\mathbb{P}}$.

Lemma 3.18 (Weak lower semicontinuity of optimal transport discrepancies). The optimal transport discrepancy $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}})$ is weakly lower semicontinuous jointly in $\mathbb{P}$ and $\hat{\mathbb{P}}$.

Proof. Assume that $\mathbb{P}_j$ and $\hat{\mathbb{P}}_j$, $j \in \mathbb{N}$, converge weakly to $\mathbb{P}$ and $\hat{\mathbb{P}}$, respectively, and define the countable ambiguity sets $\mathcal{P} = \{\mathbb{P}_j\}_{j\in\mathbb{N}}$ and $\hat{\mathcal{P}} = \{\hat{\mathbb{P}}_j\}_{j\in\mathbb{N}}$. By the definition of sequential compactness, the weak closures of $\mathcal{P}$ and $\hat{\mathcal{P}}$ are weakly compact. Prokhorov's theorem (see Theorem 3.5) thus implies that both $\mathcal{P}$ and $\hat{\mathcal{P}}$ are tight. Hence, for any $\varepsilon > 0$ there exist two compact sets $\mathcal{C}, \hat{\mathcal{C}} \subseteq \mathbb{R}^d$ with

$$\mathbb{P}_j(Z \notin \mathcal{C}) \leq \varepsilon/2 \quad \text{and} \quad \hat{\mathbb{P}}_j(\hat{Z} \notin \hat{\mathcal{C}}) \leq \varepsilon/2 \quad \text{for all } j \in \mathbb{N}.$$

Whenever $\gamma \in \Gamma(\mathbb{P}_j, \hat{\mathbb{P}}_j)$ for some $j \in \mathbb{N}$, we thus have

$$\gamma((Z, \hat{Z}) \notin \mathcal{C} \times \hat{\mathcal{C}}) \leq \mathbb{P}_j(Z \notin \mathcal{C}) + \hat{\mathbb{P}}_j(Z \notin \hat{\mathcal{C}}) \leq \varepsilon.$$

As $\mathcal{C} \times \hat{\mathcal{C}}$ is compact and as ε was chosen arbitrarily, this reveals that the union

$$\bigcup_{j\in\mathbb{N}} \Gamma(\mathbb{P}_j, \hat{\mathbb{P}}_j) \tag{3.3}$$

is tight, which in turn implies via Prokhorov's theorem that its closure is weakly compact. Now let $\gamma_j^\star$ be an optimal coupling of $\mathbb{P}_j$ and $\hat{\mathbb{P}}_j$, which solves problem (2.18), and which exists thanks to Lemma 3.17. As all these optimal couplings

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

belong to some weakly compact set (i.e. the weak closure of (3.3)), we may assume without loss of generality that $\gamma_j^\star$, $j \in \mathbb{N}$, converges weakly to some distribution γ. Otherwise, we can pass to a subsequence. Clearly we have $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$. For $\gamma^\star$ an optimal coupling of $\mathbb{P}$ and $\hat{\mathbb{P}}$, we then find

$$\begin{aligned}\liminf_{j\to\infty} \mathrm{OT}_c(\mathbb{P}_j, \hat{\mathbb{P}}_j) &= \liminf_{j\to\infty} \mathbb{E}_{\gamma_j^\star}[c(Z,\hat{Z})] \\ &\geq \mathbb{E}_{\gamma}[c(Z,\hat{Z})] \\ &\geq \mathbb{E}_{\gamma^\star}[c(Z,\hat{Z})] \\ &= \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}),\end{aligned}$$

where the two equalities follow from the definitions of $\gamma_j^\star$ and $\gamma^\star$, respectively. The first inequality holds because $\mathbb{E}_\gamma[c(Z,\hat{Z})]$ is weakly lower semicontinuous in γ thanks to Proposition 3.3, and the second inequality follows from the suboptimality of γ in (2.18). Thus $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}})$ is weakly lower semicontinuous in $\mathbb{P}$ and $\hat{\mathbb{P}}$. □

Lemma 3.18 is inspired by Clément and Desch (2008, Lemma 5.2) and Yue, Kuhn and Wiesemann (2022, Theorem 1). Next, we prove that Wasserstein ambiguity sets are weakly compact. Throughout this discussion we assume that the metric underlying the transportation cost function is induced by a norm $\|\cdot\|$ on $\mathbb{R}^d$. This assumption simplifies our derivations but could be relaxed. Recall that the p-Wasserstein distance $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})$ for $p \geq 1$ is the pth root of $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}})$, where the transportation cost function is set to $c(z,\hat{z}) = \|z - \hat{z}\|^p$; see Definition 2.18.

Theorem 3.19 (p-Wasserstein ambiguity sets). Assume that the metric $d(\cdot,\cdot)$ on $\mathcal{Z}$ is induced by some norm $\|\cdot\|$ on the ambient space $\mathbb{R}^d$. If $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ has finite pth moments (i.e. $\mathbb{E}_{\hat{\mathbb{P}}}[\|Z\|^p] < \infty$) for some exponent $p \geq 1$, then the p-Wasserstein ambiguity set $\mathcal{P}$ defined in (2.28) is weakly compact.

Proof. We first show that all distributions $\mathbb{P} \in \mathcal{P}$ have uniformly bounded pth moments. To this end, set $\hat{r} = \mathbb{E}_{\hat{\mathbb{P}}}[\|Z\|^p] < \infty$, and note that any $\mathbb{P} \in \mathcal{P}$ satisfies

$$\begin{aligned}(\mathbb{E}_{\mathbb{P}}[\|Z\|^p])^{1/p} &= \mathrm{W}_p(\mathbb{P}, \delta_0) \\ &\leq \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) + \mathrm{W}_p(\hat{\mathbb{P}}, \delta_0) \\ &= \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) + (\mathbb{E}_{\hat{\mathbb{P}}}[\|Z\|^p])^{1/p} \\ &\leq r + \hat{r}.\end{aligned}$$

Here the first inequality holds because the p-Wasserstein distance is a metric and thus satisfies the triangle inequality, and the second inequality holds because $\mathbb{P} \in \mathcal{P}$. We therefore have $\mathbb{E}_{\mathbb{P}}[\|Z\|^p] \leq (r + \hat{r})^p$ for every $\mathbb{P} \in \mathcal{P}$. In other words, the Wasserstein ball $\mathcal{P}$ is a subset of the pth-order moment ambiguity set discussed in Example 3.10. This implies that $\mathcal{P}$ is tight. Note further that $\mathcal{P}$ is defined as a sublevel set of the function $f(\mathbb{P}) = \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})$, which is weakly lower semicontinuous thanks to Lemma 3.18. Hence $\mathcal{P}$ is weakly closed. □

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Finally, we prove that the ∞-Wasserstein ambiguity set is always weakly compact.

Corollary 3.20 (∞-Wasserstein ambiguity sets). Assume that the metric $d(\cdot,\cdot)$ on $\mathcal{Z}$ is induced by some norm $\|\cdot\|$ on the ambient space $\mathbb{R}^d$. Then the ∞-Wasserstein ambiguity set defined in (2.34) is weakly compact for every $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$.

Proof. We first show that $\mathcal{P}$ is tight. To this end, select any $\varepsilon > 0$ and any compact set $\hat{\mathcal{C}} \subseteq \mathcal{Z}$ with $\hat{\mathbb{P}}(Z \notin \hat{\mathcal{C}}) \leq \varepsilon$. Note that $\hat{\mathcal{C}}$ is guaranteed to exist because $\hat{\mathbb{P}}$ is a probability distribution. Next, define $\mathcal{C}$ as the r-neighbourhood $\hat{\mathcal{C}}_r$ of $\hat{\mathcal{C}}$, that is, set

$$\mathcal{C} = \{z \in \mathcal{Z} \colon \exists \hat{z} \in \hat{\mathcal{C}} \text{ with } \|z - \hat{z}\| \leq r\}.$$

See also (2.29). One readily verifies that $\mathcal{C}$ inherits compactness from $\hat{\mathcal{C}}$. Any distribution $\mathbb{P} \in \mathcal{P}$ satisfies $\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) \leq r$. Consequently, we find

$$\mathbb{P}(Z \notin \mathcal{C}) = \mathbb{P}(Z \in \mathcal{Z}\backslash\mathcal{C}) \leq \hat{\mathbb{P}}(Z \in \mathcal{Z}\backslash\hat{\mathcal{C}}) = \hat{\mathbb{P}}(Z \notin \hat{\mathcal{C}}) \leq \varepsilon,$$

where the first inequality follows from Corollary 2.28 and the observation that the r-neighbourhood of $\mathcal{Z}\backslash\mathcal{C}$ coincides with $\mathcal{Z}\backslash\hat{\mathcal{C}}$. The second inequality follows from the definition of $\hat{\mathcal{C}}$. As ε was chosen arbitrarily, $\mathcal{P}$ is tight. It remains to be shown that $\mathcal{P}$ is weakly closed. Proposition 2.26 readily implies that $\mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) \leq r$ if and only if $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) \leq r$ for all $p \geq 1$. Thus we may conclude that

$$\mathcal{P} = \bigcap_{p \geq 1} \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d) \colon \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

That is, the ∞-Wasserstein ambiguity set can be expressed as the intersection of all p-Wasserstein ambiguity sets for $p \geq 1$, all of which are weakly closed by Theorem 3.19. Hence $\mathcal{P}$ is indeed weakly closed, and the claim follows. □

4. Duality theory for worst-case expectation problems

The DRO problem (1.2) is often interpreted as a zero-sum game between the decision-maker and a fictitious adversary. The decision-maker moves first and thus selects x *before* seeing $\mathbb{P}$. Therefore x is optimized against *all* distributions $\mathbb{P} \in \mathcal{P}$. In contrast, the adversary moves second and thus selects $\mathbb{P}$ *after* seeing x. Therefore $\mathbb{P}$ is only optimized against *one particular* decision $x \in \mathcal{X}$. Put differently, the adversary's choice may adapt to the decision-maker's choice but *not* vice versa.

In this section we develop a duality theory for the adversary's subproblem, which aims to maximize the expected loss of a fixed decision x across all distributions in a convex ambiguity set $\mathcal{P}$. To avoid clutter, we suppress the dependence of the loss function ℓ on the fixed decision x throughout this discussion, that is, we write $\ell(z)$ instead of $\ell(x, z)$. We thus address worst-case expectation problems of the form

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)]. \tag{4.1}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Note that $\mathcal{P}$ represents a convex subset of the linear space of all finite signed Borel measures on $\mathcal{Z}$. Unless $\mathcal{Z}$ is finite, (4.1) thus constitutes an infinite-dimensional convex program with a linear objective function. For this problem to be well-defined, we assume that $\ell\colon \mathcal{Z} \to \overline{\mathbb{R}}$ is a Borel function. In line with Rockafellar and Wets (2009, Section 14.E), we define $\mathbb{E}_{\mathbb{P}}[\ell(Z)] = -\infty$ if $\mathbb{E}_{\mathbb{P}}[\max\{\ell(Z), 0\}] = \infty$ and $\mathbb{E}_{\mathbb{P}}[\min\{\ell(Z), 0\}] = -\infty$. This means that infeasibility trumps unboundedness. More generally, throughout the rest of the paper, we assume that if the objective function of a minimization (maximization) problem can be expressed as the difference of two terms, both of which evaluate to ∞, then the objective function value should be interpreted as ∞ ($-\infty$). This convention is in line with the rules of extended arithmetic used by Rockafellar and Wets (2009).

In the remainder we will show that (4.1) can be dualized by using elementary tools from finite-dimensional convex analysis (Fenchel 1953, Rockafellar 1970) for a broad class of finitely-parametrized ambiguity sets including all moment ambiguity sets (Section 4.2), ϕ-divergence ambiguity sets (Section 4.3) and optimal transport ambiguity sets (Section 4.4). We broadly adopt the proof strategies developed by Shapiro (2001) and Zhang *et al.* (2024*b*) for moment and optimal transport ambiguity sets, respectively, and we extend them to ϕ-divergence ambiguity sets.

4.1. General proof strategy

In order to outline the high-level ideas for dualizing (4.1), we recall a basic result on the convexity of parametric infima; see e.g. Rockafellar (1974, Theorem 1).

Lemma 4.1 (Convexity of optimal value functions). If $\mathcal{U}$ and $\mathcal{V}$ are arbitrary real vector spaces and $H\colon \mathcal{U} \times \mathcal{V} \to \overline{\mathbb{R}}$ is a convex function, then the optimal value function $h\colon \mathcal{U} \to \overline{\mathbb{R}}$ defined by $h(u) = \inf_{v\in\mathcal{V}} H(u, v)$ is convex.

Proof. Note that h is a convex function if and only if its epigraph $\mathrm{epi}(h)$ is a convex set. By the definitions of the epigraph and the infimum operator, we find

$$\begin{aligned}\mathrm{epi}(h) &= \{(u,t) \in \mathcal{U} \times \mathbb{R}\colon\ h(u) \le t\} \\ &= \{(u,t) \in \mathcal{U} \times \mathbb{R}\colon\ \exists v \in \mathcal{V} \ \text{ with } \ H(u,v) \le t + \varepsilon \ \ \forall \varepsilon > 0\} \\ &= \bigcap_{\varepsilon>0} \{(u,t) \in \mathcal{U} \times \mathbb{R}\colon\ \exists v \in \mathcal{V} \ \text{ with } \ H(u,v) - \varepsilon \le t\}.\end{aligned}$$

Thus $\mathrm{epi}(h)$ can be obtained by projecting $\cap_{\varepsilon>0} \mathrm{epi}(H-\varepsilon)$ to $\mathcal{U}\times\mathbb{R}$. The claim then follows because $\mathrm{epi}(H - \varepsilon)$ is convex for every $\varepsilon > 0$ thanks to the convexity of H and because convexity is preserved under intersections and linear transformations; see e.g. Rockafellar (1970, Theorems 2.1, 5.7). □

The following result marks a cornerstone of convex analysis. It states that the biconjugate h^{**} (i.e. the conjugate of h^*) of a closed convex function h coincides with h. Here we adopt the standard convention that h is closed if it is lower semi-continuous and either $h(u) > -\infty$ for all $u \in \mathcal{U}$ or $h(u) = -\infty$ for all $u \in \mathcal{U}$. We use $\mathrm{cl}(h)$ to denote the closure of h, that is, the largest closed function below h.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Lemma 4.2 (Fenchel–Moreau Theorem). For any convex function $h\colon \mathbb{R}^d \to \overline{\mathbb{R}}$, we have $h \geq h^{**}$. The inequality becomes an equality on $\operatorname{rint}(\operatorname{dom}(h))$.

Proof. By Rockafellar (1970, Theorem 12.2), we have $h^{**} = \operatorname{cl}(h) \leq h$. In addition, Rockafellar (1970, Theorem 10.1) ensures that the convex function h is continuous on $\operatorname{rint}(\operatorname{dom}(h))$ and thus coincides with $\operatorname{cl}(h)$ there. Hence the claim follows. □

The main idea for dualizing the worst-case expectation problem (4.1) is to represent its optimal value as $-h(u)$, where $h(u) = \inf_{v\in\mathcal{V}} H(u,v)$, $\mathcal{U}$ is a finite-dimensional space of parameters u that encode the ambiguity set $\mathcal{P}$ (such as a set of prescribed moments or a size parameter), and $\mathcal{V}$ is an infinite-dimensional space of finite signed measures on $\mathcal{Z}$. In addition, $H(u,v)$ represents the negative expected loss if the signed measure v happens to be a probability measure in $\mathcal{P} \subseteq \mathcal{V}$ and evaluates to ∞ otherwise. If $H(u,v)$ is jointly convex on u and v, then $h(u)$ is convex by virtue of Lemma 4.1. A problem dual to (4.1) can then be constructed from the bi-conjugate $h^{**}(u)$. Lemma 4.2 provides conditions for strong duality.

4.2. Moment ambiguity sets

Recall from Section 2.1 that the generic moment ambiguity set (2.1) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}_f(\mathcal{Z})\colon \mathbb{E}_{\mathbb{P}}[f(Z)] \in \mathcal{F}\},$$

where $\mathcal{Z} \subseteq \mathbb{R}^d$ is a closed support set, $f\colon \mathcal{Z} \to \mathbb{R}^m$ is a Borel-measurable moment function, $\mathcal{F} \subseteq \mathbb{R}^m$ is a closed moment uncertainty set, and $\mathcal{P}_f(\mathcal{Z})$ denotes the family of all distributions $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ for which $\mathbb{E}_{\mathbb{P}}[f(Z)]$ is finite.[1] We may assume without loss of generality that $\mathcal{F}$ is covered by the convex set

$$\mathcal{C} = \{\mathbb{E}_{\mathbb{P}}[f(Z)]\colon \mathbb{P} \in \mathcal{P}_f(\mathcal{Z})\}$$

of all possible moments of any distribution on $\mathcal{Z}$. To rule out trivial special cases, we make the blanket assumption that $\mathcal{Z}$ and $\mathcal{F}$ are non-empty.

Clearly, problem (4.1) over the moment ambiguity set (2.1) can be recast as

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \sup_{u\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}_f(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)]\colon \mathbb{E}_{\mathbb{P}}[f(Z)] = u\} = \sup_{u\in\mathcal{F}} -h(1,u), \quad (4.2)$$

where the auxiliary function $h\colon \mathbb{R}\times\mathbb{R}^m \to \overline{\mathbb{R}}$ is defined by

$$h(u_0,u) = \inf_{v\in\mathcal{M}_{f,+}(\mathcal{Z})} \left\{ -\int_{\mathcal{Z}} \ell(z)\,\mathrm{d}v(z)\colon \int_{\mathcal{Z}} \mathrm{d}v(z) = u_0,\ \int_{\mathcal{Z}} f(z)\,\mathrm{d}v(z) = u \right\}. \quad (4.3)$$

Here the set $\mathcal{M}_{f,+}(\mathcal{Z})$ stands for the family of all Borel measures $v \in \mathcal{M}_+(\mathcal{Z})$ for which the integral $\int_{\mathcal{Z}} f(z)\,\mathrm{d}v(z)$ is finite. Put differently, $\mathcal{M}_{f,+}(\mathcal{Z})$ represents the

[1] Clearly, $\mathbb{E}_{\mathbb{P}}[f(Z)]$ must be finite to belong to the closed set $\mathcal{F}$. Therefore we may replace $\mathcal{P}(\mathcal{Z})$ with $\mathcal{P}_f(\mathcal{Z})$ in the definition of $\mathcal{P}$ without loss of generality. However, working with $\mathcal{P}_f(\mathcal{Z})$ is more convenient when we dualize the worst-case expectation problem (4.1) over $\mathcal{P}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

convex cone generated by $\mathcal{P}_f(\mathcal{Z})$. As the objective and constraint functions of the minimization problem in (4.3) are all jointly convex and jointly linear in v, u_0 and u, respectively, the equivalent reformulation that incorporates the constraints into the objective via indicator functions remains convex. This implies via Lemma 4.1 that h is convex. Under a reasonable regularity condition, one can further show that the domain of h coincides with the cone generated by $\{1\} \times \mathcal{C}$.

Lemma 4.3 (Domain of h). If $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for every $\mathbb{P} \in \mathcal{P}_f(\mathcal{Z})$, then we have

$$\operatorname{dom}(h) = \operatorname{cone}(\{1\} \times \mathcal{C}).$$

Proof. It is clear that $(u_0, u) \in \operatorname{dom}(h)$ if and only if $h(u_0, u) < \infty$, which is the case if and only if the minimization problem in (4.3) is feasible. Thus it remains to be shown that the problem in (4.3) is feasible if and only if $(u_0, u) \in \operatorname{cone}(\{1\} \times \mathcal{C})$. To this end, assume first that the problem in (4.3) is feasible at (u_0, u). This implies that there is $v \in \mathcal{M}_{f,+}(\mathcal{Z})$ with $\int_{\mathcal{Z}} \mathrm{d}v(z) = u_0$ and $\int_{\mathcal{Z}} f(z)\,\mathrm{d}v(z) = u$. Hence $u_0 \geq 0$. If $u_0 = 0$, then we must have $u = 0$. If $u_0 > 0$, on the other hand, then v/u_0 must be a probability measure in $\mathcal{P}_f(\mathcal{Z})$, which implies that $u/u_0 \in \mathcal{C}$. In either case, (u_0, u) is a non-negative multiple of a point in $\{1\} \times \mathcal{C}$ and thus belongs to $\operatorname{cone}(\{1\} \times \mathcal{C})$. Next, assume that $(u_0, u) \in \operatorname{cone}(\{1\} \times \mathcal{C})$. If $u_0 = 0$, then $u = 0$, and indeed, the zero measure in $\mathcal{M}_{f,+}(\mathcal{Z})$ is feasible in (4.3). If $u_0 > 0$, on the other hand, then $u/u_0 \in \mathcal{C}$. By the definition of $\mathcal{C}$, there exists a distribution $\mathbb{P} \in \mathcal{P}_f(\mathcal{Z})$ with $\mathbb{E}_{\mathbb{P}}[f(Z)] = u/u_0$. As $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$, this implies that $v = u_0\mathbb{P}$ is feasible in (4.3). We have thus shown that (4.3) is feasible if and only if $(u_0, u) \in \operatorname{cone}(\{1\} \times \mathcal{C})$. This observation completes the proof. □

The following proposition characterizes the bi-conjugate of h.

Proposition 4.4 (Bi-conjugate of h). The bi-conjugate of h defined in (4.3) satisfies

$$h^{**}(u_0, u) = \sup_{\lambda_0 \in \mathbb{R},\, \lambda \in \mathbb{R}^m} \{-u_0\lambda_0 - u^\top \lambda : \lambda_0 + f(z)^\top \lambda \geq \ell(z) \;\; \forall z \in \mathcal{Z}\}.$$

If additionally $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for every $\mathbb{P} \in \mathcal{P}_f(\mathcal{Z})$, then h^{**} and h match on the cone generated by $\{1\} \times \operatorname{rint}(\mathcal{C})$ except at the origin.

Proof. For any fixed $(\lambda_0, \lambda) \in \mathbb{R} \times \mathbb{R}^m$, the convex conjugate of h satisfies

$$\begin{aligned} h^*(-\lambda_0, -\lambda) &= \sup_{u_0 \in \mathbb{R},\, u \in \mathbb{R}^m} -u_0\lambda_0 - u^\top \lambda - h(u_0, u) \\ &= \begin{cases} \sup & -u_0\lambda_0 - u^\top \lambda + \displaystyle\int_{\mathcal{Z}} \ell(z)\,\mathrm{d}v(z) \\ \text{s.t.} & u_0 \in \mathbb{R},\; u \in \mathbb{R}^m,\; v \in \mathcal{M}_{f,+}(\mathcal{Z}) \\ & \displaystyle\int_{\mathcal{Z}} \mathrm{d}v(z) = u_0,\; \int_{\mathcal{Z}} f(z)\,\mathrm{d}v(z) = u \end{cases} \end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$$= \sup_{v \in \mathcal{M}_{f,+}(\mathcal{Z})} \int_{\mathcal{Z}} (\ell(z) - \lambda_0 - f(z)^\top \lambda)\, \mathrm{d}v(z)$$
$$= \begin{cases} 0 & \text{if } \ell(z) - \lambda_0 - f(z)^\top \lambda \leq 0 \ \ \forall z \in \mathcal{Z}, \\ \infty & \text{otherwise,} \end{cases}$$

where the last equality holds because $\mathcal{M}_{f,+}(\mathcal{Z})$ contains all weighted Dirac measures on $\mathcal{Z}$. Thus, for any fixed $(u_0, u) \in \mathbb{R} \times \mathbb{R}^m$, the conjugate of h^* satisfies

$$h^{**}(u_0, u) = \sup_{\lambda_0 \in \mathbb{R},\, \lambda \in \mathbb{R}^m} -u_0\lambda_0 - u^\top \lambda - h^*(-\lambda_0, -\lambda)$$
$$= \sup_{\lambda_0 \in \mathbb{R},\, \lambda \in \mathbb{R}^m} \{-u_0\lambda_0 - u^\top \lambda \colon \lambda_0 + f(z)^\top \lambda \geq \ell(z) \ \ \forall z \in \mathcal{Z}\}.$$

This establishes the desired formula for the bi-conjugate of h. Assume now that $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for every $\mathbb{P} \in \mathcal{P}_f(\mathcal{Z})$. It remains to be shown that $h(u_0, u) = h^{**}(u_0, u)$ for all $(u_0, u) \neq (0, 0)$ in the cone generated by $\{1\} \times \mathrm{rint}(\mathcal{C})$. However, this follows immediately from Lemma 4.2 and the observation that

$$\mathrm{rint}(\mathrm{dom}(h)) = \mathrm{rint}(\mathrm{cone}(\{1\} \times \mathcal{C})) = \mathrm{cone}(\{1\} \times \mathrm{rint}(\mathcal{C}))\backslash\{(0, 0)\},$$

where the two equalities hold because of Lemma 4.3 and Rockafellar (1970, Corollary 6.8.1), respectively. Therefore the claim follows. □

Proposition 4.4 implies that $h(1, u) = h^{**}(1, u)$ for all $u \in \mathrm{rint}(\mathcal{C})$. The following main theorem exploits this relation to convert the maximization problem on the right-hand side of (4.2) to an equivalent dual minimization problem.

Theorem 4.5 (Duality theory for moment ambiguity sets). If $\mathcal{P}$ is the moment ambiguity set (2.1), then the following weak duality relation holds:

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf & \lambda_0 + \delta^*_{\mathcal{F}}(\lambda) \\ \text{s.t.} & \lambda_0 \in \mathbb{R},\ \lambda \in \mathbb{R}^m \\ & \lambda_0 + f(z)^\top \lambda \geq \ell(z) \ \ \forall z \in \mathcal{Z}. \end{cases} \tag{4.4}$$

If $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}_f(\mathcal{Z})$ and $\mathcal{F} \subseteq \mathcal{C}$ is a convex and compact set with $\mathrm{rint}(\mathcal{F}) \subseteq \mathrm{rint}(\mathcal{C})$, then strong duality holds, that is, (4.4) becomes an equality.

Proof. For ease of exposition, we introduce

$$\mathcal{L} = \{(\lambda_0, \lambda) \in \mathbb{R} \times \mathbb{R}^m \colon \lambda_0 + f(z)^\top \lambda \geq \ell(z) \ \ \forall z \in \mathcal{Z}\}$$

as shorthand for the dual feasible set. Using the decomposition (4.2), we find

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \sup_{u \in \mathcal{F}} -h(1, u)$$
$$\leq \sup_{u \in \mathcal{F}} \inf_{(\lambda_0, \lambda) \in \mathcal{L}} \lambda_0 + u^\top \lambda$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$$\leq \inf_{(\lambda_0,\lambda)\in\mathcal{L}} \sup_{u\in\mathcal{F}} \lambda_0 + u^\top\lambda$$
$$= \inf_{(\lambda_0,\lambda)\in\mathcal{L}} \lambda_0 + \delta^*_{\mathcal{F}}(\lambda).$$

Here the first inequality exploits Proposition 4.4 and Lemma 4.2, which ensures that $h \geq h^{**}$, and the second inequality holds thanks to the max-min inequality. The last equality follows from the definition of the support function $\delta^*_{\mathcal{F}}$. This establishes the weak duality relation (4.4). Next, suppose that $\mathcal{F}$ is a convex compact set with $\operatorname{rint}(\mathcal{F}) \subseteq \operatorname{rint}(\mathcal{C})$. Under this additional assumption, we have

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] &= \sup_{u\in\mathcal{F}} -h(1,u) \\
&= \sup_{u\in\operatorname{rint}(\mathcal{F})} -h(1,u) \\
&= \sup_{u\in\operatorname{rint}(\mathcal{F})} \inf_{(\lambda_0,\lambda)\in\mathcal{L}} \lambda_0 + u^\top\lambda \\
&= \sup_{u\in\mathcal{F}} \inf_{(\lambda_0,\lambda)\in\mathcal{L}} \lambda_0 + u^\top\lambda \\
&= \inf_{(\lambda_0,\lambda)\in\mathcal{L}} \lambda_0 + \delta^*_{\mathcal{F}}(\lambda),
\end{aligned}$$

where the first equality exploits (4.2). The second equality follows from two observations. First, $\operatorname{rint}(\mathcal{F})$ is non-empty and convex (Rockafellar 1970, Theorem 6.2). Second, $-h(1,u)$ is concave in u, which ensures that $-h(1,u)$ cannot jump up on the boundary of its domain $\mathcal{C}$ and – in particular – on the boundary of $\mathcal{F} \subseteq \mathcal{C}$. Taken together, these observations imply that we can restrict $\mathcal{F}$ to $\operatorname{rint}(\mathcal{F})$ without reducing the supremum. The third equality follows from Proposition 4.4, which allows us to replace h with h^{**} on $\operatorname{rint}(\mathcal{F}) \subseteq \operatorname{rint}(\mathcal{C})$. The fourth equality holds because $-h^{**}(1,u)$ is concave in u, which allows us to change $\operatorname{rint}(\mathcal{F})$ back to $\mathcal{F}$. Finally, the fifth equality follows from Sion's minimax theorem (Sion 1958, Theorem 4.2), which applies because $\mathcal{F}$ is convex and compact, $\mathcal{L}$ is convex and $\lambda_0 + u^\top\lambda$ is biaffine in u and (λ_0,λ). Therefore strong duality holds. □

Theorem 4.5 shows that the worst-case expectation problem (4.1) over the moment ambiguity set (2.1) admits a semi-infinite dual. Indeed, the dual problem on the right-hand side of (4.4) accommodates finitely many decision variables but infinitely many constraints parametrized by the uncertainty realizations $z \in \mathcal{Z}$. The dual problem can also be interpreted as a robust optimization problem with uncertainty set $\mathcal{Z}$. Note that we did *not* assume $\mathcal{Z}$ to be convex. In addition, we emphasize that compactness of $\mathcal{F}$ is *not* a necessary condition for strong duality. Indeed, strong duality can also be established under Slater-type conditions (Zhen, Kuhn and Wiesemann 2023). Finally, the condition $\operatorname{rint}(\mathcal{F}) \subseteq \operatorname{rint}(\mathcal{C})$ is equivalent to the – seemingly weaker – requirement that $\mathcal{F}$ intersects $\operatorname{rint}(\mathcal{C})$. Indeed, if $\mathcal{F} \cap \operatorname{rint}(\mathcal{C}) \neq \emptyset$, then $\mathcal{F}$ is not entirely contained in the relative boundary of $\mathcal{C}$, which implies via Rockafellar (1970, Corollary 6.5.2) that $\operatorname{rint}(\mathcal{F}) \subseteq \operatorname{rint}(\mathcal{C})$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

In the remainder of this section we use Theorem 4.5 to dualize worst-case expectations problems corresponding to popular classes of moment ambiguity sets. Recall from Section 2.1.4 that the Chebyshev ambiguity set (2.4) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}_2(\mathcal{Z})\colon \mathbb{E}_{\mathbb{P}}[Z] = \mu,\ \mathbb{E}_{\mathbb{P}}[ZZ^\top] = M\ \ \forall (\mu, M) \in \mathcal{F}\},$$

where $\mathcal{F} \subseteq \mathbb{R}^d \times \mathbb{S}_+^d$ is a closed moment uncertainty set, and $\mathcal{P}_2(\mathcal{Z})$ denotes the set of all distributions in $\mathcal{P}(\mathcal{Z})$ with finite second moments. Note that $\mathcal{P}$ is an instance of the generic moment ambiguity set (2.1) with moment function $f(z) = (z, zz^\top)$.

Theorem 4.6 (Duality theory for Chebyshev ambiguity sets). If $\mathcal{P}$ is the Chebyshev ambiguity set (2.4), then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf & \lambda_0 + \delta^*_{\mathcal{F}}(\lambda, \Lambda) \\ \text{s.t.} & \lambda_0 \in \mathbb{R},\ \lambda \in \mathbb{R}^d,\ \Lambda \in \mathbb{S}^d \\ & \lambda_0 + \lambda^\top z + z^\top \Lambda z \geq \ell(z)\ \ \forall z \in \mathcal{Z}. \end{cases} \tag{4.5}$$

If $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}_2(\mathcal{Z})$ and $\mathcal{F}$ is a convex compact set with $M \succ \mu\mu^\top$ for all $(\mu, M) \in \operatorname{rint}(\mathcal{F})$, then strong duality holds, that is, (4.5) becomes an equality.

Theorem 4.6 is a direct corollary of Theorem 4.5. Thus we omit its proof. Recall that the Chebyshev ambiguity (2.4) set with uncertain moments encapsulates the support-only ambiguity set $\mathcal{P}(\mathcal{Z})$, the Markov ambiguity set (2.2) and the Chebyshev ambiguity set (2.3) with fixed moments as special cases. They are recovered by setting $\mathcal{F} = \mathbb{R}^d \times \mathbb{S}^d$, $\mathcal{F} = \{\mu\} \times \mathbb{S}^d$ and $\mathcal{F} = \{\mu\} \times \{M\}$, respectively. The following lemma characterizes the support functions of these moment uncertainty sets in closed form. The proof is elementary and is thus omitted.

Lemma 4.7 (Support functions of elementary sets). The following hold.

(i) If $\mathcal{F} = \mathbb{R}^d \times \mathbb{S}^d$, then $\delta^*_{\mathcal{F}}(\lambda, \Lambda) = \delta_{\{(0,0)\}}(\lambda, \Lambda)$.

(ii) If $\mathcal{F} = \{\mu\} \times \mathbb{S}^d$, then $\delta^*_{\mathcal{F}}(\lambda, \Lambda) = \lambda^\top \mu + \delta_{\{0\}}(\Lambda)$.

(iii) If $\mathcal{F} = \{\mu\} \times \{M\}$, then $\delta^*_{\mathcal{F}}(\lambda, \Lambda) = \lambda^\top \mu + \operatorname{Tr}(\Lambda M)$.

When combined with Theorem 4.5, Lemma 4.7 immediately leads to duality theorems for support-only, Markov and Chebyshev ambiguity sets. For brevity, we omit the details. In Section 2.1.4, we have also defined the Gelbrich ambiguity set as a Chebyshev ambiguity set with uncertain moments of the form (2.4) with $\mathcal{F}$ representing the Gelbrich uncertainty set (2.8) defined as

$$\mathcal{F} = \left\{(\mu, M) \in \mathbb{R}^d \times \mathbb{S}_+^d\colon \begin{array}{l} \exists \Sigma\ \in \mathbb{S}_+^d \text{ with } M = \Sigma + \mu\mu^\top, \\ \mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) \leq r \end{array}\right\},$$

where G is the Gelbrich distance of Definition 2.1. In the following we derive the support function $\delta^*_{\mathcal{F}}$ of the Gelbrich uncertainty set $\mathcal{F}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Lemma 4.8 (Support function of Gelbrich uncertainty sets). Let $\mathcal{F}$ be the Gelbrich uncertainty set (2.8) of radius $r \geq 0$ around $(\hat{\mu}, \hat{\Sigma}) \in \mathbb{R}^d \times \mathbb{S}_+^d$, where G is the Gelbrich distance of Definition 2.1. For any $(\lambda, \Lambda) \in \mathbb{R}^d \times \mathbb{S}^d$, we then have

$$\delta^*_{\mathcal{F}}(\lambda, \Lambda) = \begin{cases} \inf & \gamma(r^2 - \|\hat{\mu}\|^2 - \mathrm{Tr}(\hat{\Sigma})) + \mathrm{Tr}(A) + \alpha \\ \text{s.t.} & \alpha, \gamma \in \mathbb{R}_+,\ A \in \mathbb{S}_+^d \\ & \begin{bmatrix} \gamma I_d - \Lambda & \gamma \hat{\Sigma}^{1/2} \\ \gamma \hat{\Sigma}^{1/2} & A \end{bmatrix} \geq 0,\ \begin{bmatrix} \gamma I_d - \Lambda & \gamma \hat{\mu} + \lambda/2 \\ (\gamma \hat{\mu} + \lambda/2)^\top & \alpha \end{bmatrix} \geq 0. \end{cases}$$

Proof. By Proposition 2.3, which provides a semidefinite representation of the Gelbrich uncertainty set $\mathcal{F}$, the support function of $\mathcal{F}$ satisfies

$$\delta^*_{\mathcal{F}}(\lambda, \Lambda) = \begin{cases} \sup & \mu^\top \lambda + \mathrm{Tr}(M\Lambda) \\ \text{s.t.} & \mu \in \mathbb{R}^d,\ M, U \in \mathbb{S}_+^d,\ C \in \mathbb{R}^{d \times d} \\ & \mathrm{Tr}(M - 2\mu\hat{\mu}^\top - 2C) \leq r^2 - \|\hat{\mu}\|^2 - \mathrm{Tr}(\hat{\Sigma}) \\ & \begin{bmatrix} M - U & C \\ C^\top & \hat{\Sigma} \end{bmatrix} \geq 0,\ \begin{bmatrix} U & \mu \\ \mu^\top & 1 \end{bmatrix} \geq 0. \end{cases}$$

By conic duality (Ben-Tal and Nemirovski 2001, Theorem 1.4.2), the maximization problem in the above expression admits the dual minimization problem

$$\begin{array}{ll} \inf & \gamma(r^2 - \|\hat{\mu}\|^2 - \mathrm{Tr}(\hat{\Sigma})) + \mathrm{Tr}(\hat{\Sigma} A_{22}) + \alpha \\ \text{s.t.} & \alpha, \gamma \in \mathbb{R}_+,\ A_{11}, A_{22}, B \in \mathbb{S}_+^d \\ & \begin{bmatrix} A_{11} & \gamma I_d \\ \gamma I_d & A_{22} \end{bmatrix} \geq 0,\ \begin{bmatrix} B & \gamma \hat{\mu} + \lambda/2 \\ (\gamma \hat{\mu} + \lambda/2)^\top & \alpha \end{bmatrix} \geq 0,\quad \gamma I_d - \Lambda \geq A_{11} \geq B. \end{array}$$

Strong duality holds because $\alpha = \|2\gamma\hat{\mu} + \lambda\|^2$, $\gamma = \max\{\lambda_{\max}(\Lambda), 0\} + 4$, $A_{11} = 2I$, $A_{22} = \gamma^2 I$ and $B = I$ represents a Slater point for the dual problem. At optimality, we have $\gamma I_d - \Lambda = A_{11} = B$. Hence the dual problem can be further simplified to

$$\begin{array}{ll} \inf & \gamma(r^2 - \|\hat{\mu}\|^2 - \mathrm{Tr}(\hat{\Sigma})) + \mathrm{Tr}(\hat{\Sigma} A_{22}) + \alpha \\ \text{s.t.} & \alpha, \gamma \in \mathbb{R}_+,\ A_{22} \in \mathbb{S}_+^d \\ & \begin{bmatrix} \gamma I_d - \Lambda & \gamma I_d \\ \gamma I_d & A_{22} \end{bmatrix} \geq 0,\ \begin{bmatrix} \gamma I_d - \Lambda & \gamma \hat{\mu} + \lambda/2 \\ (\gamma \hat{\mu} + \lambda/2)^\top & \alpha \end{bmatrix} \geq 0. \end{array}$$

The substitution $A \leftarrow \hat{\Sigma}^{1/2} A_{22} \hat{\Sigma}^{1/2}$ and the equivalence

$$\begin{bmatrix} \gamma I_d - \Lambda & \gamma I_d \\ \gamma I_d & A_{22} \end{bmatrix} \geq 0 \iff \begin{bmatrix} I_d & 0 \\ 0 & \hat{\Sigma}^{1/2} \end{bmatrix} \begin{bmatrix} \gamma I_d - \Lambda & \gamma I_d \\ \gamma I_d & A_{22} \end{bmatrix} \begin{bmatrix} I_d & 0 \\ 0 & \hat{\Sigma}^{1/2} \end{bmatrix} \geq 0$$

then yield the desired semidefinite program. Thus the optimal value of this semidefinite program indeed equals $\delta^*_{\mathcal{F}}(\lambda, \Lambda)$. □

Armed with Theorem 4.6 and Lemma 4.8, we are now prepared to dualize the worst-case expectation problem over a Gelbrich ambiguity set.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Theorem 4.9 (Duality theory for Gelbrich ambiguity sets). If $\mathcal{P}$ is the Chebyshev ambiguity set (2.4) with $\mathcal{F}$ representing the Gelbrich uncertainty (2.8), then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf & \lambda_0 + \gamma(r^2 - \|\hat{\mu}\|^2 - \mathrm{Tr}(\hat{\Sigma})) + \mathrm{Tr}(A) + \alpha \\ \text{s.t.} & \lambda_0 \in \mathbb{R},\ \alpha, \gamma \in \mathbb{R}_+,\ \lambda \in \mathbb{R}^d,\ \Lambda \in \mathbb{S}^d,\ A \in \mathbb{S}^d_+ \\ & \lambda_0 + \lambda^\top z + z^\top \Lambda z \geq \ell(z) \quad \forall z \in \mathcal{Z} \\ & \begin{bmatrix} \gamma I_d - \Lambda & \gamma\hat{\Sigma}^{1/2} \\ \gamma\hat{\Sigma}^{1/2} & A \end{bmatrix} \geq 0, \quad \begin{bmatrix} \gamma I_d - \Lambda & \gamma\hat{\mu} + \lambda/2 \\ (\gamma\hat{\mu} + \lambda/2)^\top & \alpha \end{bmatrix} \geq 0. \end{cases} \tag{4.6}$$

If $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}_2(\mathcal{Z})$ and $r > 0$, then strong duality holds, that is, the inequality (4.6) becomes an equality.

Proof. Weak duality follows immediately from the first claim of Theorem 4.6 and Lemma 4.8. To prove strong duality, recall from Proposition 2.3 that the Gelbrich uncertainty set $\mathcal{F}$ is convex and compact. In addition, recall from the proof of Proposition 2.2 that the Gelbrich distance is continuous. As $r > 0$, this implies that

$$\mathrm{rint}(\mathcal{F}) = \left\{(\mu, M) \in \mathbb{R}^d \times \mathbb{S}^d_+ \colon M \succ \mu\mu^\top,\ \mathrm{G}((\mu, M - \mu\mu^\top), (\hat{\mu}, \hat{\Sigma})) < r\right\},$$

which in turn ensures that $M \succ \mu\mu^\top$ for all $(\mu, M) \in \mathrm{rint}(\mathcal{F})$. Therefore, strong duality follows from the second claim of Theorem 4.6. □

We close this section with some historical remarks. The classical problem of moments asks whether there exists a distribution on $\mathcal{Z}$ with a given sequence of moments. In the language of this survey, the problem of moments thus seeks to determine whether a given moment ambiguity set of the form (2.1) is nonempty, where f is a polynomial and $\mathcal{F}$ is a singleton. The analysis of moment problems has a long and distinguished history in mathematics dating back to the nineteenth century. Notable contributions were made by Chebyshev (1874), Markov (1884), Stieltjes (1894), Hamburger (1920) and Hausdorff (1923); see Shohat and Tamarkin (1950) for an early survey. The study of moment problems with tools from mathematical optimization – in particular semi-infinite duality theory – was pioneered by Isii (1960, 1962). Shapiro (2001) formulates the worst-case expectation problem over a family of distributions with prescribed moments as an infinite-dimensional conic linear program and establishes conditions for strong duality.

4.3. ϕ-divergence ambiguity sets

Recall from Section 2.2 that the ϕ-divergence ambiguity set (2.10) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

Here $\mathcal{Z}$ is a closed support set, $r \geq 0$ is a size parameter, ϕ is an entropy function in the sense of Definition 2.4, D_ϕ is the corresponding ϕ-divergence in the sense

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

of Definition 2.5, and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. It is expedient to extend D_ϕ to arbitrary measures. By slight abuse of notation, we thus define the ϕ-divergence of $\nu \in \mathcal{M}_+(\mathcal{Z})$ with respect to $\hat{\nu} \in \mathcal{M}_+(\mathcal{Z})$ as

$$\mathrm{D}_\phi(\nu, \hat{\nu}) = \int_{\mathcal{Z}} \phi^\pi \left(\frac{\mathrm{d}\nu}{\mathrm{d}\rho}(z), \frac{\mathrm{d}\hat{\nu}}{\mathrm{d}\rho}(z) \right) \mathrm{d}\rho(z),$$

where $\rho \in \mathcal{M}_+(\mathcal{Z})$ is a dominating measure with $\nu, \hat{\nu} \ll \rho$. An obvious generalization of Proposition 2.6 implies that $\mathrm{D}_\phi(\nu, \hat{\nu})$ is convex in $(\nu, \hat{\nu})$ and independent of the choice of ρ. By using the extension of D_ϕ to general measures, the worst-case expectation problem (4.1) over the ambiguity set (2.10) can now be recast as

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = -h(1, r),$$

where the auxiliary function $h\colon \mathbb{R}^2 \to \overline{\mathbb{R}}$ is defined by

$$h(u_0, u) = \inf_{\nu \in \mathcal{M}_+(\mathcal{Z})} \left\{ -\int_{\mathcal{Z}} \ell(z)\, \mathrm{d}\nu(z) \colon \int_{\mathcal{Z}} \mathrm{d}\nu(z) = u_0,\ \mathrm{D}_\phi(\nu, \hat{\mathbb{P}}) \le u \right\}. \tag{4.7}$$

As the objective and constraint functions of the minimization problem in (4.7) are jointly convex in ν, u_0 and u, Lemma 4.1 implies that h is convex. Clearly we have $\mathrm{dom}(h) \subseteq \mathbb{R}^2_+$. Under mild regularity conditions, one can additionally show that $\{1\} \times \mathbb{R}_{++} \subseteq \mathrm{rint}(\mathrm{dom}(h))$.

Lemma 4.10 (Domain of h). If $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$ and ϕ is continuous at 1, then

$$\{1\} \times \mathbb{R}_{++} \subseteq \mathrm{rint}(\mathrm{dom}(h)).$$

Proof. If $u_0 = 1$ and $u > 0$, then $\nu = \hat{\mathbb{P}}$ is feasible in (4.7). Indeed, $\hat{\mathbb{P}}$ obeys both constraints, and its objective function value satisfies $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. If we perturb u_0 and u locally, then $u_0\hat{\mathbb{P}}$ satisfies the equality constraint, and the objective function does *not* evaluate to $-\infty$ for all $u_0 \ge 0$. The inequality constraint, on the other hand, is satisfied for all $u > 0$ and all u_0 that are sufficiently close to 1 because

$$\mathrm{D}_\phi(u_0\hat{\mathbb{P}}, \hat{\mathbb{P}}) = \phi^\pi(u_0, 1) = \phi(u_0) < u.$$

Here the first equality follows from the definition of D_ϕ with $\rho = \hat{\mathbb{P}}$, the second equality follows from the definition of the perspective function ϕ^π, and the inequality holds because $\phi(1) = 0$, $u > 0$ and $\phi(u_0)$ is continuous at $u_0 = 1$. This confirms that $(1, u) \in \mathrm{rint}(\mathrm{dom}(h))$ for every $u > 0$, and thus the claim follows. □

The following two lemmas are instrumental to deriving the bi-conjugate of h.

Lemma 4.11 (Conjugates of scaled perspective functions). If ϕ is an entropy function in the sense of Definition 2.4, $t \in \mathbb{R}$, $\beta \in \mathbb{R}_+$ and $\lambda \in \mathbb{R}_{++}$, then we have

$$\sup_{\alpha \in \mathbb{R}} t\alpha - \lambda\phi^\pi(\alpha, \beta) = \begin{cases} \beta\lambda\phi^*(t/\lambda) & \text{if } \beta > 0, \\ \lambda\delta_{\mathrm{cl}(\mathrm{dom}(\phi^*))}(t/\lambda) & \text{if } \beta = 0. \end{cases}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Proof. If $\beta > 0$, then we have

$$\sup_{\alpha\in\mathbb{R}} t\alpha - \lambda\phi^{\pi}(\alpha,\beta) = \sup_{\alpha\in\mathbb{R}} t\alpha - \lambda\beta\phi(\alpha/\beta) = \beta \sup_{\alpha\in\mathbb{R}} t\alpha - \lambda\phi(\alpha) = \beta\lambda\phi^*(t/\lambda),$$

where the three equalities follow from the definition of the perspective function ϕ^{π}, the substitution $\alpha \leftarrow \alpha/\beta$ and the replacement of t by $\lambda t/\lambda$, respectively. Note that these manipulations are admissible because $\beta, \lambda > 0$. If $\beta = 0$, then we have

$$\begin{aligned}\sup_{\alpha\in\mathbb{R}} t\alpha - \lambda\phi^{\pi}(\alpha,\beta) &= \sup_{\alpha\in\mathbb{R}} t\alpha - \lambda\phi^{\infty}(\alpha)\\ &= \sup_{\alpha\in\mathbb{R}} t\alpha - \lambda\delta^*_{\mathrm{dom}(\phi^*)}(\alpha)\\ &= \lambda\delta_{\mathrm{cl}(\mathrm{dom}(\phi^*))}(t/\lambda),\end{aligned}$$

where the first equality again holds because of the definition of ϕ^{π}, and the second equality exploits Rockafellar (1970, Theorem 13.3). The third equality replaces t with $\lambda t/\lambda$ and exploits the elementary observation that the conjugate of the support function of a convex set coincides with the indicator function of the closure of this set (Rockafellar 1970, Theorem 13.2). Thus the claim follows. □

Lemma 4.12 (Domain of conjugate entropy functions). If ϕ is an entropy function in the sense of Definition 2.4, then we have

$$\mathrm{cl}(\mathrm{dom}(\phi^*)) = \begin{cases}(-\infty, \phi^{\infty}(1)] & \text{if } \phi^{\infty}(1) < \infty,\\ \mathbb{R} & \text{if } \phi^{\infty}(1) = \infty.\end{cases}$$

Proof. As ϕ is proper, convex and closed, Rockafellar (1970, Theorem 8.5) implies that its recession function ϕ^{∞} is positive homogeneous. Recall that $\phi(s) = \infty$ for every $s < 0$. We may thus conclude that $\phi^{\infty}(t) = t\,\phi^{\infty}(1)$ for $t > 0$, $\phi^{\infty}(t) = 0$ for $t = 0$ and $\phi^{\infty}(t) = \infty$ for $t < 0$. In addition, Rockafellar (1970, Theorem 13.3) implies that the support function of $\mathrm{dom}(\phi^*)$ coincides with the recession function ϕ^{∞}. The indicator function of $\mathrm{cl}(\mathrm{dom}(\phi^*))$ is known to coincide with the conjugate of the support function of $\mathrm{dom}(\phi^*)$, and therefore it satisfies

$$\delta_{\mathrm{cl}(\mathrm{dom}(\phi^*))}(s) = \sup_{t\in\mathbb{R}_+} (s - \phi^{\infty}(1))t = \begin{cases}0 & \text{if } s \le \phi^{\infty}(1),\\ \infty & \text{otherwise.}\end{cases}$$

This shows that $\mathrm{cl}(\mathrm{dom}(\phi^*)) = (-\infty, \phi^{\infty}(1)]$ if $\phi^{\infty}(1) < \infty$ and that $\mathrm{cl}(\mathrm{dom}(\phi^*)) = \mathbb{R}$ otherwise. Hence the claim follows. □

Proposition 4.13 (Bi-conjugate of h). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. Then the bi-conjugate of h defined in (4.7) satisfies

$$h^{**}(u_0, u) = \begin{cases}\sup\limits_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} & -\lambda_0 u_0 - \lambda u - \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^{\pi}(\ell(Z) - \lambda_0, \lambda)]\\ \text{s.t.} & \lambda_0 + \lambda\,\phi^{\infty}(1) \ge \sup\limits_{z\in\mathcal{Z}} \ell(z),\end{cases}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Table 4.1. Examples of entropy functions, their asymptotic slopes and their conjugates. Here, for any $c \in \mathbb{R}$, we use the $[c]_+$ as shorthand for $\max\{c, 0\}$.

Divergence	$\phi(s)$ $(s \geq 0)$	$\phi^\infty(1)$	$\phi^*(t)$
Kullback–Leibler	$s\log(s) - s + 1$	∞	$\mathrm{e}^t - 1$
Likelihood	$-\log(s) + s - 1$	1	$-\log(1-t)$
Total variation	$\frac{1}{2}\lvert s-1\rvert$	$\frac{1}{2}$	$\max\{t, -1/2\} + \delta_{(-\infty,1/2]}(t)$
Pearson χ^2	$(s-1)^2$	∞	$(t/2+1)_+^2 - 1$
Neyman χ^2	$\frac{1}{s}(s-1)^2$	1	$2 - 2\sqrt{1-t}$
Cressie–Read for $\beta \in (0,1)$	$\dfrac{s^\beta - \beta s + \beta - 1}{\beta(\beta-1)}$	1	$\dfrac{[(\beta-1)t+1]_+^{\beta/(\beta-1)}}{\beta}$
Cressie–Read for $\beta > 1$	$\dfrac{s^\beta - \beta s + \beta - 1}{\beta(\beta-1)}$	∞	$\dfrac{[(\beta-1)t+1]_+^{\beta/(\beta-1)}}{\beta}$

where the product $\lambda\,\phi^\infty(1)$ is assumed to evaluate to ∞ if $\lambda = 0$ and $\phi^\infty(1) = \infty$. If ϕ is continuous at 1, then h^{**} coincides with h on $\{1\} \times \mathbb{R}_{++}$.

As $\phi(1) = 0$, we have $\phi^*(\tau) = \sup_{\alpha\in\mathbb{R}} \tau\alpha - \phi(\alpha) \geq \tau$ for all $\tau \in \mathbb{R}$. This readily implies that $(\phi^*)^\pi(\tau, \lambda) \geq \tau$ for all $\tau, \lambda \in \mathbb{R}$. Hence $\mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\ell(Z) - \lambda_0, \lambda)] \geq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z) - \lambda_0]$. In addition, ϕ^* is non-decreasing because $\mathrm{dom}(\phi) \subseteq \mathbb{R}_+$. Examples of common entropy functions and their conjugates are listed in Table 4.1.

Proof of Proposition 4.13. For any fixed $(\lambda_0, \lambda) \in \mathbb{R}^2$, the conjugate of h satisfies

$$\begin{aligned} h^*(-\lambda_0, -\lambda) &= \sup_{(u_0,u)\in\mathbb{R}^2} -\lambda_0 u_0 - \lambda u - h(u_0, u) \\ &= \sup_{u\in\mathbb{R}_+, v\in\mathcal{M}_+(\mathcal{Z})} \left\{ -\lambda u + \int_{\mathcal{Z}} (\ell(z) - \lambda_0)\,\mathrm{d}v(z) \colon \mathrm{D}_\phi(v, \hat{\mathbb{P}}) \leq u \right\}, \end{aligned}$$

where the second equality holds because $\int_{\mathcal{Z}} \mathrm{d}v(z) = u_0$ and $\mathrm{D}_\phi(v, \hat{\mathbb{P}}) \geq 0$. As $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$, the resulting maximization problem over u is unbounded whenever $\lambda < 0$. If $\lambda > 0$, on the other hand, then we find

$$\begin{aligned} h^*(-\lambda_0, -\lambda) &= \sup_{v\in\mathcal{M}_+(\mathcal{Z})} \int_{\mathcal{Z}} (\ell(z) - \lambda_0)\,\mathrm{d}v(z) - \lambda \mathrm{D}_\phi(v, \hat{\mathbb{P}}) \\ &= \begin{cases} \sup & \displaystyle\int_{\mathcal{Z}} (\ell(z) - \lambda_0) \frac{\mathrm{d}v}{\mathrm{d}\rho}(z) - \lambda\phi^\pi\left(\frac{\mathrm{d}v}{\mathrm{d}\rho}(z), \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z)\right) \mathrm{d}\rho(z) \\ \text{s.t.} & v, \rho \in \mathcal{M}_+(\mathcal{Z}),\ v \ll \rho,\ \hat{\mathbb{P}} \ll \rho, \end{cases} \end{aligned}$$

where the second equality exploits the definition of D_ϕ. Note that $\mathrm{d}v/\mathrm{d}\rho(z)$ and $\mathrm{d}\hat{\mathbb{P}}/\mathrm{d}\rho(z)$ belong to the space $\mathcal{L}_1(\rho)$ of all ρ-integrable Borel functions that can

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

be represented as the Radon–Nikodym derivative of some measure in $\mathcal{M}_+(\mathcal{Z})$ with respect to ρ. Introducing auxiliary decision variables $\alpha, \beta \in \mathcal{L}_1(\rho)$ for the Radon–Nikodym derivatives of v and $\hat{\mathbb{P}}$, respectively, and eliminating the measure v yields

$$h^*(-\lambda_0, -\lambda) = \begin{cases} \sup & \displaystyle\int_{\mathcal{Z}} (\ell(z) - \lambda_0)\alpha(z) - \lambda\phi^\pi(\alpha(z), \beta(z))\, \mathrm{d}\rho(z) \\ \text{s.t.} & \rho \in \mathcal{M}_+(\mathcal{Z}),\ \alpha, \beta \in \mathcal{L}_1(\rho) \\ & \dfrac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho} = \beta \ \ \rho\text{-a.s.} \end{cases} \tag{4.8}$$

For any $\rho \in \mathcal{M}_+(\mathcal{Z})$ and $\beta \in \mathcal{L}_1(\rho)$ with $\beta = \mathrm{d}\hat{\mathbb{P}}/\mathrm{d}\rho$ ρ-almost surely, we then find

$$\begin{aligned} &\sup_{\alpha \in \mathcal{L}_1(\rho)} \int_{\mathcal{Z}} (\ell(z) - \lambda_0)\alpha(z) - \lambda\phi^\pi(\alpha(z), \beta(z))\, \mathrm{d}\rho(z) \\ &\quad = \int_{\mathcal{Z}} \sup_{\alpha \in \mathbb{R}} \{(\ell(z) - \lambda_0)\alpha - \lambda\phi^\pi(\alpha, \beta(z))\}\, \mathrm{d}\rho(z), \end{aligned} \tag{4.9}$$

where the equality follows from Rockafellar and Wets (2009, Theorem 14.60), which applies because the negation of the function in curly brackets in (4.9) constitutes a normal integrand in the sense of Rockafellar and Wets (2009, Definition 14.27). This can be verified by recalling that sums and perspectives of normal integrands are normal integrands (Rockafellar and Wets 2009, Proposition 14.45 and Example 14.48). Next, we partition $\mathcal{Z}$ into $\mathcal{Z}_+(\beta) = \{z \in \mathcal{Z} : \beta(z) > 0\}$ and $\mathcal{Z}_0(\beta) = \{z \in \mathcal{Z} : \beta(z) = 0\}$. By Lemma 4.11, the integral (4.9) equals

$$\int_{\mathcal{Z}_+(\beta)} \lambda\phi^*\left(\frac{\ell(z) - \lambda_0}{\lambda}\right)\beta(z)\, \mathrm{d}\rho(z) + \int_{\mathcal{Z}_0(\beta)} \lambda\delta_{\mathrm{cl}(\mathrm{dom}(\phi^*))}\left(\frac{\ell(z) - \lambda_0}{\lambda}\right) \mathrm{d}\rho(z).$$

As $\beta = \mathrm{d}\hat{\mathbb{P}}/\mathrm{d}\rho$ ρ-almost surely, and as $\hat{\mathbb{P}}(Z \in \mathcal{Z}_+(\beta)) = 1$, the first of these integrals simply reduces to an expectation with respect to the reference distribution and is thus independent of β. The second integral still depends on β through the integration domain $\mathcal{Z}_0(\beta)$. Thus, partially maximizing over α allows us to recast (4.8) as

$$\begin{aligned} h^*(-\lambda_0, -\lambda) = \mathbb{E}_{\hat{\mathbb{P}}}\left[\lambda\phi^*\left(\frac{\ell(Z) - \lambda_0}{\lambda}\right)\right] \\ + \sup_{\substack{\rho \in \mathcal{M}_+(\mathcal{Z}), \\ \beta \in \mathcal{L}_1(\rho)}} \left\{\int_{\mathcal{Z}_0(\beta)} \lambda\delta_{\mathrm{cl}(\mathrm{dom}(\phi^*))}\left(\frac{\ell(z) - \lambda_0}{\lambda}\right) \mathrm{d}\rho(z) : \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho} = \beta\ \rho\text{-a.s.}\right\}. \end{aligned}$$

If there exists $z_0 \in \mathcal{Z}$ with $(\ell(z_0) - \lambda_0)/\lambda \notin \mathrm{cl}(\mathrm{dom}(\phi^*))$, then $h^*(-\lambda_0, -\lambda) = \infty$. To see this, assume first that z_0 is an atom of $\hat{\mathbb{P}}$. In this case, the expectation in the first line evaluates to ∞. If z_0 is *not* an atom of $\hat{\mathbb{P}}$, then the supremum in the second line evaluates to ∞ because we may set $\rho = \hat{\mathbb{P}} + \delta_{z_0}$ and define $\beta \in \mathcal{L}_1(\rho)$ through

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$\beta(z) = 1$ if $z \neq z_0$ and $\beta(z_0) = 0$. Hence we may conclude that

$$h^*(-\lambda_0, -\lambda) = \begin{cases} \mathbb{E}_{\hat{\mathbb{P}}}\left[\lambda\phi^*\left(\dfrac{\ell(Z)-\lambda_0}{\lambda}\right)\right] & \text{if } \dfrac{\ell(z)-\lambda_0}{\lambda} \in \mathrm{cl}(\mathrm{dom}(\phi^*)) \ \ \forall z \in \mathcal{Z}, \\ \infty & \text{otherwise.} \end{cases}$$

Note that this formula was derived under the assumption that $\lambda > 0$. Note also that, by Lemma 4.12, the condition $(\ell(z)-\lambda_0)/\lambda \in \mathrm{cl}(\mathrm{dom}(\phi^*))$ is equivalent to the requirement that $\lambda_0 + \lambda\,\phi^\infty(1)$ is larger than or equal to $\sup_{z\in\mathcal{Z}} \ell(z)$. We claim that

$$h^*(-\lambda_0, -\lambda) = \begin{cases} \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\ell(Z)-\lambda_0, \lambda)] & \text{if } \lambda \geq 0 \text{ and } \lambda_0 + \lambda\,\phi^\infty(1) \geq \sup\limits_{z\in\mathcal{Z}} \ell(z), \\ \infty & \text{otherwise,} \end{cases} \tag{4.10}$$

for all $\lambda_0, \lambda \in \mathbb{R}$. Indeed, the above reasoning and the definition of the perspective function $(\phi^*)^\pi$ ensure that (4.10) holds whenever $\lambda \neq 0$. Note that h^* is convex and closed thanks to Rockafellar (1970, Theorem 12.2). The expression on the right-hand side of (4.10) is also convex and closed in (λ_0, λ). In particular, it is lower semicontinuous thanks to Fatou's lemma, which applies because $\phi(1) = 0$ such that $(\phi^*)^\pi(t, \lambda) \geq t$ for all $t \in \mathbb{R}$ and $\lambda \in \mathbb{R}_+$ and because $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. Observe also that $(\phi^*)^\pi$ is proper, closed and convex thanks to Rockafellar (1970, pp. 35, 67, Theorem 13.3). Hence (4.10) must indeed hold for all $\lambda_0, \lambda \in \mathbb{R}$.

Given (4.10), we finally obtain

$$h^{**}(u_0, u) = \sup_{\lambda_0,\lambda\in\mathbb{R}} -\lambda_0 u_0 - \lambda u - h^*(-\lambda_0, -\lambda) = \begin{cases} \sup\limits_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} & -\lambda_0 u_0 - \lambda u - \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\ell(Z)-\lambda_0, \lambda)] \\ \text{s.t.} & \lambda_0 + \lambda\,\phi^\infty(1) \geq \sup\limits_{z\in\mathcal{Z}} \ell(z), \end{cases}$$

which establishes the desired formula for the bi-conjugate of h. It remains to be shown that if ϕ is continuous at 1, then $h(1, u) = h^{**}(1, u)$ for all $u \in \mathbb{R}_{++}$. However, this follows immediately from Lemmas 4.2 and 4.10. □

The following main theorem uses Proposition 4.13 to dualize the worst-case expectation problem (4.1) with a ϕ-divergence ambiguity set.

Theorem 4.14 (Duality theory for ϕ-divergence ambiguity sets). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. If $\mathcal{P}$ is the ϕ-divergence ambiguity set (2.10), then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf\limits_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} & \lambda_0 + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\ell(Z)-\lambda_0, \lambda)] \\ \text{s.t.} & \lambda_0 + \lambda\,\phi^\infty(1) \geq \sup\limits_{z\in\mathcal{Z}} \ell(z). \end{cases} \tag{4.11}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Here the product $\lambda\,\phi^\infty(1)$ is assumed to evaluate to ∞ if $\lambda = 0$ and $\phi^\infty(1) = \infty$. If additionally $r > 0$ and ϕ is continuous at 1, then strong duality holds, that is, the inequality (4.11) collapses to an equality.

Proof. Recall first that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = -h(1,r) \leq -h^{**}(1,r),$$

where the inequality holds because of Lemma 4.2. Weak duality thus follows from the first claim in Proposition 4.13. If ϕ is additionally continuous at 1, and if $r > 0$, then strong duality follows from the second claim in Proposition 4.13. □

Recall now that the *restricted* ϕ-divergence ambiguity set (2.11) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathbb{P} \ll \hat{\mathbb{P}},\ \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

That is, $\mathcal{P}$ contains all distributions from within the (unrestricted) ϕ-divergence ambiguity set (2.10) that are absolutely continuous with respect to $\hat{\mathbb{P}}$. The worst-case expected loss over $\mathcal{P}$ can again be expressed as $-h(1,r)$, where $h(u_0,u)$ is now defined as the infimum of the optimization problem (4.7) with the additional constraint $v \ll \hat{\mathbb{P}}$. One readily verifies that h remains convex and that $\{1\}\times\mathbb{R}_{++}$ is still contained in $\mathrm{rint}(\mathrm{dom}(h))$ despite this restriction. Indeed, the proof of Lemma 4.10 remains valid almost verbatim.

Theorem 4.15 (Duality theory for restricted ϕ-divergence ambiguity sets). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. If $\mathcal{P}$ is the restricted ϕ-divergence ambiguity set (2.11), then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \lambda_0 + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\ell(Z)-\lambda_0,\lambda)]. \tag{4.12}$$

If additionally $r > 0$ and ϕ is continuous at 1, then strong duality holds, that is, the inequality (4.12) collapses to an equality.

Note that if (λ_0,λ) is feasible in (4.12), then $(\ell(Z)-\lambda_0,\lambda)$ belongs $\hat{\mathbb{P}}$-almost surely to $\mathrm{dom}((\phi^*)^\pi)$. Otherwise, its objective function value equals ∞. In view of Lemma 4.12, this implies that $\lambda_0 + \lambda\,\phi^\infty(1) \geq \operatorname{ess\,sup}_{\hat{\mathbb{P}}}[\ell(Z)]$. In contrast, if (λ_0,λ) is feasible in (4.11), then it satisfies the constraint $\lambda_0 + \lambda\,\phi^\infty(1) \geq \sup_{z\in\mathcal{Z}} \ell(z)$, which is more restrictive unless $\phi^\infty(1) = \infty$. Hence the dual problem in (4.12) has a (weakly) larger feasible set and a (weakly) smaller infimum than the dual problem in (4.11). This is perhaps unsurprising because (4.12) corresponds to the worst-case expectation problem over the restricted ϕ-divergence ambiguity set, which is (weakly) smaller than the corresponding *un*restricted ϕ-divergence ambiguity set. Note also that the solution of a worst-case expectation problem over an unrestricted ϕ-divergence ambiguity set depends on $\mathcal{Z}$ and not just on the support of $\hat{\mathbb{P}}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Proof of Theorem 4.15. If $h(u_0, u)$ is defined as the infimum of the optimization problem (4.7) with the additional constraint $\nu \ll \hat{\mathbb{P}}$, then one can show that

$$h^{**}(u_0, u) = \sup_{\lambda_0, \lambda \in \mathbb{R}} -\lambda_0 u_0 - \lambda u - \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)'(\ell(Z) - \lambda_0, \lambda)].$$

Indeed, one can proceed as in the proof of Proposition 4.13. However, the reasoning simplifies significantly because the additional constraint $\nu \ll \hat{\mathbb{P}}$ allows us to set the dominating measure ρ in the definition of D_ϕ to $\hat{\mathbb{P}}$. Thus the Radon–Nikodym derivative $\beta = \mathrm{d}\hat{\mathbb{P}}/\mathrm{d}\rho$ is $\hat{\mathbb{P}}$-almost surely equal to 1. This in turn implies that the calculation of h^* requires no case distinction, that is, the set $\mathcal{Z}_0(\beta)$ is empty.

Given the bi-conjugate of h, both weak and strong duality can then be established exactly as in the proof of Theorem 4.14. Details are omitted for brevity. □

Van Parys *et al.* (2021, Proposition 5) establish a strong duality result for worst-case expectations over likelihood ambiguity sets as introduced in Section 2.2.2. Theorem 4.14 extends this result to general ϕ-divergence ambiguity sets with a significantly shorter proof that only uses tools from convex analysis. Ben-Tal *et al.* (2013) establish a strong duality result akin to Theorem 4.15 for restricted ϕ-divergence ambiguity sets under the assumption that the reference distribution $\hat{\mathbb{P}}$ is discrete. Shapiro (2017) extends this result to general reference distributions by using tools from infinite-dimensional analysis. In contrast, our proof of Theorem 4.15 establishes the same duality result using finite-dimensional convex analysis.

4.4. Optimal transport ambiguity sets

Recall from Section 2.3 that the optimal transport ambiguity set (2.27) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

Here $\mathcal{Z}$ is a closed support set, $r \geq 0$ is a size parameter, c is a transportation cost function in the sense of Definition 2.14, OT_c is the corresponding optimal transport discrepancy in the sense of Definition 2.15, and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. In analogy to Section 4.3, the worst-case expectation problem (4.1) over the ambiguity set (2.27) can now be reformulated as

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = -h(r),$$

where the auxiliary function $h \colon \mathbb{R} \to \overline{\mathbb{R}}$ is defined by

$$h(u) = \inf_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \{-\mathbb{E}_{\mathbb{P}}[\ell(Z)] \colon \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \leq u\}. \tag{4.13}$$

As the objective and constraint functions of the minimization problem in (4.13) are jointly convex in $\mathbb{P}$ and u, Lemma 4.1 implies that h is convex. Recall also that c is non-negative and satisfies $c(z, z) = 0$ for all $z \in \mathcal{Z}$. If $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$, it is therefore easy to show that $\mathrm{dom}(h) = \mathbb{R}_+$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The following lemma will be instrumental for deriving the bi-conjugate of h. Recall that $\Gamma(\mathbb{P}, \hat{\mathbb{P}})$ denotes the set of all couplings of $\mathbb{P}$ and $\hat{\mathbb{P}}$; see Definition 2.15.

Lemma 4.16 (Interchangeability principle). If c is a transportation cost function, ℓ is upper semicontinuous and $\lambda \geq 0$, then we have

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \sup_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} \mathbb{E}_\gamma[\ell(Z) - \lambda c(Z, \hat{Z})] = \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c(z, \hat{Z})\right].$$

One can show that Lemma 4.16 remains valid, for example, if $\mathcal{Z}$ is a Polish (separable metric) space equipped with its Borel σ-algebra and even if c and ℓ fail to be lower and upper semicontinuous, respectively (Zhang *et al.* 2024*b*, Proposition 1).

Proof of Lemma 4.16. Define $L\colon \mathcal{Z} \to \overline{\mathbb{R}}$ through $L(\hat{z}) = \sup_{z\in\mathcal{Z}} \ell(z) - \lambda c(z, \hat{z})$. If $\lambda = 1$, then L reduces to the c-transform of ℓ defined in (2.25). Note first that $\lambda c(z, \hat{z}) - \ell(z)$ is lower semicontinuous in $(z, \hat{z})$ and thus constitutes a normal integrand thanks to Rockafellar and Wets (2009, Example 14.31). This implies via Rockafellar and Wets (2009, Theorem 14.37) that $L(\hat{z})$ is Borel-measurable.

Observe next that, by the definition of L, we have $\ell(z) - \lambda c(z, \hat{z}) \leq L(\hat{z})$ for all $z, \hat{z} \in \mathcal{Z}$. This inequality persists if we integrate both sides with respect to any coupling $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$ for any distribution $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$, and therefore we obtain

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \sup_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} \mathbb{E}_\gamma[\ell(Z) - \lambda c(Z, \hat{Z})] \leq \mathbb{E}_{\hat{\mathbb{P}}}[L(\hat{Z})].$$

It remains to prove the reverse inequality. To this end, observe that

$$\begin{aligned}
\mathbb{E}_{\hat{\mathbb{P}}}[L(\hat{Z})] &= \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c(z, \hat{Z})\right] \\
&= \sup_{f\in\mathcal{F}} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(f(\hat{Z})) - \lambda c(f(\hat{Z}), \hat{Z})] \\
&\leq \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \sup_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} \mathbb{E}_\gamma[\ell(Z) - \lambda c(Z, \hat{Z})],
\end{aligned}$$

where $\mathcal{F}$ denotes the family of all Borel functions $f\colon \mathcal{Z} \to \mathcal{Z}$. The second equality follows from Rockafellar and Wets (2009, Theorem 14.60), which applies because $\lambda c(z, \hat{z}) - \ell(z)$ is a normal integrand. Note that the joint distribution of $f(\hat{Z})$ and $\hat{Z}$ under $\hat{\mathbb{P}}$ coincides with the pushforward distribution $\gamma = \hat{\mathbb{P}} \circ g^{-1}$, where $g\colon \mathcal{Z} \to \mathcal{Z} \times \mathcal{Z}$ is defined through $g(\hat{z}) = (f(\hat{z}), \hat{z})$. By construction, we have $\gamma \in \Gamma(\hat{\mathbb{P}} \circ f^{-1}, \hat{\mathbb{P}})$. The inequality in the above expression therefore holds because $\hat{\mathbb{P}} \circ f^{-1} \in \mathcal{P}(\mathcal{Z})$. This observation completes the proof. □

Proposition 4.17 (Bi-conjugate of h). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$ and that ℓ is upper semicontinuous. Then the bi-conjugate of h defined in (4.13) satisfies

$$h^{**}(u) = \sup_{\lambda\geq 0} -\lambda r - \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c(z, \hat{Z})\right].$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

In addition, h^{**} coincides with h on $\mathbb{R}_{++}$.

Proof. For any fixed $\lambda \in \mathbb{R}$, the conjugate of h satisfies

$$\begin{aligned} h^*(-\lambda) &= \sup_{u\in\mathbb{R}} \; -\lambda u - h(u) \\ &= \sup_{u\in\mathbb{R}_+, \mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{-\lambda u + \mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \le u\}, \end{aligned}$$

where the second equality holds because $\mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \ge 0$. As $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$, the resulting maximization problem is unbounded if $\lambda < 0$. If $\lambda > 0$, then we find

$$\begin{aligned} h^*(-\lambda) &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \; \mathbb{E}_{\mathbb{P}}[\ell(Z)] - \lambda \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \\ &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \; \sup_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} \; \mathbb{E}_{\mathbb{P}}[\ell(Z)] - \lambda \mathbb{E}_{\gamma}[c(Z, \hat{Z})] \\ &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \; \sup_{\gamma\in\Gamma(\mathbb{P},\hat{\mathbb{P}})} \; \mathbb{E}_{\gamma}[\ell(Z) - \lambda c(Z, \hat{Z})] \\ &= \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \; \ell(z) - \lambda c(z, \hat{Z})\right], \end{aligned} \tag{4.14}$$

where the second equality follows from Definition 2.15, the third equality holds because the marginal distribution of Z under γ is given by $\mathbb{P}$, and the fourth equality exploits Lemma 4.16. The above reasoning implies that $h^*(-\lambda)$ coincides with (4.14) for all $\lambda > 0$. However, this formula remains valid at $\lambda = 0$. To see this, note that h^* is convex and closed thanks to Rockafellar (1970, Theorem 12.2). The last expectation in (4.14) is also convex and closed in λ thanks to Fatou's lemma, which applies because $\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c(z, \hat{z})$ is larger than or equal to $\ell(\hat{z})$ and lower semicontinuous in λ for every $\hat{z} \in \mathcal{Z}$ and because $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$. Hence the last expectation in (4.14) is indeed convex and lower-semicontinuous in λ, and thus it indeed coincides with $h^*(-\lambda)$ for all $\lambda \in \mathbb{R}_+$.

Given (4.14), we finally obtain the following formula for the bi-conjugate of h:

$$h^{**}(u) = \sup_{\lambda\ge 0} \; -\lambda u - h^*(-\lambda) = \sup_{\lambda\ge 0} \; -\lambda u - \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \; \ell(z) - \lambda c(z, \hat{Z})\right].$$

Here the first equality holds because $h^*(-\lambda) = \infty$ whenever $\lambda < 0$. The second equality follows from (4.14), which holds for any $\lambda \ge 0$. This establishes the desired formula for h^{**}. Lemma 4.2 and our earlier observation that $\mathrm{dom}(h) = \mathbb{R}_+$ finally imply that $h(u) = h^{**}(u)$ for all $u \in \mathbb{R}_{++}$. □

The following main theorem uses Proposition 4.17 to dualize the worst-case expectation problem (4.1) with an optimal transport ambiguity set.

Theorem 4.18 (Duality theory for optimal transport ambiguity sets). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$ and ℓ is upper semicontinuous. If $\mathcal{P}$ is the optimal transport

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

ambiguity set defined in (2.27), then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \inf_{\lambda\in\mathbb{R}_+} \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c(z,\hat{Z})\right]. \tag{4.15}$$

If $r > 0$, then strong duality holds, that is, (4.15) collapses to an equality.

Proof. Recall first that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = -h(r) \leq -h^{**}(r),$$

where the inequality holds because of Lemma 4.2. Weak duality thus follows from the first claim in Proposition 4.17. If $r > 0$, then strong duality follows from the second claim in Proposition 4.17. This concludes the proof. □

Mohajerin Esfahani and Kuhn (2018) and Zhao and Guan (2018) use semi-infinite duality theory to prove Theorem 4.18 in the special case when OT_c is the 1-Wasserstein distance and when the reference distribution $\hat{\mathbb{P}}$ is discrete. Blanchet and Murthy (2019) and Gao and Kleywegt (2023) prove a generalization of Theorem 4.18 by leveraging a Fenchel duality theorem in Banach spaces and by devising a constructive argument, respectively. They both allow for arbitrary optimal transport discrepancies as well as arbitrary reference distributions on Polish spaces. The proof shown here, which exploits the interchangeability principle of Lemma 4.16 and elementary tools from convex analysis, is due to Zhang *et al.* (2024*b*).

5. Duality theory for worst-case risk problems

The standard DRO problem (1.2) assumes that the decision-maker is risk-neutral and ambiguity-averse. Risk-neutrality means that if the distribution of Z is known, then decisions are ranked by their *expected* loss. Ambiguity aversion means that if the distribution of Z is ambiguous, then expectations are evaluated under a distribution in the ambiguity set $\mathcal{P}$ that is *most detrimental* to the decision-maker.

If low-probability events have a disproportionate negative impact on the decision-maker, then it is *in*appropriate to use the expected loss as a decision criterion even if the distribution of Z is known. Instead, it is expedient to rank decisions by the *risk* of their loss with respect to a law-invariant risk measure. A law-invariant risk measure ϱ assigns each (univariate) loss distribution in $\mathcal{P}(\mathbb{R})$ a riskiness index. If the loss is representable as $\ell(Z)$, where $\ell\colon \mathbb{R}^d \to \mathbb{R}$ is a Borel function and Z is a d-dimensional random vector with probability distribution $\mathbb{P}$, then the distribution of the loss $\ell(Z)$ is given by the pushforward distribution $\mathbb{P}\circ\ell^{-1}$. Throughout this paper, we use $\varrho_{\mathbb{P}}[\ell(Z)]$ to denote the risk $\varrho(\mathbb{P}\circ\ell^{-1})$ of such a loss distribution. These conventions are formalized in the following definition. Here and in the remainder we use $\mathcal{L}(\mathbb{R}^d)$ to denote the family of all Borel functions $\ell\colon \mathbb{R}^d \to \mathbb{R}$.

Definition 5.1 (Law-invariant risk measure). A law-invariant risk measure is a function $\varrho\colon \mathcal{P}(\mathbb{R}) \to \overline{\mathbb{R}}$. We use $\varrho_{\mathbb{P}}[\ell(Z)]$ to denote $\varrho(\mathbb{P}\circ\ell^{-1})$ for any Borel function $\ell \in \mathcal{L}(\mathbb{R}^d)$, Borel distribution $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ and dimension $d \in \mathbb{N}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

A law-invariant risk measure ϱ has the property that if $\mathbb{P}_1 \circ \ell_1^{-1} = \mathbb{P}_2 \circ \ell_2^{-1}$ for two different Borel functions ℓ_1 and ℓ_2 and two different distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively, then $\varrho_{\mathbb{P}_1}[\ell_1(Z_1)] = \varrho_{\mathbb{P}_2}[\ell_2(Z_2)]$. In fact, this property is the very reason why ϱ is called 'law-invariant'.

The notation $\varrho_{\mathbb{P}}[\ell(Z)]$ is consistent with our usual conventions for the expected value $\mathbb{E}_{\mathbb{P}}[\ell(Z)]$, which is a special instance of a law-invariant risk measure. Also, it makes the dependence of the risk on $\mathbb{P}$ explicit, which is necessary when $\mathbb{P}$ is ambiguous. We stress that, in contrast to most of the literature on risk measures, our definition of a law-invariant risk measure ϱ is *not* tied to a particular probability space. A prime example of a law-invariant risk measure is the value-at-risk.

Definition 5.2 (Value-at-risk). The value-at-risk (VaR) at level $\beta \in (0, 1)$ of an uncertain loss $\ell(Z)$ with $\ell \in \mathcal{L}(\mathbb{R}^d)$ and $Z \sim \mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ is given by

$$\beta\text{-VaR}_{\mathbb{P}}[\ell(Z)] = \inf\{\tau \in \mathbb{R} \colon \mathbb{P}(\ell(Z) \leq \tau) \geq 1 - \beta\}. \tag{5.1}$$

The VaR is indeed law-invariant because $\mathbb{P}(\ell(Z) \leq \tau) = F(\tau)$ depends on ℓ and $\mathbb{P}$ only indirectly through the cumulative distribution function F associated with the pushfoward distribution $\mathbb{P} \circ \ell^{-1}$. Note that the infimum in (5.1) is attained because F is non-decreasing and right-continuous. By construction, the VaR at level β represents the smallest number $\tau^\star$ that weakly exceeds the loss with probability $1-\beta$. Thus it coincides with the leftmost $(1-\beta)$-quantile of the loss distribution F. For later reference we remark that the β-VaR can be reformulated as

$$\beta\text{-VaR}_{\mathbb{P}}[\ell(Z)] = \inf\{\tau \in \mathbb{R} \colon \mathbb{P}(\ell(Z) \geq \tau) \leq \beta\}. \tag{5.2}$$

However, the infimum in (5.2) may *not* be attained. Note that the VaR is well-defined and finite for *any* loss function $\ell \in \mathcal{L}(\mathbb{R}^d)$ and for *any* distribution $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$. Nonetheless, other law-invariant risk measures are finite only for certain sub-classes of loss functions and distributions. In the remainder of this paper we will often study risk measures that display some or all of the following structural properties.

Definition 5.3 (Properties of risk measures). A law-invariant risk measure ϱ is

(i) *translation-invariant* if

$$\varrho_{\mathbb{P}}[\ell(Z) + c] = \varrho_{\mathbb{P}}[\ell(Z)] + c \quad \forall \ell \in \mathcal{L}(\mathbb{R}^d),\ \forall c \in \mathbb{R},\ \forall \mathbb{P} \in \mathcal{P}(\mathbb{R}^d);$$

(ii) *scale-invariant* if

$$\varrho_{\mathbb{P}}[c\ell(Z)] = c\varrho_{\mathbb{P}}[\ell(Z)] \quad \forall \ell \in \mathcal{L}(\mathbb{R}^d),\ \forall c \in \mathbb{R}_+,\ \forall \mathbb{P} \in \mathcal{P}(\mathbb{R}^d);$$

(iii) *monotone* if

$$\varrho_{\mathbb{P}}[\ell_1(Z)] \leq \varrho_{\mathbb{P}}[\ell_2(Z)]$$
$$\forall \ell_1, \ell_2 \in \mathcal{L}(\mathbb{R}^d) \text{ with } \ell_1(Z) \leq \ell_2(Z)\ \ \mathbb{P}\text{-a.s.},\ \forall \mathbb{P} \in \mathcal{P}(\mathbb{R}^d);$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

(iv) *convex* if

$$\varrho_{\mathbb{P}}[\theta\ell_1(Z) + (1-\theta)\ell_2(Z)] \leq \theta\varrho_{\mathbb{P}}[\ell_1(Z)] + (1-\theta)\varrho_{\mathbb{P}}[\ell_2(Z)]$$
$$\forall \ell_1, \ell_2 \in \mathcal{L}(\mathbb{R}^d),\ \forall \theta \in [0,1],\ \forall \mathbb{P} \in \mathcal{P}(\mathbb{R}^d).$$

A *coherent* risk measure is translation-invariant, scale-invariant, monotone as well as convex (Artzner *et al.* 1999). In addition, a *convex* risk measure is translation-invariant, monotone and convex (but not necessarily scale-invariant).

Any law-invariant risk measure ϱ gives rise to a risk-averse DRO problem

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(x,Z)]. \tag{5.3}$$

This problem seeks a decision x that minimizes the worst-case risk of the random loss $\ell(x,Z)$ with respect to all distributions of Z in the ambiguity set $\mathcal{P}$. Below we will show that the duality theory for worst-case expectation problems developed in Section 4 has ramifications for a broad class of worst-case risk problems of the form

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)]. \tag{5.4}$$

Here we suppress as usual the dependence of the loss function on x to avoid clutter.

5.1. Optimized certainty equivalents

We now describe a class of law-invariant risk measures for which the *risk-averse* DRO problem (5.3) can be converted to an equivalent *risk-neutral* DRO problem of the form (1.2). This will show that many risk-averse DRO problems are susceptible to methods developed for risk-neutral problems. The risk measures studied in this section are induced by disutility functions in the sense of the following definition.

Definition 5.4 (Disutility function). A disutility function $g:\mathbb{R}\to\mathbb{R}$ is a convex (and therefore continuous) function with $g(0)=0$ and $g(\tau)>\tau$ for all $\tau\neq 0$.

Ben-Tal and Teboulle (1986) use disutility functions to construct a class of law-invariant risk measures, which they term optimized certainty equivalents. Recall that if the objective function of a minimization (maximization) problem can be expressed as the difference of two terms, both of which evaluate to ∞ (e.g. the positive and negative parts of an integral), then it should be interpreted as ∞ ($-\infty$).

Definition 5.5 (Optimized certainty equivalent). The optimized certainty equivalent induced by the disutility function g is the law-invariant risk measure ϱ with

$$\varrho_{\mathbb{P}}[\ell(Z)] = \inf_{\tau\in\mathbb{R}} \tau + \mathbb{E}_{\mathbb{P}}[g(\ell(Z)-\tau)]. \tag{5.5}$$

The expected disutility $\mathbb{E}_{\mathbb{P}}[g(\ell(Z))]$ represents a deterministic present loss that the decision-maker considers to be equally (un)desirable as the random future loss $\ell(Z)$. If it is possible to shift a deterministic portion τ of the loss $\ell(Z)$ to the present, then the decision-maker will solve the minimization problem in (5.5) in order to

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

strike an optimal trade-off between present and future losses. Hence it is natural to interpret $\varrho_{\mathbb{P}}[\ell(Z)]$ as an 'optimized certainty equivalent'.

There is also an intimate relation between optimized certainty equivalents and a class of ϕ-divergences. To see this, let ϕ be an entropy function in the sense of Definition 2.4 with $\phi^\infty(1) = \infty$. Assume also that ϕ is twice continuously differentiable on a neighbourhood of 1 with $\phi'(1) = 0$ and $\phi''(1) > 0$. Under these conditions, ϕ^* constitutes a disutility function in the sense of Definition 5.4. Indeed, ϕ^* is real-valued because $\phi^\infty(1) = \infty$ and satisfies $\phi^*(t) \geq t$ for all $t \in \mathbb{R}$ because $\phi(1) = 0$. Finally, we have $\phi^*(0) = 0$ because $\phi'(1) = 0$ and $\phi^*(t) > t$ for all $t \neq 0$ because $\phi''(t) > 0$. If $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$, then the optimized certainty equivalent induced by the disutility function $g = \phi^*$ satisfies

$$\inf_{\lambda_0 \in \mathbb{R}} \lambda_0 + \mathbb{E}_{\hat{\mathbb{P}}}[\phi^*(\ell(Z) - \lambda_0)] = \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(Z)] - \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \tag{5.6}$$

and thus coincides with the optimal value of a penalty-based distributionally robust optimization model with a ϕ-divergence penalty. The equality in the above expression follows from Ben-Tal and Teboulle (2007, Theorem 4.2), which is reminiscent of the strong duality theorem for worst-case expectation problems over restricted ϕ-divergence ambiguity sets (see Theorem 4.15). The assumption that $\phi^\infty(1) = \infty$ indeed ensures that $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}})$ is finite only if $\mathbb{P} \ll \hat{\mathbb{P}}$. We also remark that if g is a disutility function in the sense of Definition 5.4 and if g is non-decreasing, then g^* constitutes an entropy function in the sense of Definition 2.4.

We will see below that the optimized certainty equivalents encapsulate several widely used risk measures as special cases. Notable examples include the mean–variance risk measure, the mean–median risk measure, the conditional value-at-risk or the entropic risk measure. More generally, Rockafellar, Uryasev and Zabarankin (2006, 2008) show that virtually any regular risk measure admits a representation of the form (5.5) provided that the expected disutility is replaced with a more general measure of regret; see also Rockafellar and Royset (2014, 2015) and the survey papers by Rockafellar and Royset (2013) and Royset (2022).

Definition 5.6 (Mean–variance risk measure). The mean–variance risk measure with risk-aversion coefficient $\beta \in (0, \infty)$ is the law-invariant risk measure ϱ with

$$\varrho_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\mathbb{P}}[\ell(Z)] + \beta \cdot \mathbb{V}_{\mathbb{P}}[\ell(Z)],$$

where $\mathbb{V}_{\mathbb{P}}[\ell(Z)]$ denotes the variance of $\ell(Z)$ under $\mathbb{P}$.

We call a function $f : \mathbb{R} \to \overline{\mathbb{R}}$ coercive if $\lim_{i\to\infty} f(\tau_i) = \infty$ for every sequence $\{\tau_i\}_{i\in\mathbb{N}}$ with $\lim_{i\to\infty} |\tau_i| = \infty$. Coercivity will play a key role in re-expressing worst-case optimized certainty equivalents in terms of worst-case expectations.

Proposition 5.7 (Mean–variance risk measure). The mean–variance risk measure ϱ with risk-aversion coefficient $\beta \in (0, \infty)$ is the optimized certainty equivalent

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

induced by the disutility function $g(\tau) = \tau + \beta\tau^2$. The objective function of problem (5.5) is coercive in τ and is uniquely minimized by $\tau^\star = \mathbb{E}_\mathbb{P}[\ell(Z)]$.

Proof. The objective function of problem (5.5) corresponding to the disutility function g is given by $\mathbb{E}_\mathbb{P}[\ell(Z) + \beta(\ell(Z) - \tau)^2]$. This function is ostensibly coercive in τ and is minimized by $\tau^\star = \mathbb{E}_\mathbb{P}[\ell(Z)]$. Substituting $\tau^\star$ back into the objective function shows that the optimized certainty equivalent induced by g indeed coincides with the mean–variance risk measure with risk-aversion coefficient β. □

Definition 5.8 (Mean–MAD risk measure). The mean–median absolute deviation (MAD) risk measure with risk-aversion coefficient $\beta \in (0, \infty)$ is the law-invariant risk measure ϱ with

$$\varrho_\mathbb{P}[\ell(Z)] = \mathbb{E}_\mathbb{P}[\ell(Z)] + \beta \cdot \mathbb{E}_\mathbb{P}[|\ell(Z) - \mathbb{M}_\mathbb{P}[\ell(Z)]|],$$

where $\mathbb{M}_\mathbb{P}[\ell(Z)]$ denotes the median of $\ell(Z)$ under $\mathbb{P}$.

Proposition 5.9 (Mean–MAD risk measure). The mean–MAD risk measure ϱ with risk-aversion coefficient $\beta \in (0, \infty)$ is the optimized certainty equivalent induced by the disutility function $g(\tau) = \tau + \beta|\tau|$. The objective function of problem (5.5) is coercive in τ and is minimized by $\tau^\star = \mathbb{M}_\mathbb{P}[\ell(Z)]$.

Proof. The objective function of problem (5.5) corresponding to the disutility function g is given by $\mathbb{E}_\mathbb{P}[\ell(Z) + \beta|\ell(Z) - \tau|]$. This function is ostensibly coercive in τ and is minimized by $\tau^\star = \mathbb{M}_\mathbb{P}[\ell(Z)]$. Substituting $\tau^\star$ back into the objective function yields the mean–MAD risk measure with risk-aversion coefficient β. □

Definition 5.10 (Conditional value-at-risk). The conditional VaR (CVaR) at level $\beta \in (0, 1)$ is the law-invariant risk measure denoted as β-CVaR with

$$\beta\text{-CVaR}_\mathbb{P}[\ell(Z)] = \inf_{\tau \in \mathbb{R}} \tau + \frac{1}{\beta}\mathbb{E}_\mathbb{P}[\max\{\ell(Z) - \tau, 0\}]. \tag{5.7}$$

Note that β-$\text{CVaR}_\mathbb{P}[\ell(Z)]$ converges to $\mathbb{E}_\mathbb{P}[\ell(Z)]$ as β tends to 1. One can further show that it converges to the essential supremum $\operatorname{ess\,sup}_\mathbb{P}[\ell(Z)]$ as β tends to 0.

Proposition 5.11 (CVaR). The CVaR at level $\beta \in (0, 1)$ is the optimized certainty equivalent induced by the disutility function $g(\tau) = \beta^{-1}\max\{\tau, 0\}$. The objective function of problem (5.5) is coercive in τ and is minimized by $\tau^\star = \beta\text{-VaR}_\mathbb{P}[\ell(Z)]$.

Proof. It is evident that problem (5.7) is an instance of problem (5.5) corresponding to the given disutility function g. In addition, as $\beta \in (0, 1)$, it is evident that the objective function of problem (5.7) is coercive in τ. Finally, one readily verifies that $\tau^\star = \beta\text{-VaR}_\mathbb{P}[\ell(Z)]$ solves the first-order optimality condition of the unconstrained convex program (5.7) and thus constitutes a minimizer. □

By substituting $\tau^\star = \beta\text{-VaR}_\mathbb{P}[\ell(Z)]$ into the objective function of problem (5.7), it becomes now clear that $\beta\text{-CVaR}_\mathbb{P}[\ell(Z)] \geq \beta\text{-VaR}_\mathbb{P}[\ell(Z)]$. If the loss $\ell(Z)$ has a

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

continuous distribution under $\mathbb{P}$, then one can further use (5.7) to show that

$$\beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\mathbb{P}}[\ell(Z) \mid \ell(Z) \geq \beta\text{-VaR}_{\mathbb{P}}[\ell(Z)]].$$

Hence the CVaR at level β coincides with the expectation of the upper β-tail of the loss distribution, which implies that β-$\text{CVaR}_{\mathbb{P}}[\ell(Z)]$ is generically *strictly* larger than β-$\text{VaR}_{\mathbb{P}}[\ell(Z)]$. For details we refer to Rockafellar and Uryasev (2000, 2002).

Definition 5.12 (Entropic risk measure). The entropic risk measure with risk-aversion parameter $\beta \in (0, \infty)$ is the law-invariant risk measure denoted as β-ERM with

$$\beta\text{-ERM}_{\mathbb{P}}[\ell(Z)] = \frac{1}{\beta} \log \mathbb{E}_{\mathbb{P}}[\exp(\beta \ell(Z))]. \tag{5.8}$$

Using a Taylor expansion, one can show that β-$\text{ERM}_{\mathbb{P}}[\ell(Z)]$ converges to the expected value $\mathbb{E}_{\mathbb{P}}[\ell(Z)]$ as β tends to 0. Similarly, one can show that β-$\text{ERM}_{\mathbb{P}}[\ell(Z)]$ converges to the essential supremum $\operatorname{ess\,sup}_{\mathbb{P}}[\ell(Z)]$ as β tends to ∞.

Proposition 5.13 (Entropic risk measure). The entropic risk measure with risk-aversion parameter $\beta \in (0, 1)$ is the optimized certainty equivalent induced by the disutility function $g(\tau) = \beta^{-1}(\exp(\beta\tau) - 1)$. The objective function of problem (5.5) is coercive in τ and is minimized by $\tau^\star = \beta^{-1} \log(\mathbb{E}_{\mathbb{P}}[\exp(\beta \ell(Z))])$.

Proof. By the definition of g, we have

$$\begin{aligned}
\inf_{\tau \in \mathbb{R}} \tau + \mathbb{E}_{\mathbb{P}}[g(\ell(Z) - \tau)] &= \inf_{\tau \in \mathbb{R}} \tau + \frac{1}{\beta} \mathbb{E}_{\mathbb{P}}[\exp(\beta(\ell(Z) - \tau)) - 1] \\
&= \frac{1}{\beta} \log \mathbb{E}_{\mathbb{P}}[\exp(\beta \ell(Z))] \\
&= \beta\text{-ERM}_{\mathbb{P}}[\ell(Z)].
\end{aligned}$$

The second equality holds because the unconstrained convex minimization problem over τ is uniquely solved by $\tau^\star = \beta^{-1} \log(\mathbb{E}_{\mathbb{P}}[\exp(\beta \ell(Z))])$, which can be verified by inspecting the problem's first-order optimality condition. In addition, as $\beta \in (0, 1)$, it is clear that the problem's objective function is coercive in τ. □

Kupper and Schachermayer (2009) show that, with the exception of the expected value, the entropic risk measure is the only relevant law-invariant risk measure that obeys the tower property. That is, for any random vectors Z_1 and Z_2 it satisfies

$$\beta\text{-ERM}_{\mathbb{P}}[\ell(Z_2)] = \beta\text{-ERM}_{\mathbb{P}}[\beta\text{-ERM}_{\mathbb{P}}[\ell(Z_2) \mid Z_1]],$$

where the *conditional* entropic risk measure β-$\text{ERM}_{\mathbb{P}}[Z_1 \mid Z_2]$ is defined in the obvious way by replacing the unconditional expectation in (5.8) with a conditional expectation. The entropic risk measure is often used for modelling risk-aversion in *dynamic* optimization problems, where the dynamic consistency of the decisions taken at different points in time is a concern. For example, it occupies centre stage in finance (Föllmer and Schied 2008), risk-sensitive control (Whittle 1990, Başar and Bernhard 1995) and economics (Hansen and Sargent 2008).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Proposition 5.14 (Dual representation of the entropic risk measure). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. Then the entropic risk measure admits the dual representation

$$\beta\text{-ERM}_{\hat{\mathbb{P}}}[\ell(Z)] = \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(Z)] - \frac{1}{\beta}\cdot \mathrm{KL}(\mathbb{P},\hat{\mathbb{P}}).$$

Proof. Let ϕ be the entropy function of the Kullback–Leibler divergence. Thus we have $\phi^*(t) = \mathrm{e}^t - 1$ for all $t \in \mathbb{R}$; see Table 4.1. By Proposition 5.13, the entropic value-at-risk is the optimized certainty equivalent induced by the disutility function

$$g(t) = \beta^{-1}(\exp(\beta t) - 1) = \beta^{-1}\phi^*(\beta t) = (\beta^{-1}\phi)^*(t),$$

where the last equality uses Theorem 16.1 of Rockafellar (1970). This implies that

$$\begin{aligned}\beta\text{-ERM}_{\hat{\mathbb{P}}}[\ell(Z)] &= \inf_{\tau\in\mathbb{R}} \tau + \mathbb{E}_{\mathbb{P}}[(\beta^{-1}\phi)^*(\ell(Z)-\tau)]\\ &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(Z)] - \mathrm{D}_{\beta^{-1}\phi}(\mathbb{P},\hat{\mathbb{P}})\\ &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(Z)] - \beta^{-1}\mathrm{KL}(\mathbb{P},\hat{\mathbb{P}}).\end{aligned}$$

Here the second equality follows from the strong duality relation (5.6), which applies because $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$, and the third equality holds because the entropy function ϕ was assumed to induce the Kullback–Leibler divergence. □

We remark that Proposition 5.14 can also be proved by leveraging the Donsker–Varadhan formula from Proposition 2.9 in lieu of the duality relation (5.6).

One can show that every optimized certainty equivalent ϱ is translation-invariant and convex. If the underlying disutility function g is non-decreasing, then ϱ is also monotone, and if g is positive homogeneous, then ϱ is also scale-invariant.

In the remainder we will show that if ϱ is any optimized certainty equivalent, then the worst-case risk problem (5.4) can be reduced to a worst-case expectation problem of the form (4.1). This reduction is predicated on a lopsided minimax theorem to be derived below, and it allows us to extend the duality theory for worst-case expectation problems of Section 4 to a rich class of worst-case risk problems.

5.2. *Lopsided minimax theorems*

A generic minimax problem can be represented as

$$\inf_{u\in\mathcal{U}} \sup_{v\in\mathcal{V}} H(u,v),$$

where $\mathcal{U}$ and $\mathcal{V}$ are arbitrary spaces, and $H\colon \mathcal{U}\times\mathcal{V}\to\overline{\mathbb{R}}$ is an arbitrary function. A minimax theorem provides conditions under which the infimum and supremum operators can be interchanged without changing the problem's optimal value. The following minimax theorem inspired by Rockafellar (1974, Example 13) will be

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

essential for solving worst-case risk problems with optimized certainty equivalents. Recall from Section 4.1 that a convex function is closed if it is either proper and lower semicontinuous or identically equal to $+\infty$ or to $-\infty$.

Theorem 5.15 (Lopsided minimax theorem). Suppose that $\mathcal{U}$ is an arbitrary vector space and $\mathcal{V}$ is a locally convex topological vector space. Assume also that the function $H\colon \mathcal{U} \times \mathcal{V} \to \overline{\mathbb{R}}$ is such that $H(u, v)$ is convex in u and such that $-H(u, v)$ is convex and closed in v. If $\sup_{v\in\mathcal{V}} \inf_{u\in\mathcal{U}} H(u, v) > -\infty$ and for every $\alpha \in \mathbb{R}$ there exists $u \in \mathcal{U}$ such that $\{v \in \mathcal{V}\colon H(u, v) \geq \alpha\}$ is compact, then we have

$$\inf_{u\in\mathcal{U}} \sup_{v\in\mathcal{V}} H(u, v) = \sup_{v\in\mathcal{V}} \inf_{u\in\mathcal{U}} H(u, v).$$

Proof. Let $\mathcal{V}^*$ be the topological dual of $\mathcal{V}$, and define the bilinear form $\langle\cdot, \cdot\rangle\colon \mathcal{V}^* \times \mathcal{V} \to \mathbb{R}$ through $\langle v^*, v\rangle = v^*(v)$. If we equip $\mathcal{V}^*$ with the weak topology induced by $\mathcal{V}$, then $\langle\cdot, v\rangle$ is a continuous linear functional on $\mathcal{V}^*$ for every $v \in \mathcal{V}$, and every continuous linear functional on $\mathcal{V}^*$ can be represented in this way.

Define $F\colon \mathcal{U} \times \mathcal{V}^* \to \overline{\mathbb{R}}$ through $F(u, v^*) = \sup_{v\in\mathcal{V}} H(u, v) - \langle v^*, v\rangle$, which is jointly convex in u and v^* thanks to Lemma 4.1. Thus $F(u, v^*) = (-H)^*(u, -v^*)$, where the conjugate of $-H(u, v)$ is evaluated with respect to its second argument v only. As $-H(u, v)$ is convex and closed in v, this implies via Lemma 4.2 that $F^*(u, v) = -H(u, -v)$. Here, again, the conjugate of $F(u, v^*)$ is evaluated with respect to its second argument v^* only. In addition, define $h\colon \mathcal{V}^* \to \overline{\mathbb{R}}$ through $h(v^*) = \inf_{u\in\mathcal{U}} F(u, v^*)$, which is convex in v^*. Thus we find

$$h(0) = \inf_{u\in\mathcal{U}} F(u, 0) = \inf_{u\in\mathcal{U}} \sup_{v\in\mathcal{V}} H(u, v),$$

where the two equalities follow from the definitions of h and F, respectively. In addition, we also have

$$\begin{aligned} h^{**}(0) &= \sup_{v\in\mathcal{V}} -h^*(-v) \\ &= \sup_{v\in\mathcal{V}} \inf_{v^*\in\mathcal{V}^*} \langle v^*, v\rangle + h(v^*) \\ &= \sup_{v\in\mathcal{V}} \inf_{u\in\mathcal{U}} \inf_{v^*\in\mathcal{V}^*} \langle v^*, v\rangle + F(u, v^*) \\ &= \sup_{v\in\mathcal{V}} \inf_{u\in\mathcal{U}} -F^*(u, -v) \\ &= \sup_{v\in\mathcal{V}} \inf_{u\in\mathcal{U}} H(u, v), \end{aligned}$$

where the first two equalities follow from the definitions of the bi-conjugate h^{**} and the conjugate h^*, respectively, and the third equality exploits the definition of h. The fourth equality follows from the definition of the conjugate F^*, and the last equality holds because $F^*(u, v) = -H(u, -v)$. Thus the desired minimax result holds if we manage to prove that $h(0) = h^{**}(0)$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

By the definitions of h^* and h and by the relation $F^*(u,v) = -H(u,-v)$, we have

$$\begin{aligned}\{v \in \mathcal{V}\colon h^*(v) \leq \alpha\} &= \left\{v \in \mathcal{V}\colon \sup_{v^* \in \mathcal{V}^*} \langle v^*, v\rangle - h(v^*) \leq \alpha\right\} \\ &= \left\{v \in \mathcal{V}\colon \sup_{u \in \mathcal{U}} \sup_{v^* \in \mathcal{V}^*} \langle v^*, v\rangle - F(u,v^*) \leq \alpha\right\} \\ &= \left\{v \in \mathcal{V}\colon \sup_{u \in \mathcal{U}} -H(u,-v) \leq \alpha\right\} \\ &= -\bigcap_{u \in \mathcal{U}} \{v \in \mathcal{V}\colon H(u,v) \geq -\alpha\}\end{aligned}$$

for any $\alpha \in \mathbb{R}$. Hence $\{v \in \mathcal{V}\colon h^*(v) \leq \alpha\}$ is representable as an intersection of closed sets, at least one of which is compact. Therefore the intersection is also compact. Selecting $\alpha > \inf_{v \in \mathcal{V}} h^*(v)$, which is possible because $\sup_{v \in \mathcal{V}} \inf_{u \in \mathcal{U}} H(u,v) > -\infty$ implies that $\inf_{v \in \mathcal{V}} h^*(v) < \infty$, we further ensure that the compact set $\{v \in \mathcal{V}\colon h^*(v) \leq \alpha\}$ is non-empty. This implies via Rockafellar (1974, Theorem 10 (b)) that $h^{**}(v^*)$ and $h(v^*)$ are both bounded above on a neighbourhood of 0. By Rockafellar (1974, Theorem 17 (a)), this in turn implies that $h(0) = h^{**}(0)$, which establishes the desired minimax equality. □

Swapping the roles of u and v leads to the following immediate corollary.

Corollary 5.16 (Reverse lopsided minimax theorem). Suppose that $\mathcal{U}$ is a locally convex topological vector space and $\mathcal{V}$ is an arbitrary vector space. Assume also that the function $H\colon \mathcal{U} \times \mathcal{V} \to \overline{\mathbb{R}}$ is such that $H(u,v)$ is convex and closed in u and such that $-H(u,v)$ is convex in v. If $\inf_{u \in \mathcal{U}} \sup_{v \in \mathcal{V}} H(u,v) < \infty$ and for every $\alpha \in \mathbb{R}$ there exists $v \in \mathcal{V}$ such that $\{u \in \mathcal{U}\colon H(u,v) \leq \alpha\}$ is compact, then we have

$$\inf_{u \in \mathcal{U}} \sup_{v \in \mathcal{V}} H(u,v) = \sup_{v \in \mathcal{V}} \inf_{u \in \mathcal{U}} H(u,v).$$

A function $h_v(u) = H(u,v)$ whose sublevel sets $\{u \in \mathcal{U}\colon h_v(u) \leq \alpha\}$ are all compact is commonly referred to as *inf-compact* (Hartung 1982). The following lemma provides an easily checkable sufficient condition for the inf-compactness of $h_v(u)$ when $\mathcal{U}$ is a Euclidean space. To this end, recall that a function h_v is *coercive* if, for every sequence $\{u_i\}_{i \in \mathbb{N}}$ with $\lim_{i \to \infty} \|u_i\|_2 = \infty$, we have $\lim_{i \to \infty} h_v(u_i) = \infty$.

Lemma 5.17 (Inf-compactness). Suppose that $\mathcal{U}$ is a Euclidean space and $H\colon \mathcal{U} \times \mathcal{V} \to \overline{\mathbb{R}}$ is lower semicontinuous and coercive in its first argument. Then the sublevel sets $\{u \in \mathcal{U}\colon H(u,v) \leq \alpha\}$ are compact for all $v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$.

Proof. To show that the sublevel set $\mathcal{U}_\alpha(v) = \{u \in \mathcal{U}\colon H(u,v) \leq \alpha\}$ is compact, note first that $\mathcal{U}_\alpha(v)$ is closed because $H(u,v)$ is lower semicontinuous in u. In order to prove that $\mathcal{U}_\alpha(v)$ is also bounded, assume for the sake of contradiction that there exists a sequence $\{u_i\}_{i \in \mathbb{N}} \in \mathcal{U}_\alpha(v)$ with $\lim_{i \to \infty} \|u_i\| = \infty$. As $H(u,v)$ is coercive in u, we have $\lim_{i \to \infty} H(u_i,v) = \infty$. However, this contradicts the assumption that $H(u_i,v) \leq \alpha$ for all $i \in \mathbb{N}$. Thus $\mathcal{U}_\alpha(v)$ must be bounded and compact. □

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Note that if $H_0 : \mathcal{U}_0 \times \mathcal{V}_0 \to \overline{\mathbb{R}}$ is defined on convex sets $\mathcal{U}_0 \subseteq \mathcal{U}$ and $\mathcal{V}_0 \subseteq \mathcal{V}$, then it can be extended to a function $H : \mathcal{U} \times \mathcal{V} \to \overline{\mathbb{R}}$ on the underlying vector spaces $\mathcal{U}$ and $\mathcal{V}$ by setting

$$H(u,v) = \begin{cases} H_0(u,v) & \text{if } u \in \mathcal{U}_0 \text{ and } v \in \mathcal{V}_0, \\ +\infty & \text{if } u \notin \mathcal{U}_0 \text{ and } v \in \mathcal{V}_0, \\ -\infty & \text{if } v \notin \mathcal{V}_0. \end{cases}$$

This construction guarantees that

$$\inf_{u\in\mathcal{U}} \sup_{v\in\mathcal{V}} H(u,v) = \inf_{u\in\mathcal{U}_0} \sup_{v\in\mathcal{V}_0} H_0(u,v) \quad \text{and} \quad \sup_{v\in\mathcal{V}} \inf_{u\in\mathcal{U}} H(u,v) = \sup_{v\in\mathcal{V}_0} \inf_{u\in\mathcal{U}_0} H_0(u,v).$$

It also guarantees that if $H_0(u,v)$ is convex and closed in u and concave in v, then so is $H(u,v)$. Thus the feasible sets in any convex–concave minimax problem can always be extended to the underlying vector spaces without changing the problem.

We now leverage Corollary 5.16 to derive a minimax theorem for optimized certainty equivalents. This result exploits the inf-compactness of the objective function of problem (5.5) in τ. Shafiee and Kuhn (2024) establish similar minimax theorems for a more general class of regular risk and deviation measures introduced by Rockafellar and Uryasev (2013).

Theorem 5.18 (Minimax theorem for optimized certainty equivalents). Suppose that $\mathcal{P} \subseteq \mathcal{P}(\mathcal{Z})$ is non-empty and convex, ϱ is any optimized certainty equivalent induced by a disutility function g, $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[g(\ell(Z))] < \infty$, and $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}$. Then $G(\tau,\mathbb{P}) = \tau + \mathbb{E}_{\mathbb{P}}[g(\ell(Z) - \tau)]$ for $\tau \in \mathbb{R}$ and $\mathbb{P} \in \mathcal{P}$ satisfies

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] = \sup_{\mathbb{P}\in\mathcal{P}} \inf_{\tau\in\mathbb{R}} G(\tau,\mathbb{P}) = \inf_{\tau\in\mathbb{R}} \sup_{\mathbb{P}\in\mathcal{P}} G(\tau,\mathbb{P}).$$

Proof. Note first that $G(\tau,\mathbb{P})$ is convex in τ and concave (in fact, linear) in $\mathbb{P}$. In addition, $G(\tau,\mathbb{P})$ is closed in τ. To see this, observe that

$$\begin{aligned} \liminf_{\tau'\to\tau} G(\tau',\mathbb{P}) &= \liminf_{\tau'\to\tau} \mathbb{E}_{\mathbb{P}}[\tau' + g(\ell(Z) - \tau')] \\ &\geq \mathbb{E}_{\mathbb{P}}\left[\liminf_{\tau'\to\tau} \tau' + g(\ell(Z) - \tau')\right] \\ &\geq \mathbb{E}_{\mathbb{P}}[\tau + g(\ell(Z) - \tau)] \\ &= G(\tau,\mathbb{P}), \end{aligned}$$

where the two inequalities follow from Fatou's lemma and the continuity of g, respectively. Fatou's lemma applies because any disutility function satisfies $g(\tau) \geq \tau$ for all $\tau \in \mathbb{R}$, which implies that $\tau + g(\ell(z) - \tau) \geq \ell(z)$ for all $z \in \mathcal{Z}$ and $\tau \in \mathbb{R}$. Note also that $\mathbb{E}_{\mathbb{P}}[\ell(Z)]$ is finite by assumption. Next, we show that $G(\tau,\mathbb{P})$ is inf-compact in τ. To this end, recall that $g(0) = 0$ and $g(\tau) > \tau$ for all $\tau \neq 0$. As g is also convex, this implies that $g(\tau)$ must grow faster than τ as τ tends to $+\infty$ and that $g(\tau)$ must decay more slowly than τ as τ tends to $-\infty$. Hence there exists

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$\varepsilon > 0$ with $g(\tau) \geq (1+\varepsilon)\tau - 1$ and $g(\tau) \geq (1-\varepsilon)\tau - 1$ for all $\tau \in \mathbb{R}$. For a formal proof of this assertion we refer to Zhen *et al.* (2023, Lemma C.10). This implies that

$$G(\tau, \mathbb{P}) \geq \tau + (1+\varepsilon)(\mathbb{E}_{\mathbb{P}}[\ell(Z)] - \tau) - 1 = -\varepsilon\tau + (1+\varepsilon)\mathbb{E}_{\mathbb{P}}[\ell(Z)] - 1$$

and

$$G(\tau, \mathbb{P}) \geq \tau + (1-\varepsilon)(\mathbb{E}_{\mathbb{P}}[\ell(Z)] - \tau) - 1 = \varepsilon\tau + (1+\varepsilon)\mathbb{E}_{\mathbb{P}}[\ell(Z)] - 1$$

for all $\tau \in \mathbb{R}$, and thus $\{\tau \in \mathbb{R} \colon G(\tau, \mathbb{P}) \leq \alpha\}$ is compact for every $\alpha \in \mathbb{R}$.

Next, set $\mathcal{U} = \mathbb{R}$, and define $\mathcal{V} = \mathcal{M}(\mathbb{R}^d)$ as the space of all finite signed Borel measures on $\mathbb{R}^d$. In addition, define the function $H \colon \mathcal{U} \times \mathcal{V} \to \overline{\mathbb{R}}$ through

$$H(u, v) = \begin{cases} G(u, v) & \text{if } v \in \mathcal{P}, \\ -\infty & \text{if } v \notin \mathcal{P}. \end{cases}$$

By construction, $H(u, v)$ is convex and closed in u and concave in v. Recall from Section 4.1 that a convex function is closed if it is either proper and lower semicontinuous or identically equal to $+\infty$ or to $-\infty$. In addition, we have

$$\sup_{v \in \mathcal{V}} H(0, v) = \sup_{\mathbb{P} \in \mathcal{P}} G(0, \mathbb{P}) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[g(\ell(Z))] < \infty,$$

and the sublevel sets $\{u \in \mathcal{U} \colon H(u, v) \leq \alpha\}$ are compact for every $\alpha \in \mathbb{R}$ provided that $v \in \mathcal{P}$. The claim thus follows from Corollary 5.16. □

Theorem 5.18 implies that if $\beta \in (0, 1)$, then the worst-case β-CVaR satisfies

$$\sup_{\mathbb{P} \in \mathcal{P}} \beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)] = \inf_{\tau \in \mathbb{R}} \tau + \frac{1}{\beta} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\max\{\ell(Z) - \tau, 0\}] \tag{5.9}$$

for *any* non-empty convex ambiguity set $\mathcal{P} \subseteq \mathcal{P}(\mathcal{Z})$ provided that $\mathbb{E}_{\mathbb{P}}[|\ell(Z)|] < \infty$ for all $\mathbb{P} \in \mathcal{P}$. In the extant literature, the interchange of the supremum over $\mathbb{P}$ and the infimum over τ is often justified with Sion's minimax theorem (Sion 1958). However, many studies overlook that Sion's minimax theorem only applies if $\mathcal{P}$ is weakly compact and $\mathbb{E}_{\mathbb{P}}[\max\{\ell(Z) - \tau, 0\}]$ is weakly upper semicontinuous in $\mathbb{P}$. As shown in Section 3, unfortunately, many popular ambiguity sets fail to be weakly compact. In addition, $\mathbb{E}_{\mathbb{P}}[\max\{\ell(Z) - \tau, 0\}]$ fails to be weakly upper semicontinuous unless the loss function ℓ is upper semicontinuous and bounded on $\mathcal{Z}$; see Proposition 3.3. All non-trivial convex loss functions on $\mathbb{R}^d$ violate this condition. In contrast, Theorem 5.18 offers a more general result that exploits the inf-compactness in τ but obviates any restrictive topological conditions on $\mathcal{P}$ or ℓ.

5.3. Moment ambiguity sets

Recall that the generic moment ambiguity set (2.1) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}_f(\mathcal{Z}) \colon \mathbb{E}_{\mathbb{P}}[f(Z)] \in \mathcal{F}\},$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where $\mathcal{Z} \subseteq \mathbb{R}^d$ is a non-empty closed support set, $f\colon \mathcal{Z} \to \mathbb{R}^m$ is a Borel-measurable moment function, $\mathcal{F} \subseteq \mathbb{R}^m$ is a non-empty closed moment uncertainty set, and $\mathcal{P}_f(\mathcal{Z})$ denotes the family of all distributions $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ for which $\mathbb{E}_{\mathbb{P}}[f(Z)]$ is finite. Recall also that $\mathcal{C} = \{\mathbb{E}_{\mathbb{P}}[f(Z)] : \mathbb{P} \in \mathcal{P}_f(\mathcal{Z})\}$ represents the family of all possible moments of any distribution on $\mathcal{Z}$. The next theorem establishes a duality result for the worst-case risk problem (5.4) with a moment ambiguity set.

Theorem 5.19 (Duality theory for moment ambiguity sets II). If $\mathcal{P}$ is the moment ambiguity set (2.1) and ϱ is an optimized certainty equivalent induced by a disutility function g, then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf & \tau + \lambda_0 + \delta^*_{\mathcal{F}}(\lambda) \\ \text{s.t.} & \tau, \lambda_0 \in \mathbb{R},\ \lambda \in \mathbb{R}^m \\ & \lambda_0 + f(z)^\top \lambda \geq g(\ell(z) - \tau) \ \ \forall z \in \mathcal{Z} \end{cases} \tag{5.10}$$

If $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[g(\ell(Z))] < \infty$, $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}_f(\mathcal{Z})$, and $\mathcal{F} \subseteq \mathcal{C}$ is a convex and compact set with $\operatorname{rint}(\mathcal{F}) \subseteq \operatorname{rint}(\mathcal{C})$, then strong duality holds, that is, the inequality (5.10) becomes an equality.

Proof. The max-min inequality implies that

$$\begin{aligned} \sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] &= \sup_{\mathbb{P}\in\mathcal{P}} \inf_{\tau\in\mathbb{R}} \tau + \mathbb{E}_{\mathbb{P}}[g(\ell(Z) - \tau)] \\ &\leq \inf_{\tau\in\mathbb{R}} \sup_{\mathbb{P}\in\mathcal{P}} \tau + \mathbb{E}_{\mathbb{P}}[g(\ell(Z) - \tau)]. \end{aligned}$$

The inner maximization problem in the resulting upper bound constitutes a worst-case expectation problem. Hence it is bounded above by the dual problem derived in Theorem 4.5. Substituting this dual problem into the above expression yields (5.10). Strong duality follows from the minimax theorem for optimized certainty equivalents (Theorem 5.18) and the strong duality result for worst-case expectation problems (Theorem 4.5), which apply under the given assumptions. □

The semi-infinite constraint in (5.10) involves the composite function $g(\ell(z)-\tau)$, which fails to be concave in z even if g is non-decreasing and ℓ is concave. Thus, checking whether a given $(\tau, \lambda_0, \lambda)$ satisfies the semi-infinite constraint in (5.10) is generically hard. In fact Chen and Sim (2024, Theorem 1) prove that evaluating the worst-case entropic risk is NP-hard even if ℓ is linear and $\mathcal{P}$ is a Markov ambiguity set. Hence, while providing theoretical insights, Theorem 5.18 does not necessarily pave the way towards an efficient method for solving worst-case risk problems of the form (5.4). Nevertheless, Theorem 5.18 provides a concise reformulation for (5.4) that is susceptible to approximate iterative solution procedures.

5.4. *ϕ-divergence ambiguity sets*

Recall that the ϕ-divergence ambiguity set (2.10) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) : \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq r\},$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where $\mathcal{Z}$ is a closed support set, $r \geq 0$ is a size parameter, ϕ is an entropy function in the sense of Definition 2.4, D_ϕ is the corresponding ϕ-divergence in the sense of Definition 2.5, and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. The next theorem establishes a duality result for worst-case risk problems over ϕ-divergence ambiguity sets. The proof follows from Theorems 4.14 and 5.18 and is thus omitted.

Theorem 5.20 (Duality theory for ϕ-divergence ambiguity sets II). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. If $\mathcal{P}$ is the ϕ-divergence ambiguity set (2.10), and ϱ is an optimized certainty equivalent induced by a disutility function g, then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf\limits_{\tau,\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} & \tau + \lambda_0 + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(g(\ell(Z)-\tau)-\lambda_0,\lambda)] \\ \text{s.t.} & \lambda_0 + \lambda\,\phi^\infty(1) \geq \sup\limits_{z\in\mathcal{Z}} g(\ell(z)-\tau). \end{cases}$$

If $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[g(\ell(Z))] < \infty$, $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}$, $r > 0$ and ϕ is continuous at 1, then strong duality holds, that is, the inequality becomes an equality.

A duality result akin to Theorem 5.20 also holds for worst-case risk problems over *restricted* ϕ-divergence ambiguity sets of the form

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathbb{P} \ll \hat{\mathbb{P}},\ \mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}}) \leq r\}.$$

The proof of the next theorem follows immediately from Theorems 4.15 and 5.18.

Theorem 5.21 (Duality theory for restricted ϕ-divergence ambiguity sets II). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$. If $\mathcal{P}$ is the restricted ϕ-divergence ambiguity set (2.11), and ϱ is an optimized certainty equivalent induced by a disutility function g, then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] \leq \inf_{\tau,\lambda_0\in\mathbb{R},\,\lambda\in\mathbb{R}_+} \tau + \lambda_0 + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(g(\ell(Z)-\tau)-\lambda_0,\lambda)].$$

If $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[g(\ell(Z))] < \infty$, $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}$, $r > 0$ and ϕ is continuous at 1, then strong duality holds, that is, the inequality becomes an equality.

5.5. Optimal transport ambiguity sets

Recall that the optimal transport ambiguity set (2.27) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathrm{OT}_c(\mathbb{P},\hat{\mathbb{P}}) \leq r\},$$

where $\mathcal{Z}$ is a closed support set, $r \geq 0$ is a size parameter, c is a transportation cost function in the sense of Definition 2.14, OT_c is the corresponding optimal transport discrepancy in the sense of Definition 2.15, and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. The next theorem establishes a duality result for worst-case risk problems over optimal transport ambiguity sets. Its proof follows immediately from Theorems 4.18 and 5.18 and is thus omitted.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Theorem 5.22 (Duality theory for optimal transport ambiguity sets II). Assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$ and ℓ is upper semicontinuous. If $\mathcal{P}$ is the optimal transport ambiguity set defined in (2.27) and ϱ is an optimized certainty equivalent induced by a disutility function g, then the following weak duality relation holds:

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] \leq \inf_{\tau\in\mathbb{R},\, \lambda\in\mathbb{R}_+} \tau + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} g(\ell(z)-\tau) - \lambda c(z,\hat{Z})\right].$$

If $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[g(\ell(Z))] < \infty$, $\mathbb{E}_{\mathbb{P}}[\ell(Z)] > -\infty$ for all $\mathbb{P} \in \mathcal{P}$ and $r > 0$, then strong duality holds, that is, the inequality becomes an equality.

Worst-case risk problems with optimal transport ambiguity sets are studied by Pflug and Wozabal (2007), Pichler (2013) and Wozabal (2014) in the context of portfolio selection with linear loss functions and by Mohajerin Esfahani *et al.* (2018) in the context of inverse optimization using the CVaR. Sadana, Delage and Georghiou (2024) investigate worst-case entropic risk measures over ∞-Wasserstein balls and establish tractable reformulations under standard convexity assumptions. Kent, Li, Blanchet and Glynn (2021) and Sheriff and Mohajerin Esfahani (2024) develop customized Frank–Wolfe algorithms in the space of probability distribution to address worst-case risk problems involving generic loss functions and risk measures. Specifically, Kent *et al.* (2021) work with Wasserstein gradient flows and use the corresponding notions of smoothness to establish the convergence of their Frank–Wolfe algorithm. In contrast, Sheriff and Mohajerin Esfahani (2024) work with Gâteaux derivatives, which leads to a different notion of smoothness and thus to a different convergence analysis. Both algorithms display sublinear convergence rates. When the reference distribution $\hat{\mathbb{P}}$ is discrete or when only samples from $\hat{\mathbb{P}}$ are used, the algorithms' iterates represent discrete distributions with progressively increasing bit sizes. Theorem 5.22 provides a compact, albeit potentially non-convex, reformulation of the worst-case risk problem. This reformulation is amenable to primal–dual gradient methods in the finite-dimensional space of the dual variables, which are guaranteed to converge to a stationary point.

Worst-case risk problems represent special instances of optimization problems over spaces of probability distributions. The mainstream methods to address such problems leverage the machinery of Wasserstein gradient flows (Ambrosio, Gigli and Savaré 2008). Wasserstein gradient flows have recently been used in the context of distributionally robust optimization problems (Lanzetti, Bolognani and Dörfler 2022, Lanzetti, Terpin and Dörfler 2024, Xu, Lee, Cheng and Xie 2024), non-convex optimization (Chizat and Bach 2018, Chizat 2022) and variational inference (Jiang, Chewi and Pooladian 2024, Lambert *et al.* 2022, Diao, Balasubramanian, Chewi and Salim 2023, Zhang and Zhou 2020). The results of this section are new and complementary to these existing works.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

6. Analytical solutions of nature's subproblem

A key challenge in DRO is to handle the worst-case expectation problem embedded in (1.2). This problem is solved by the fictitious adversary – commonly thought of as *nature* – once the decision-maker has committed to an $x \in \mathcal{X}$. It maximizes a linear function over a convex subset of an infinite-dimensional space of measures and thus appears to be intractable. Therefore considerable research effort has been devoted to identifying conditions under which this problem is efficiently solvable. We now show that it can actually be solved *analytically* in interesting situations.

The duality theory derived in Section 4 motivates the following simple strategy for finding analytical solutions of nature's subproblem. Construct feasible solutions for the primal worst-case expectation problem and its dual, and show that their objective function values match. If such matching solutions can be found, then both of them must be optimal in their respective optimization problems thanks to weak duality. As we will see below, this simple strategy succeeds surprisingly often. In addition, we will see that analytical solutions for worst-case *expectation* problems can sometimes be generalized to analytical solutions for worst-case *risk* problems of the form (5.4). The material reviewed in this section covers several decades of research in DRO from the 1950s until the present day.

6.1. Jensen bound

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathbb{E}_{\mathbb{P}}[Z] = \mu\}, \tag{6.1a}$$

which maximizes the expected value of $\ell(Z)$ over the Markov ambiguity set of all distributions supported on $\mathcal{Z}$ with mean μ. The Markov ambiguity set is a moment ambiguity set of the form (2.1) with $f(z) = z$ and $\mathcal{F} = \{\mu\}$. By Theorem 4.5 and as the support function of $\mathcal{F}$ is linear, the problem dual to (6.1a) is given by

$$\inf_{\lambda_0\in\mathbb{R},\, \lambda\in\mathbb{R}^d} \{\lambda_0 + \lambda^\top \mu : \lambda_0 + \lambda^\top z \geq \ell(z) \;\; \forall z \in \mathcal{Z}\}. \tag{6.1b}$$

Intuitively, this dual problem aims to find an affine function $a(z) = \lambda_0 + \lambda^\top z$ that majorizes the loss function $\ell(z)$ on $\mathcal{Z}$ and has minimal expected value $\mathbb{E}_{\mathbb{P}}[a(Z)]$ under any distribution $\mathbb{P}$ feasible in the primal problem (6.1a).

Proposition 6.1 (Jensen bound). Suppose that $\mathcal{Z}$ is convex, $\mu \in \mathcal{Z}$, ℓ is concave, and $\lambda^\star$ is any supergradient of ℓ at μ. Then the primal problem (6.1a) is solved by $\mathbb{P}^\star = \delta_\mu$, and the dual problem (6.1b) is solved by $(\lambda_0^\star, \lambda^\star)$, where $\lambda_0^\star = \ell(\mu) - \mu^\top \lambda^\star$. In addition, the optimal values of (6.1a) and (6.1b) both equal $\ell(\mu)$.

Proof. By construction, $\mathbb{P}^\star$ is feasible in the primal worst-case expectation problem, and its objective function value amounts to $\ell(\mu)$. In addition, $(\lambda_0^\star, \lambda^\star)$ is feasible in the dual robust optimization problem because $\lambda^\star$ is a supergradient of ℓ at μ, and its objective function value amounts to $\ell(\mu)$, too. Hence, by weak duality

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

as established in Theorem 4.5, $\mathbb{P}^\star$ is primal optimal, and $(\lambda_0^\star, \lambda^\star)$ is dual optimal. □

Proposition 6.1 implies Jensen's inequality $\mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] = \ell(\mathbb{E}_{\mathbb{P}}[Z])$, which holds for all distributions $\mathbb{P}$ feasible in (6.1a) (Jensen 1906). Proposition 6.1 further shows that (6.1b) is solved by any affine function tangent to ℓ at μ.

If the loss function $\ell(x, z)$ in the DRO problem (1.2) is concave in z for any fixed $x \in \mathcal{X}$, then Proposition 6.1 implies that the *same* distribution $\mathbb{P}^\star$ solves the inner maximization problem in (1.2) for every $x \in \mathcal{X}$. Hence the DRO problem (1.2) reduces to the (non-robust) stochastic program $\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[\ell(x, Z)]$.

Jensen's inequality is traditionally used to approximate hard *stochastic* optimization problems of the form $\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)]$, where $\mathbb{P}$ is a known continuous distribution of Z. Proposition 6.1 implies that if $\ell(x, z)$ is concave in z for any $x \in \mathcal{X}$, then replacing $\mathbb{P}$ with $\mathbb{P}^\star = \delta_{\mathbb{E}_{\mathbb{P}}[Z]}$ leads to a conservative approximation of this stochastic program. As $\mathbb{P}^\star$ is discrete (in fact, a Dirac distribution), the resulting approximate problem is much easier to solve. Its approximation quality can be improved by partitioning $\mathcal{Z}$ into finitely many convex cells and constructing separate Jensen bounds for all cells (Birge and Louveaux 2011, Section 10.1).

6.2. *Edmundson–Madansky bound*

The worst-case expectation problem (6.1a) over a Markov ambiguity set and its dual (6.1b) can also be solved in closed form if ℓ is convex and $\mathcal{Z}$ is a simplex.

Proposition 6.2 (Edmundson–Madansky bound). Suppose that $\mathcal{Z}$ is the probability simplex in $\mathbb{R}^d$ with vertices e_i, $i \in [d]$, $\mu \in \operatorname{rint}(\mathcal{Z})$, and ℓ is convex and real-valued. Then the primal problem (6.1a) is solved by $\mathbb{P}^\star = \sum_{i=1}^d \mu_i \delta_{e_i}$, and the dual problem (6.1b) is solved by $(\lambda_0^\star, \lambda^\star)$, where $\lambda_0^\star = 0$ and $\lambda_i^\star = \ell(e_i)$ for all $i \in [d]$. In addition, the optimal values of (6.1a) and (6.1b) both equal $\sum_{i=1}^d \mathbb{E}_{\mathbb{P}}[Z_i]\ell(e_i)$.

Proof. As μ belongs to the probability simplex, $\mathbb{P}^\star$ is feasible in the primal worst-case expectation problem with objective function value $\sum_{i=1}^d \mu_i \ell(e_i)$. Also, as ℓ is convex, Jensen's inequality implies that

$$\lambda_0^\star + z^\top \lambda^\star = \sum_{i=1}^d z_i \ell(e_i) \geq \ell\left(\sum_{i=1}^d z_i e_i\right) = \ell(z) \quad \text{for all } z \in \mathcal{Z}.$$

We conclude that $(\lambda_0^\star, \lambda^\star)$ is feasible in the dual robust optimization problem, and its objective function value amounts to $\sum_{i=1}^d \mu_i \ell(e_i)$, too. Hence, by weak duality as established in Theorem 4.5, $\mathbb{P}^\star$ is primal optimal, and $(\lambda_0^\star, \lambda^\star)$ is dual optimal. □

Proposition 6.2 implies the Edmundson–Madansky inequality, which states that $\mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] = \sum_{i=1}^d \mathbb{E}_{\mathbb{P}}[Z_i]\ell(e_i)$ for all distributions $\mathbb{P}$ feasible in (6.1a) (Edmundson 1956, Madansky 1959), and it shows that (6.1b) is solved by an affine

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

function that touches ℓ at the vertices e_i, $i \in [d]$, of $\mathcal{Z}$. We emphasize, however, that Proposition 6.2 remains valid with minor modifications if $\mathcal{Z}$ is an arbitrary regular simplex in $\mathbb{R}^d$, that is, the convex hull of $d+1$ affinely independent vectors $v_i \in \mathbb{R}^d$, $i \in [d+1]$; see Birge and Wets (1986) and Gassmann and Ziemba (1986).

If the loss function $\ell(x, z)$ in (1.2) is convex in z for any fixed $x \in \mathcal{X}$, then Proposition 6.2 implies that the DRO problem (1.2) is equivalent to the stochastic program $\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[\ell(x, Z)]$, where $\mathbb{P}^\star$ is *independent* of x. As $\mathbb{P}^\star$ is a discrete distribution with d atoms, this stochastic program is usually easy to solve.

6.3. *Barycentric approximation*

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{V}\times\mathcal{W})} \big\{\mathbb{E}_{\mathbb{P}}[\ell(V, W)] : \mathbb{E}_{\mathbb{P}}[V] = \bar{v},\ \mathbb{E}_{\mathbb{P}}[W] = \bar{w},\ \mathbb{E}_{\mathbb{P}}[VW^\top] = C\big\}, \tag{6.2a}$$

which maximizes the expected value of $\ell(V, W)$ across all distributions of $Z = (V, W)$ on $\mathcal{V} \times \mathcal{W}$ under which V and W have mean vectors $\bar{v}$ and $\bar{w}$, respectively, and cross-moment matrix C. Note that if V and W are uncorrelated, then $C = \bar{v}\bar{w}^\top$. Problem (6.2a) optimizes over a moment ambiguity set of the form (2.1) with $f(v, w) = (v, w, vw^\top)$ and $\mathcal{F} = \{\bar{v}\} \times \{\bar{w}\} \times \{C\}$. By Theorem 4.5 and as the support function of $\mathcal{F}$ is linear, the problem dual to (6.2a) is given by

$$\begin{array}{ll} \inf & \lambda_0 + \lambda_v^\top \bar{v} + \lambda_w^\top \bar{w} + \langle \Lambda, C\rangle \\ \text{s.t.} & \lambda_0 \in \mathbb{R},\ \lambda_v \in \mathbb{R}^{d_v},\ \lambda_w \in \mathbb{R}^{d_w},\ \Lambda \in \mathbb{R}^{d_v\times d_w} \\ & \lambda_0 + \lambda_v^\top v + \lambda_w^\top w + v^\top \Lambda w \geq \ell(v, w)\ \ \forall v \in \mathcal{V},\ \forall w \in \mathcal{W}. \end{array} \tag{6.2b}$$

This dual problem seeks a bi-affine function $b(v, w) = \lambda_0 + \lambda_v^\top v + \lambda_w^\top w + v^\top \Lambda w$ that majorizes the loss function $\ell(v, w)$ on $\mathcal{V} \times \mathcal{W}$ and minimizes $\mathbb{E}_{\mathbb{P}}[b(V, W)]$ under any distribution $\mathbb{P}$ feasible in (6.2a). The following proposition shows that problems (6.2a) and (6.2b) can be solved in closed form if ℓ is a concave–convex saddle function and $\mathcal{W}$ is a simplex. Below, we use e_i to denote the ith standard basis vector in $\mathbb{R}^{d_w}$, $i \in [d_w]$, and e to denote the vector of ones in $\mathbb{R}^{d_w}$.

Proposition 6.3 (Barycentric approximation). Suppose that $\mathcal{V} \subseteq \mathbb{R}^{d_v}$ is convex and $\mathcal{W} \subseteq \mathbb{R}^{d_w}$ is the probability simplex with vertices e_i, $i \in [d_w]$. Suppose also that the loss function $\ell(v, w)$ is concave and superdifferentiable in v for any fixed w and convex in w for any fixed v. In addition, suppose that $\bar{v} \in \mathcal{V}$, $\bar{w} \in \operatorname{rint}(\mathcal{W})$ and $Ce = \bar{v}$ and that problem (6.2a) is feasible. Then (6.2a) is solved by

$$\mathbb{P}^\star = \sum_{i=1}^{d_w} \bar{w}_i\, \delta_{(Ce_i/\bar{w}_i, e_i)}.$$

If $\Lambda_i^\star$ is any supergradient in $\partial_v \ell(Ce_i/\bar{w}_i, e_i)$ for all $i \in [d_w]$ and

$$\lambda_{w,i}^\star = \ell(Ce_i/\bar{w}_i, e_i) - (\Lambda_i^\star)^\top Ce_i/\bar{w}_i \quad \text{for all } i \in [d_w],$$

then the dual problem (6.2b) is solved by $(\lambda_0^\star, \lambda_v^\star, \lambda_w^\star, \Lambda^\star)$, where $\lambda_0^\star = 0$ and

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$\lambda_v^\star = 0$, while $\lambda_w^\star$ has elements $\lambda_{w,i}^\star$ and $\Lambda^\star$ has columns $\Lambda_i^\star$, $i \in [d_w]$. The optimal values of (6.2a) and (6.2b) coincide and are both equal to

$$\sum_{i=1}^{d_w} \mu_{w,i}\, \ell(Ce_i/\bar{w}_i, e_i).$$

The condition $Ce = \bar{v}$ is necessary for (6.2a) to be feasible. Indeed, if $\mathbb{P}$ is feasible in (6.2a), then we have $Ce = \mathbb{E}_{\mathbb{P}}[VW^\top e] = \mathbb{E}_{\mathbb{P}}[V] = \bar{v}$. Here the second equality holds because $\mathbb{P}(W \in \mathcal{W}) = 1$ and $\mathcal{W}$ is the probability simplex in $\mathbb{R}^{d_w}$. However, the condition $Ce = \bar{v}$ is *not* sufficient for (6.2a) to be feasible. Indeed, if the support set $\mathcal{V} = \{\bar{v}\}$ is a singleton, then $C = \mathbb{E}_{\mathbb{P}}[VW^\top] = \bar{v}\bar{w}^\top$. That is, V and W must be uncorrelated. Hence $\mathcal{V}$ and C cannot be selected independently. To circumvent this problem, Proposition 6.3 requires (6.2a) to be feasible.

Proof of Proposition 6.3. Note that $\bar{w} > 0$ and $\mathrm{e}^\top \bar{w} = 1$ because $\bar{w}$ belongs to the relative interior of the probability simplex $\mathcal{W}$. Thus $\mathbb{P}^\star$ is indeed a well-defined probability distribution, that is, the atoms of $\mathbb{P}^\star$ have positive probabilities that sum to 1. In addition, $\mathbb{P}^\star$ is supported on $\mathcal{V} \times \mathcal{W}$ because

$$Ce_i/\bar{w}_i = \mathbb{E}_{\mathbb{P}}\left[\frac{VW_i}{\mathbb{E}_{\mathbb{P}}[W_i]}\right] = \mathbb{E}_{\mathbb{P}}\left[V\,\frac{\mathbb{E}_{\mathbb{P}}[W_i|V]}{\mathbb{E}_{\mathbb{P}}[W_i]}\right] \in \mathcal{V} \quad \text{and} \quad e_i \in \mathcal{W}$$

for all $i \in [d_w]$, where $\mathbb{P}$ is any distribution feasible in (6.2a). Note also that if V and W are uncorrelated, in which case $C = \bar{v}\bar{w}^\top$, then the ith generalized barycentre $Ce_i/\bar{w}_i$ of $\mathcal{V}$ simplifies to $\bar{v}$ for every $i \in [d_w]$. Recalling that $Ce = \bar{v}$, we further have

$$\mathbb{E}_{\mathbb{P}^\star}[V] = \sum_{i=1}^{d_w} \bar{w}_i\, Ce_i/\bar{w}_i = \bar{v}, \quad \mathbb{E}_{\mathbb{P}^\star}[W] = \sum_{i=1}^{d_w} \bar{w}_i\, e_i = \bar{w}$$

and

$$\mathbb{E}_{\mathbb{P}^\star}[VW^\top] = \sum_{i=1}^{d_w} \bar{w}_i\, Ce_i e_i^\top/\bar{w}_i = C.$$

In summary, we have shown that $\mathbb{P}^\star$ is feasible in (6.2a). A similar calculation reveals that the objective function value of $\mathbb{P}^\star$ in (6.2a) is given by the formula in the proposition statement. Details are omitted for brevity.

To show that $(\lambda_0^\star, \lambda_v^\star, \lambda_w^\star, \Lambda^\star)$ is feasible in (6.2b), note first that

$$\begin{aligned}
\lambda_0^\star + (\lambda_v^\star)^\top v + (\lambda_w^\star)^\top w + v^\top \Lambda^\star w &= \sum_{i=1}^{d_w} w_i [\ell(Ce_i/\bar{w}_i, e_i) + (\Lambda_i^\star)^\top (v - Ce_i/\bar{w}_i)] \\
&\geq \sum_{i=1}^{d_w} w_i\, \ell(v, e_i) \\
&\geq \ell(v, w)
\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$. The first inequality follows from the concavity of $\ell(v, w)$ in v and the definition of $\Lambda_i^\star$ as a supergradient, while the second inequality follows from the convexity of $\ell(v, w)$ in w and Jensen's inequality. Hence $(\lambda_0^\star, \lambda_v^\star, \lambda_w^\star, \Lambda^\star)$ is indeed feasible in (6.2b). A similar calculation reveals that the objective function value of $(\lambda_0^\star, \lambda_v^\star, \lambda_w^\star, \Lambda^\star)$ in (6.2b) is given by the formula in the proposition statement. Consequently, by weak duality as established in Theorem 4.5, we have shown that $\mathbb{P}^\star$ is primal optimal and $(\lambda_0^\star, \lambda_v^\star, \lambda_w^\star, \Lambda^\star)$ is dual optimal. □

Proposition 6.3 remains valid with obvious minor modifications if $\mathcal{W}$ is defined as an arbitrary regular simplex in $\mathbb{R}^{d_w}$ (Frauendorfer 1992). If $z = (v, w)$ and the loss function $\ell(x, z) = \ell(x, v, w)$ in (1.2) is concave in v and convex in w for any fixed $x \in \mathcal{X}$, then Proposition 6.3 implies that the DRO problem (1.2) is equivalent to the stochastic program $\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[\ell(x, U, V)]$, where $\mathbb{P}^\star$ is *independent* of x. As $\mathbb{P}^\star$ is a discrete distribution with d_w atoms, this stochastic program is usually easy to solve. Traditionally, the distribution $\mathbb{P}^\star$ is used to approximate hard *stochastic* optimization problems of the form $\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, V, W)]$, where $\mathbb{P}$ is a *known* continuous distribution of (V, W). Proposition 6.3 implies that if $\ell(x, v, w)$ is concave in v and convex in w for any $x \in \mathcal{X}$, then replacing $\mathbb{P}$ with $\mathbb{P}^\star$ leads to a conservative approximation, which is termed the *upper barycentric approximation* of the original stochastic program (Frauendorfer 1992). Barycentric approximations for more general stochastic programs involving loss functions that may fail to be convex and/or concave are derived by Kuhn (2005).

6.4. Ben-Tal and Hochman bound

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathbb{E}_{\mathbb{P}}[Z] = \mu,\ \mathbb{E}_{\mathbb{P}}[|Z - \mu|] = \sigma\}, \tag{6.3a}$$

which maximizes the expected value of $\ell(Z)$ over the family of all univariate distributions supported on $\mathcal{Z}$ with mean μ and mean absolute deviation σ. Note that problem (6.3a) optimizes over a moment ambiguity set of the form (2.1) with $f(z) = (z, |z - \mu|)$ and $\mathcal{F} = \{\mu\} \times \{\sigma\}$. By Theorem 4.5 and as the support function of $\mathcal{F}$ is linear, the problem dual to (6.3a) is given by

$$\inf_{\lambda_0,\lambda_1,\lambda_2\in\mathbb{R}} \{\lambda_0 + \lambda_1\mu + \lambda_2\sigma : \lambda_0 + \lambda_1 z + \lambda_2|z - \mu| \geq \ell(z)\ \ \forall z \in \mathcal{Z}\}. \tag{6.3b}$$

Intuitively, this dual problem aims to approximate the loss function from above with a piecewise linear continuous function that has a kink at μ. The problems (6.3a) and (6.3b) can be solved in closed form if ℓ is convex.

Proposition 6.4 (Ben-Tal and Hochman bound). Assume that $\mathcal{Z} = [0, 1]$, $\mu \in (0, 1)$ and $\sigma \in [0, 2\mu(1 - \mu)]$. Suppose also that ℓ is a real-valued convex function. Then the primal problem (6.3a) is solved by

$$\mathbb{P}^\star = \frac{\sigma}{2\mu}\,\delta_0 + \left(1 - \frac{\sigma}{2\mu} - \frac{\sigma}{2(1-\mu)}\right)\delta_\mu + \frac{\sigma}{2(1-\mu)}\,\delta_1,$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

and the dual problem (6.3b) is solved by

$$\lambda_0^\star = \frac{(1-\mu)\ell(0) + \ell(\mu) - \mu\ell(1)}{2(1-\mu)},$$
$$\lambda_1^\star = \frac{(\mu-1)\ell(0) + (1-2\mu)\ell(\mu) + \mu\ell(1)}{2\mu(1-\mu)},$$
$$\lambda_2^\star = \frac{(1-\mu)\ell(0) - \ell(\mu) + \mu\ell(1)}{2\mu(1-\mu)}.$$

In addition, the optimal values of (6.3a) and (6.3b) coincide and are both equal to

$$\frac{\sigma}{2\mu}\,\ell(0) + \left(1 - \frac{\sigma}{2\mu} - \frac{\sigma}{2(1-\mu)}\right)\ell(\mu) + \frac{\sigma}{2(1-\mu)}\,\ell(1).$$

Proof. The assumptions about μ and σ imply that $\mathbb{P}^\star$ is supported on $\mathcal{Z}$ and that the probabilities of the three atoms are non-negative and sum to 1. Also, we have

$$\mathbb{E}_{\mathbb{P}^\star}[Z] = \left(\mu - \frac{\sigma}{2} - \frac{\sigma\mu}{2(1-\mu)}\right) + \frac{\sigma}{2(1-\mu)} = \mu \quad \text{and} \quad \mathbb{E}_{\mathbb{P}^\star}[|Z-\mu|] = \sigma.$$

Thus $\mathbb{P}^\star$ is feasible in (6.3a). In addition, one readily verifies that the objective function value of $\mathbb{P}^\star$ in (6.3a) is given by the formula in the proposition statement.

Next, note that the piecewise linear function $\lambda_0^\star + \lambda_1^\star z + \lambda_2^\star |z-\mu|$ coincides with the loss function $\ell(z)$ for every $z \in \{0, \mu, 1\}$. As the loss function is convex, we may thus conclude that $\lambda_0^\star + \lambda_1^\star z + \lambda_2^\star |z-\mu|$ majorizes $\ell(z)$ for every $z \in [0,1] = \mathcal{Z}$. This shows that $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ is feasible in (6.3b). An elementary calculation further reveals that the objective function value of $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ in (6.3b) is given by the formula in the proposition statement. Weak duality as established in Theorem 4.5 thus implies that $\mathbb{P}^\star$ is primal optimal and that $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ is dual optimal. □

Proposition 6.4 readily extends to support sets of the form $\mathcal{Z} = [a, b]$ for any $a, b \in \mathbb{R}$ with $a < \mu < b$ by applying a linear coordinate transformation. If $\ell(x, z)$ in (1.2) is convex in z for any fixed $x \in \mathcal{X}$, then Proposition 6.4 implies that the DRO problem (1.2) is equivalent to the stochastic program $\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[\ell(x, Z)]$, where the three-point distribution $\mathbb{P}^\star$ is *independent* of x. Traditionally, this stochastic program is used as a conservative approximation for a stochastic program of the form $\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)]$, where $\mathbb{P}$ is a known continuous distribution (Ben-Tal and Hochman 1972). Unlike the Jensen and Edmundson–Madansky bounds, which only use information about the location of $\mathbb{P}$, and unlike the barycentric approximation, which only uses information about the location and certain cross-moments of $\mathbb{P}$, the Ben-Tal and Hochman bound uses information about the location as well as the dispersion of $\mathbb{P}$. Thus it provides a tighter approximation.

If Z is a d-dimensional random vector with *independent* components Z_i, $i \in [d]$, each of which has a known mean and mean absolute deviation, then one can show that the worst-case expected value of a convex loss function is attained by $\mathbb{P}^\star = \otimes_{i=1}^d \mathbb{P}_i^\star$, where each $\mathbb{P}_i^\star$ is a three-point distribution constructed as in Proposition 6.4

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

(Ben-Tal and Hochman 1972). In this case, $\mathbb{P}^\star$ is a discrete distribution with 3^d atoms. Hence, evaluating expected values with respect to $\mathbb{P}^\star$ is generically hard but becomes tractable for a class of exponential loss functions that offer safe approximations for chance constraints (Postek *et al.* 2018).

6.5. *Scarf's bound*

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathbb{E}_{\mathbb{P}}[Z] = 0,\ \mathbb{E}_{\mathbb{P}}[Z^2] = \sigma^2\}, \tag{6.4a}$$

which maximizes the expected value of $\ell(Z)$ over the Chebyshev ambiguity set of all univariate distributions supported on $\mathcal{Z}$ with mean 0 and variance σ^2. This Chebyshev ambiguity set is a moment ambiguity set of the form (2.1) with $f(z) = (z, z^2)$ and $\mathcal{F} = \{0\} \times \{\sigma^2\}$. By Theorem 4.5 and as the support function of $\mathcal{F}$ is linear, the problem dual to (6.4a) is given by

$$\inf_{\lambda_0,\lambda_1,\lambda_2\in\mathbb{R}} \{\lambda_0 + \lambda_2\sigma^2 : \lambda_0 + \lambda_1 z + \lambda_2(z-\mu)^2 \geq \ell(z)\ \ \forall z \in \mathcal{Z}\}. \tag{6.4b}$$

This dual problem seeks a quadratic function $q(z) = \lambda_0 + \lambda_1 z + \lambda_2 z^2$ that majorizes the loss function $\ell(z)$ throughout $\mathcal{Z}$ and has minimal expectation $\mathbb{E}_{\mathbb{P}}[q(Z)]$ under any distribution $\mathbb{P}$ with mean 0 and variance σ^2. The problems (6.4a) and (6.4b) can be solved in closed form if ℓ is a ramp function.

Proposition 6.5 (Scarf's bound). If $\mathcal{Z} = \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$ and $\ell(z) = \max\{z - a, 0\}$ is a ramp function with a kink at $a \in \mathbb{R}$, then the primal problem (6.4a) is solved by

$$\mathbb{P}^\star = \frac{1}{2}\left(1 + \frac{a}{\sqrt{a^2+\sigma^2}}\right)\delta_{a-\sqrt{a^2+\sigma^2}} + \frac{1}{2}\left(1 - \frac{a}{\sqrt{a^2+\sigma^2}}\right)\delta_{a+\sqrt{a^2+\sigma^2}},$$

and the dual problem (6.4b) is solved by

$$\lambda_0^\star = \frac{\left(a - \sqrt{a^2+\sigma^2}\right)^2}{4\sqrt{a^2+\sigma^2}}, \quad \lambda_1^\star = -\frac{a - \sqrt{a^2+\sigma^2}}{2\sqrt{a^2+\sigma^2}} \quad \text{and} \quad \lambda_2^\star = \frac{1}{4\sqrt{a^2+\sigma^2}}.$$

The optimal values of (6.4a) and (6.4b) are both equal to $\frac{1}{2}(\sqrt{a^2+\sigma^2} - a)$.

Proof. Note that the two-point distribution $\mathbb{P}^\star$ is well-defined, that is, its atoms have non-negative probabilities that sum to 1. By the definition of $\mathbb{P}^\star$, we also have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^\star}[Z] &= \frac{1}{2}\left(1 + \frac{a}{\sqrt{a^2+\sigma^2}}\right)\left(a - \sqrt{a^2+\sigma^2}\right) \\ &\quad + \frac{1}{2}\left(1 - \frac{a}{\sqrt{a^2+\sigma^2}}\right)\left(a + \sqrt{a^2+\sigma^2}\right) \\ &= 0.\end{aligned}$$

Similarly, it is easy to verify that $\mathbb{E}_{\mathbb{P}^\star}[Z^2] = \sigma^2$. This shows that $\mathbb{P}^\star$ is feasible in

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

(6.4a). The objective function value of $\mathbb{P}^\star$ is

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] = \mathbb{E}_{\mathbb{P}^\star}[\max\{Z - a, 0\}] = \frac{1}{2}\big(\sqrt{a^2 + \sigma^2} - a\big).$$

Next, observe that the dual variables $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ defined in the proposition statement give rise to the quadratic function

$$q^\star(z) = \lambda_0^\star + \lambda_1^\star z + \lambda_2^\star z^2 = \frac{1}{4\sqrt{a^2 + \sigma^2}}\big(z - a + \sqrt{a^2 + \sigma^2}\big)^2.$$

We will now show that $q^\star(z) \geq \max\{z - a, 0\} = \ell(z)$ for all $z \in \mathcal{Z}$. Clearly, $q^\star$ is non-negative and evaluates to 0 at $a - \sqrt{a^2 + \sigma^2}$. In addition, $q^\star$ touches the affine function $z - a$ at $a + \sqrt{a^2 + \sigma^2}$. To see this, note that

$$q^\star\big(a + \sqrt{a^2 + \sigma^2}\big) = \sqrt{a^2 + \sigma^2} \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}z} q^\star\big(a + \sqrt{a^2 + \sigma^2}\big) = 1.$$

Hence $q^\star$ majorizes the ramp function $\ell(z)$, implying that $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ is dual feasible. Also, the objective function value of $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ is given by

$$\lambda_0^\star + \lambda_2^\star \sigma^2 = \frac{1}{4\sqrt{a^2 + \sigma^2}}\Big(\sigma^2 + \big(a - \sqrt{a^2 + \sigma^2}\big)^2\Big) = \frac{1}{2}\big(\sqrt{a^2 + \sigma^2} - a\big).$$

As the objective function values of $\mathbb{P}^\star$ and $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ match, weak duality as established in Theorem 4.5 thus implies that $\mathbb{P}^\star$ is primal optimal and that $(\lambda_0^\star, \lambda_1^\star, \lambda_2^\star)$ is dual optimal. This observation completes the proof. □

Proposition 6.5 was first derived by Scarf (1958) in his pioneering treatise on the distributionally robust newsvendor problem; see also Jagannathan (1977, Theorem 1). Note that if the mean of Z is known to equal $\mu \neq 0$ instead of 0, then Scarf's bound remains valid if we replace a with $a - \mu$. Gallego and Moon (1993) extend Scarf's bound to more general loss functions such as wedge functions or ramp functions with a discontinuity, whereas Natarajan *et al.* (2018) extend Scarf's bound to more general ambiguity sets that contain information not only about the mean and variance of Z but also about its *semi*variance. In addition, Das, Dhara and Natarajan (2021) discuss variants of Scarf's bound that rely on information about the mean and the αth moment of Z for any $\alpha > 1$.

Proposition 6.5 is often used to reformulate DRO problems of the form (1.2) whose objective function is given by the expected value of a ramp function. Examples include distributionally robust newsvendor, support vector machine or mean–CVaR portfolio selection problems. In most of these applications, the location a of the kind of the ramp function is a decision variable or a function of the decision variables. Thus the worst-case distribution $\mathbb{P}^\star$ is decision-dependent, which means that Proposition 6.5 does *not* enable us to reduce the DRO problem (1.2) to a stochastic program with a single fixed worst-case distribution.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

6.6. Marshall and Olkin bound

Consider the worst-case probability problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{P}(Z \in \mathcal{C}) \colon \mathbb{E}_{\mathbb{P}}[Z] = 0,\ \mathbb{E}_{\mathbb{P}}[ZZ^\top] = I_d\}, \tag{6.5a}$$

which maximizes the probability of the event $Z \in \mathcal{C}$ over the Chebyshev ambiguity set of all distributions on $\mathcal{Z} = \mathbb{R}^d$ with mean 0 and covariance matrix I_d. This Chebyshev ambiguity set is a moment ambiguity set of the form (2.1) with $f(z) = (z, zz^\top)$ and $\mathcal{F} = \{0\} \times \{I_d\}$. If we set ℓ to the characteristic function of $\mathcal{C}$ defined by $\ell(z) = \mathbb{1}_{z\in\mathcal{C}}$ for all $z \in \mathcal{Z}$, then the worst-case probability problem (6.5a) can be recast as a worst-case expectation problem. By Theorem 4.5 and as the support function of $\mathcal{F}$ is linear, the corresponding dual problem is thus given by

$$\inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}^d,\Lambda\in\mathbb{S}^d} \{\lambda_0 + \langle\Lambda, I_d\rangle \colon \lambda_0 + \lambda^\top z + z^\top\Lambda z \geq \ell(z)\ \ \forall z \in \mathcal{Z}\}. \tag{6.5b}$$

The problems (6.5a) and (6.5b) can be solved analytically if $\mathcal{C}$ is convex and closed.

Proposition 6.6 (Marshall and Olkin bound). Suppose that $\mathcal{Z} = \mathbb{R}^d$, $\mathcal{C} \subseteq \mathbb{R}^d$ is convex and closed, and ℓ is the characteristic function of $\mathcal{C}$. Set $\Delta = \min_{z\in\mathcal{C}} \|z\|_2$, and let $z_0 \in \mathbb{R}^d$ be the unique minimizer of this problem. Then the optimal values of (6.5a) and (6.5b) are both equal to $(1 + \Delta^2)^{-1}$. If $\Delta = 0$, then the supremum of (6.5a) may not be attained. However, if $\Delta > 0$, then (6.5a) is solved by

$$\mathbb{P}^\star = \frac{1}{1+\Delta^2}\,\delta_{z_0} + \frac{\Delta^2}{1+\Delta^2}\mathbb{Q},$$

where $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ is an arbitrary distribution with mean $-z_0/\Delta^2$ and covariance matrix

$$\frac{1+\Delta^2}{\Delta^2}\,(I_d - z_0 z_0^\top/\Delta^2).$$

For any $\Delta \geq 0$, problem (6.4b) is solved by

$$\lambda_0^\star = \frac{1}{(1+\Delta^2)^2}, \quad \lambda^\star = \frac{2z_0}{(1+\Delta^2)^2} \quad \text{and} \quad \Lambda^\star = \frac{z_0 z_0^\top}{(1+\Delta^2)^2}.$$

Proof. Assume first that $\Delta = 0$, that is, $0 \in \mathcal{C}$. For every $j \in \mathbb{N}$, let $\mathbb{Q}_j \in \mathcal{P}(\mathcal{Z})$ be any distribution with mean 0 and covariance matrix jI_d, and set

$$\mathbb{P}_j = (1 - 1/j)\,\delta_0 + (1/j)\,\mathbb{Q}_j.$$

We thus have $\mathbb{E}_{\mathbb{P}_j}[Z] = 0$ and $\mathbb{E}_{\mathbb{P}_j}[ZZ^\top] = I_d$, which implies that $\mathbb{P}_j$ is feasible in (6.5a). In addition, the objective function value of $\mathbb{P}_j$ in (6.5a) satisfies

$$\mathbb{P}_j(Z \in \mathcal{C}) = 1 - 1/j + \mathbb{Q}_j(Z \in \mathcal{Z})/j \geq 1 - j^{-1}.$$

Driving j to infinity reveals that problem (6.5a) is trivial for $\Delta = 0$ and that its supremum equals 1. Assume now that $\Delta > 0$, and let $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ be an arbitrary

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

distribution with mean $-z_0/\Delta^2$ and covariance matrix

$$\frac{1+\Delta^2}{\Delta^2}\,(I_d - z_0 z_0^\top/\Delta^2).$$

Such a distribution is guaranteed to exist because $I_d \succeq z_0 z_0^\top/\Delta^2$. In addition, define $\mathbb{P}^\star$ as in the proposition statement. By construction, we have $\mathbb{E}_{\mathbb{P}^\star}[Z] = 0$ and

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^\star}[ZZ^\top] &= \frac{z_0 z_0^\top}{1+\Delta^2} + \frac{\Delta^2}{1+\Delta^2}\,\mathbb{E}_{\mathbb{Q}}[ZZ^\top] \\ &= \frac{z_0 z_0^\top}{1+\Delta^2} + I_d - \frac{z_0 z_0^\top}{\Delta^2} + \frac{\Delta^2}{1+\Delta^2}\,\frac{z_0 z_0^\top}{\Delta^4} \\ &= I_d.\end{aligned}$$

Also, the objective function value of $\mathbb{P}^\star$ in (6.5a) is given by $\mathbb{P}^\star(Z \in \mathcal{C}) = (1+\Delta^2)^{-1}$. Next, use $(\lambda_0^\star, \lambda^\star, \Lambda^\star)$ defined in the proposition to construct the quadratic function

$$q^\star(z) = \lambda_0^\star + (\lambda^\star)^\top z + z^\top \Lambda^\star z = \frac{(z_0^\top z + 1)^2}{(1+\Delta^2)^2}.$$

Note that $q^\star$ is non-negative and constant on any hyperplane perpendicular to z_0. If $\Delta > 0$, we have $q^\star(z_0) = 1$ as well as $q^\star(-z_0/\Delta^2) = 0$. Thus, at every $z \in \mathcal{Z}$ with $z_0^\top z \geq -1$, the quadratic function $q^\star(z)$ is non-decreasing in the direction of z_0. As z_0 minimizes the differentiable convex function $\|z\|_2^2$ over the convex closed set $\mathcal{C}$, we have $z_0^\top(z - z_0) \geq 0$ for all $z \in \mathcal{C}$. By the monotonicity properties of $q^\star$, this implies that $q^\star(z) \geq 1$ for every $z \in \mathcal{C}$. Hence the quadratic function $q^\star$ majorizes the indicator function ℓ on $\mathcal{Z}$, which implies that $(\lambda_0^\star, \lambda^\star, \Lambda^\star)$ is dual feasible. If $\Delta = 0$, then $q^\star(z) = 1$ for all $z \in \mathcal{Z}$, and $(\lambda_0^\star, \lambda^\star, \Lambda^\star)$ is also dual feasible. In any case, one readily verifies that its objective function value is given by

$$\lambda_0^\star + \langle \Lambda^\star, I_d \rangle = (1+\Delta^2)^{-1}.$$

As the objective function values of $\mathbb{P}^\star$ and $(\lambda_0^\star, \lambda^\star, \Lambda^\star)$ for $\Delta > 0$ match, weak duality as established in Theorem 4.5 implies that $\mathbb{P}^\star$ is primal optimal and that $(\lambda_0^\star, \lambda^\star, \Lambda^\star)$ is dual optimal. If $\Delta = 0$, then the optimal value 1 of the primal problem also matches the objective function value of $(\lambda_0^\star, \lambda^\star, \Lambda^\star)$ in (6.5b). Hence $(\lambda_0^\star, \lambda^\star, \Lambda^\star)$ remains dual optimal even though the supremum of the primal problem may not be attained. This observation completes the proof. □

6.7. *Chebyshev risk*

Analytical solutions of worst-case *expectation* problems sometimes enable us to evaluate the worst-case *risk* of a random variable if the underlying risk measure is law-invariant, translation-invariant as well as scale-invariant; see Definition 5.3. For example, it is elementary to verify that the β-VaR and β-CVaR constitute law-invariant, translation-invariant as well as scale-invariant risk measures for every fixed $\beta \in (0, 1)$. If the distribution of Z is unknown except for its mean $\mu \in \mathbb{R}^d$ and

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

covariance matrix $\Sigma \in \mathbb{S}_+^d$, then it is natural to quantify the riskiness of an uncertain loss $\ell(Z)$ under a law-invariant risk measure ϱ by the corresponding *Chebyshev risk*. Specifically, the Chebyshev risk of $\ell(Z)$ is defined as the worst-case risk

$$\sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}[\ell(Z)],$$

where $\mathcal{P}(\mu,\Sigma)$ denotes the Chebyshev ambiguity set that contains all probability distributions on $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{S}_+^d$.

We now describe a powerful tool for analysing the Chebyshev risk with respect to any law-, translation- and scale-invariant risk measure. To this end, recall that if Z follows some distribution $\mathbb{P}$ on $\mathbb{R}^d$, then $L = \ell(Z)$ follows the pushforward distribution $\mathbb{P} \circ \ell^{-1}$ on $\mathbb{R}$. If $\mathbb{P}$ is uncertain and only known to belong to some ambiguity set $\mathcal{P}$, then the distribution of $L = \ell(Z)$ is also uncertain and only known to belong to the pushforward ambiguity set $\mathcal{P} \circ \ell^{-1} = \{\mathbb{P} \circ \ell^{-1} : \mathbb{P} \in \mathcal{P}\}$. The following proposition due to Popescu (2007) shows that linear pushforwards of Chebyshev ambiguity sets are again Chebyshev ambiguity sets.

Proposition 6.7 (Pushforwards of Chebyshev ambiguity sets). If $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{S}_+^d$, $\theta \in \mathbb{R}^d$, and $\ell : \mathbb{R}^d \to \mathbb{R}$ is the linear transformation defined by $\ell(z) = \theta^\top z$, then the pushforward of the Chebyshev ambiguity set $\mathcal{P}(\mu,\Sigma)$ is the Chebyshev ambiguity set of all distributions on $\mathbb{R}$ with mean $\theta^\top\mu$ and variance $\theta^\top\Sigma\theta$, that is,

$$\mathcal{P}(\mu,\Sigma) \circ \ell^{-1} = \mathcal{P}(\theta^\top\mu, \theta^\top\Sigma\theta).$$

Proof. First select any distribution $\mathbb{P} \in \mathcal{P}(\mu,\Sigma)$. If the random vector Z follows $\mathbb{P}$, then the random variable $L = \ell(Z)$ follows $\mathbb{P} \circ \ell^{-1}$. Thus we have

$$\mathbb{E}_{\mathbb{P}\circ\ell^{-1}}[L] = \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\mathbb{P}}[\theta^\top Z] = \theta^\top\mu,$$

where the first equality follows from the measure-theoretic change of variables formula. Similarly, one can show that $\mathbb{E}_{\mathbb{P}\circ\ell^{-1}}[(L-\theta^\top\mu)^2] = \theta^\top\Sigma\theta$. Thus we find

$$\mathcal{P}(\mu,\Sigma) \circ \ell^{-1} \subseteq \mathcal{P}(\theta^\top\mu, \theta^\top\Sigma\theta).$$

Next, select any $\mathbb{Q}_L \in \mathcal{P}(\theta^\top\mu, \theta^\top\Sigma\theta)$. If $\theta^\top\Sigma\theta = 0$, then $\mathbb{Q}_L = \delta_{\theta^\top\mu}$, which coincides with the pushforward distribution $\mathbb{P} \circ \ell^{-1}$ for any $\mathbb{P} \in \mathcal{P}(\mu,\Sigma)$. In the remainder of the proof we may thus assume that $\theta^\top\Sigma\theta \neq 0$. Let now L be a random variable governed by $\mathbb{Q}_L$, and let M be a d-dimensional random vector governed by an arbitrary distribution $\mathbb{Q}_M \in \mathcal{P}(\mathbb{R}^d)$ with mean μ and covariance matrix Σ. For example, we can set $\mathbb{Q}_M$ to the normal distribution $\mathcal{N}(\mu,\Sigma)$. Assume L and M are independent. Then the distribution $\mathbb{P}$ of the d-dimensional random vector

$$Z = \frac{1}{\theta^\top\Sigma\theta}\Sigma\theta\, L + \left(I_d - \frac{1}{\theta^\top\Sigma\theta}\Sigma\theta\theta^\top\right)M$$

belongs to $\mathcal{P}(\mu,\Sigma)$. By the construction of L and M, we indeed have

$$\mathbb{E}_{\mathbb{P}}[Z] = \frac{1}{\theta^\top\Sigma\theta}\Sigma\theta\,\mathbb{E}_{\mathbb{Q}_L}[L] + \left(I_d - \frac{1}{\theta^\top\Sigma\theta}\Sigma\theta\theta^\top\right)\mathbb{E}_{\mathbb{Q}_M}[M] = \mu$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

and

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[(Z-\mu)(Z-\mu)^\top] &= \frac{1}{\theta^\top\Sigma\theta}\Sigma\theta\theta^\top\Sigma + \left(I_d - \frac{1}{\theta^\top\Sigma\theta}\Sigma\theta\theta^\top\right)\Sigma\left(I_d - \frac{1}{\theta^\top\Sigma\theta}\theta\theta^\top\Sigma\right)\\ &= \Sigma.\end{aligned}$$

The first equality in the above expression holds because L and M are independent, L has variance $\theta^\top\Sigma\theta$ and M has covariance matrix Σ. By construction, we further have $\ell(Z) = \theta^\top Z = L$, which implies that $\mathbb{P} \circ \ell^{-1} = \mathbb{Q}_L$. We have thus shown that for every $\mathbb{Q}_L \in \mathcal{P}(\theta^\top\mu, \theta^\top\Sigma\theta)$ there exists $\mathbb{P} \in \mathcal{P}(\mu, \Sigma)$ with $\mathbb{P} \circ \ell^{-1} = \mathbb{Q}_L$, that is,

$$\mathcal{P}(\mu, \Sigma) \circ \ell^{-1} \supseteq \mathcal{P}(\theta^\top\mu, \theta^\top\Sigma\theta).$$

This observation completes the proof. □

Generalizations of Proposition 6.7 to multi-dimensional affine transformations and to subfamilies of the Chebyshev ambiguity set that contain only distributions with certain structural properties (such as symmetry, linear unimodality or log-concavity) are presented by Yu, Li, Schuurmans and Szepesvári (2009); see also Chen *et al.* (2011).

We now show that if the risk measure ϱ is law-, translation- and scale-invariant and the loss function ℓ is linear, then the Chebyshev risk reduces to a mean–standard deviation risk measure, which involves the standard risk coefficient of ϱ.

Definition 6.8 (Standard risk coefficient). The standard risk coefficient of a law-invariant risk measure ϱ is given by $\alpha = \sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \varrho_{\mathbb{Q}}[L]$.

Thus the standard risk coefficient of ϱ is defined as the worst-case risk of an uncertain loss L whose distribution $\mathbb{Q}$ is only known to have mean 0 and variance 1.

Proposition 6.9 (Chebyshev risk). If ϱ is a law-, translation- and scale-invariant risk measure with standard risk coefficient α, there is $\theta \in \mathbb{R}^d$ with $\ell(z) = \theta^\top z$ for all $z \in \mathbb{R}^d$, and $\mathcal{P}(\mu, \Sigma)$ is the Chebyshev ambiguity set of all distributions on $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{S}_+^d$, then the Chebyshev risk satisfies

$$\sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}[\ell(Z)] = \theta^\top\mu + \alpha\sqrt{\theta^\top\Sigma\theta}.$$

Proof. If $\theta^\top\Sigma\theta = 0$, then

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}[\theta^\top Z] &= \theta^\top\mu + \sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}[\theta^\top(Z-\mu)]\\ &= \theta^\top\mu + \sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}[0]\\ &= \theta^\top\mu,\end{aligned}$$

where the first equality holds because ϱ is translation-invariant, whereas the second equality holds because $\theta^\top(Z-\mu)$ equals 0 in law under any $\mathbb{P} \in \mathcal{P}(\mu, \Sigma)$ and because

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

ϱ is law-invariant. Finally, the third equality follows from the scale-invariance of ϱ. If $\theta^\top \Sigma \theta > 0$, on the other hand, then we have

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}[\theta^\top Z] &= \theta^\top \mu + \sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}[\theta^\top (Z-\mu)] \\ &= \theta^\top \mu + \sup_{\mathbb{P}\in\mathcal{P}(\mu,\Sigma)} \varrho_{\mathbb{P}}\left[\frac{\theta^\top (Z-\mu)}{\sqrt{\theta^\top \Sigma \theta}}\right] \sqrt{\theta^\top \Sigma \theta} \\ &= \theta^\top \mu + \alpha \sqrt{\theta^\top \Sigma \theta},\end{aligned}$$

where the first two equalities follow from the translation- and scale-invariance of ϱ, respectively. The third equality follows from Proposition 6.7, the law-invariance of ϱ and the definition of α. Indeed, the pushforward of the multivariate Chebyshev ambiguity set $\mathcal{P}(\mu, \Sigma)$ under the transformation $\ell(z) = \theta^\top (z-\mu)/\sqrt{\theta^\top \Sigma \theta}$ coincides with the univariate standard Chebyshev ambiguity set $\mathcal{P}(0, 1)$. □

The standard risk coefficient of a generic law-invariant risk measure may be difficult to compute. We now show, however, that the standard risk coefficients of the VaR and the CVaR match and are available in closed form.

Proposition 6.10 (Standard risk coefficients of VaR and CVaR). For any $\beta \in (0, 1)$, the standard risk coefficients of the β-VaR and the β-CVaR coincide, that is,

$$\sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \beta\text{-CVaR}_{\mathbb{Q}}[L] = \sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \beta\text{-VaR}_{\mathbb{Q}}[L] = \sqrt{\frac{1-\beta}{\beta}}.$$

Proof. As β-$\text{CVaR}_{\mathbb{Q}}[L]$ upper bounds β-$\text{VaR}_{\mathbb{Q}}[L]$ for every $\mathbb{Q} \in \mathcal{P}(0, 1)$, we have

$$\sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \beta\text{-CVaR}_{\mathbb{Q}}[L] \geq \sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \beta\text{-VaR}_{\mathbb{Q}}[L]. \tag{6.6}$$

The rest of the proof proceeds as follows. We first derive an analytical formula for the worst-case β-VaR on the right-hand side (Step 1). Next, we prove that the same analytical formula provides an upper bound on the worst-case β-CVaR on the left-hand side (Step 2). The claim then follows from the above inequality.

Step 1. We first express the worst-case β-VaR as its smallest upper bound to find

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \beta\text{-VaR}_{\mathbb{Q}}[L] &= \inf_{\tau\in\mathbb{R}} \{\tau : \beta\text{-VaR}_{\mathbb{Q}}(L) \leq \tau \ \forall \mathbb{Q} \in \mathcal{P}(0,1)\} \\ &= \inf_{\tau\in\mathbb{R}} \{\tau : \mathbb{Q}(L \geq \tau) \leq \beta \ \forall \mathbb{Q} \in \mathcal{P}(0,1)\} \\ &= \inf_{\tau\in\mathbb{R}} \left\{\tau : \frac{1}{1+\tau^2} \leq \beta\right\} \\ &= \sqrt{\frac{1-\beta}{\beta}}.\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The second equality in the above derivation follows from (5.2), and the third equality follows from the Marshall and Olkin bound of Proposition 6.6. The final formula is obtained by analytically solving the minimization problem over τ.

Step 2. The max-min inequality[2] and the definition of the β-CVaR imply that

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \beta\text{-CVaR}_{\mathbb{Q}}[L] &\leq \inf_{\tau\in\mathbb{R}} \sup_{\mathbb{Q}\in\mathcal{P}(0,1)} \tau + \frac{1}{\beta}\mathbb{E}_{\mathbb{Q}}[\max\{L-\tau,0\}] \\ &= \inf_{\tau\in\mathbb{R}} \tau + \frac{1}{2\beta}\left(\sqrt{1+\tau^2}-\tau\right) \\ &= \sqrt{\frac{1-\beta}{\beta}},\end{aligned}$$

where the first equality follows from Scarf's bound derived in Proposition 6.5, and the last equality is obtained by analytically solving the convex minimization problem over τ. The unique minimizer is given by

$$\tau^\star = \frac{1-2\beta}{2\sqrt{\beta(1-\beta)}}.$$

This completes Step 2. The claim then follows by combining the analytical formula for the worst-case β-VaR found in Step 1 and the analytical upper bound on the worst-case β-CVaR found in Step 2 with the elementary inequality (6.6). □

Propositions 6.9 and 6.10 provide an analytical formula for the Chebyshev risk of a linear loss function provided that the underlying risk measure is the VaR or the CVaR. The formula for the worst-case VaR was first derived by Lanckriet *et al.* (2001, 2002) and El Ghaoui *et al.* (2003); see also Calafiore and El Ghaoui (2006). The equality of the worst-case VaR and the worst-case CVaR was discovered by Zymler *et al.* (2013*a*). It holds not only for linear but also for arbitrary concave and arbitrary quadratic (not necessarily concave) loss functions. Proposition 6.9 follows from Nguyen *et al.* (2021). The standard risk coefficient can be characterized in closed form for a wealth of law-, translation- and scale-invariant risk measures other than the VaR and the CVaR. It is available, for instance, for all spectral risk measures and all risk measures that admit a Kusuoka representation (Li 2018) as well as all distortion risk measures (Cai *et al.* 2023); see also Nguyen *et al.* (2021).

6.8. *Gelbrich risk*

Let $\mathcal{G}_r(\hat{\mu},\hat{\Sigma})$ denote the Gelbrich ambiguity set of all distributions $\mathbb{P}\in\mathcal{P}(\mathbb{R}^d)$ whose mean–covariance pairs $(\mu,\Sigma)\in\mathbb{R}^d\times\mathbb{S}_+^d$ reside in a ball of radius $r\geq 0$ around $(\hat{\mu},\hat{\Sigma})\in\mathbb{R}^d\times\mathbb{S}_+^d$ with respect to the Gelbrich distance; see Definition 2.1.

[2] The Chebyshev ambiguity set $\mathcal{P}(0,1)$ is *not* weakly compact (see Example 3.9). Therefore Sion's minimax theorem does not allow us to interchange the infimum over τ and the supremum over $\mathbb{Q}$. While we could instead invoke Theorem 5.18, this is actually not needed to prove Proposition 6.10.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Recall from Section 2.1.4 that the Gelbrich ambiguity set accounts for moment ambiguity and thus often provides a more realistic account of uncertainty than a naïve Chebyshev ambiguity set. If the distribution of Z is only known to have a mean–covariance pair close to $(\hat{\mu}, \hat{\Sigma})$, then it is natural to quantify the riskiness of an uncertain loss $\ell(Z)$ under a law-invariant risk measure ϱ by the *Gelbrich risk*

$$\sup_{\mathbb{P} \in \mathcal{G}_r(\hat{\mu}, \hat{\Sigma})} \varrho_{\mathbb{P}}[\ell(Z)].$$

By construction, $\mathcal{G}_r(\hat{\mu}, \hat{\Sigma})$ is the union of all Chebyshev ambiguity sets $\mathcal{P}(\mu, \Sigma)$ corresponding to a mean–covariance pair (μ, Σ) with $\mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) \leq r$. This decomposition of the Gelbrich ambiguity set into Chebyshev ambiguity sets allows us via Proposition 6.9 to derive an analytical formula for the Gelbrich risk.

Proposition 6.11 (Gelbrich risk). Assume that ϱ is a law-, translation- and scale-invariant risk measure with standard risk coefficient $\alpha \in \mathbb{R}_+$, there is $\theta \in \mathbb{R}^d$ with $\ell(z) = \theta^\top z$ for all $z \in \mathbb{R}^d$, and $\mathcal{G}_r(\hat{\mu}, \hat{\Sigma})$ is the Gelbrich ambiguity set of all distributions on $\mathbb{R}^d$ whose mean–covariance pairs have a Gelbrich distance of at most $r \geq 0$ from $(\hat{\mu}, \hat{\Sigma}) \in \mathbb{R}^d \times \mathbb{S}_+^d$. Then the Gelbrich risk satisfies

$$\sup_{\mathbb{P} \in \mathcal{G}_r(\hat{\mu}, \hat{\Sigma})} \varrho_{\mathbb{P}}[\theta^\top Z] = \hat{\mu}^\top \theta + \alpha \sqrt{\theta^\top \hat{\Sigma} \theta} + r \sqrt{1 + \alpha^2}\, \|\theta\|_2. \tag{6.7}$$

Proof. Assume first that $\hat{\Sigma} > 0$. If $\theta = 0$, then the claim holds trivially because ϱ is law- and scale-invariant. If $r = 0$, then the claim follows immediately from Proposition 6.9. We may thus assume that $\theta \neq 0$ and $r > 0$. In this case we have

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{G}_r(\hat{\mu}, \hat{\Sigma})} \varrho_{\mathbb{P}}[\theta^\top Z] &= \begin{cases} \sup & \sup_{\mathbb{P} \in \mathcal{P}(\mu, \Sigma)} \varrho_{\mathbb{P}}[\theta^\top Z] \\ \text{s.t.} & \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{S}_+^d,\ \mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) \leq r \end{cases} \\
&= \begin{cases} \sup & \mu^\top \theta + \alpha \sqrt{\theta^\top \Sigma \theta} \\ \text{s.t.} & \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{S}_+^d \\ & \|\mu - \hat{\mu}\|^2 + \mathrm{Tr}\big(\Sigma + \hat{\Sigma} - 2(\hat{\Sigma}^{1/2} \Sigma \hat{\Sigma}^{1/2})^{1/2}\big) \leq r^2, \end{cases}
\end{aligned}$$

where the first equality exploits the decomposition of the Gelbrich ambiguity set into Chebyshev ambiguity sets. The second equality follows from Proposition 6.9 and Definition 2.1. By dualizing the resulting convex optimization problem, we find

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{G}_r(\hat{\mu}, \hat{\Sigma})} \varrho_{\mathbb{P}}[\theta^\top Z] = \inf_{\gamma \in \mathbb{R}_+} \Big\{ \gamma(r^2 - \mathrm{Tr}(\hat{\Sigma})) + \sup_{\mu \in \mathbb{R}^d} \big\{ \mu^\top \theta - \gamma \|\mu - \hat{\mu}\|^2 \big\} \\
+ \sup_{\Sigma \in \mathbb{S}_+^d} \big\{ \alpha \sqrt{\theta^\top \Sigma \theta} - \gamma\, \mathrm{Tr}\big(\Sigma - 2(\hat{\Sigma}^{1/2} \Sigma \hat{\Sigma}^{1/2})^{1/2}\big) \big\} \Big\}.
\end{aligned} \tag{6.8}$$

Strong duality holds because $r > 0$, which implies that $(\hat{\mu}, \hat{\Sigma})$ constitutes a Slater point for the primal maximization problem. If $\gamma = 0$, then the maximization problems over μ and Σ in (6.8) are unbounded. We may thus restrict γ to be strictly

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

positive. For any fixed $\gamma > 0$, the maximization problem over μ can be solved in closed form. Its optimal value is given by $\hat{\mu}^\top\theta + \|\theta\|^2/(4\gamma)$. By introducing an auxiliary variable t, the maximization problem over Σ can be reformulated as

$$\begin{array}{ll} \sup & \alpha t - \gamma \operatorname{Tr}\big(\Sigma - 2(\hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2})^{1/2}\big) \\ \text{s.t.} & t \in \mathbb{R}_+,\ \Sigma \in \mathbb{S}_+^d,\ t^2 - \theta^\top\Sigma\theta \leq 0. \end{array} \tag{6.9}$$

Note that $t = 0$ and $\Sigma = \theta\theta^\top$ form a Slater point for (6.9) because $\theta \neq 0$. Thus problem (6.9) admits a strong dual. The variable substitution $B \leftarrow (\hat{\Sigma}^{1/2}\Sigma\hat{\Sigma}^{1/2})^{1/2}$ allows us to reformulate this dual problem more concisely as

$$\inf_{\lambda\in\mathbb{R}_+} \sup_{t\in\mathbb{R}_+} \alpha t - \lambda t^2 + \sup_{B\in\mathbb{S}_+^d} \operatorname{Tr}(B^2\Delta_\lambda) + 2\gamma \operatorname{Tr}(B), \tag{6.10}$$

where

$$\Delta_\lambda = \hat{\Sigma}^{-1/2}(\lambda\theta\theta^\top - \gamma I_d)\hat{\Sigma}^{-1/2} \quad \text{for any } \lambda \geq 0.$$

Note that Δ_λ is well-defined because $\hat{\Sigma} \succ 0$. Recall now that the standard risk coefficient α was assumed to be non-negative. If $\lambda > 0$, then the supremum over t in (6.10) evaluates to $\alpha^2/(4\lambda)$. Otherwise, if $\lambda = 0$, then this supremum evaluates to $+\infty$. Therefore we may restrict the outer minimization problem in (6.10) to strictly positive λ. Similarly, if $\Delta_\lambda \not\prec 0$, then the supremum over B in (6.10) evaluates to $+\infty$. From now on, we may thus restrict the outer minimization problem in (6.10) to λ that satisfy $\gamma I_d - \lambda\theta\theta^\top \succ 0$. This constraint is equivalent to $\lambda < \gamma\|\theta\|^{-2}$ and guarantees that $\Delta_\lambda \prec 0$. As $\lambda > 0$, this in turn implies that $B^\star = -\gamma\Delta_\lambda^{-1}$ is positive definite and satisfies the first-order optimality condition $B\Delta_\lambda + \Delta_\lambda B + 2\gamma I_d = 0$. Note that this optimality condition can be interpreted as a continuous Lyapunov equation, and therefore its solution $B^\star$ is in fact unique; see e.g. Hespanha (2019, Theorem 12.5). By making the implicit constraints on λ explicit and by evaluating the two suprema over t and B analytically, problem (6.10) can finally be reformulated as

$$\begin{aligned} &\inf_{0<\lambda<\gamma\|\theta\|^{-2}} \frac{\alpha^2}{4\lambda} + \gamma^2 \operatorname{Tr}\big(\hat{\Sigma}^{1/2}(\gamma I_d - \lambda\theta\theta^\top)^{-1}\hat{\Sigma}^{1/2}\big) \\ &\quad = \inf_{0<\lambda<\gamma\|\theta\|^{-2}} \frac{\alpha^2}{4\lambda} + \gamma \operatorname{Tr}(\hat{\Sigma}) + \frac{\theta^\top\hat{\Sigma}\theta}{\lambda^{-1} - \|\theta\|^2/\gamma} \\ &\quad = \gamma \operatorname{Tr}(\hat{\Sigma}) + \frac{\alpha^2}{4}\frac{\|\theta\|^2}{\gamma} + \alpha\sqrt{\theta^\top\hat{\Sigma}\theta}. \end{aligned}$$

Here the first equality exploits the Sherman–Morrison formula (Bernstein 2009, Corollary 2.8.8) to rewrite the inverse matrix, and the second equality is obtained by solving the minimization problem over λ analytically. Indeed, the infimum is attained at the unique solution $\lambda^\star$ of the first-order condition

$$\frac{1}{\lambda} = \frac{\|\theta\|^2}{\gamma} + \frac{2}{\alpha}\sqrt{\theta^\top\hat{\Sigma}\theta}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

in the interior of the feasible set. In summary, we have solved both embedded subproblems in (6.8) analytically. Substituting their optimal values into (6.8) yields

$$\sup_{\mathbb{P}\in\mathcal{G}_r(\hat{\mu},\hat{\Sigma})} \varrho_{\mathbb{P}}[\theta^\top Z] = \inf_{\gamma\geq 0}\ \hat{\mu}^\top\theta + \alpha\sqrt{\theta^\top\hat{\Sigma}\theta} + \gamma r^2 + \frac{1+\alpha^2}{4}\frac{\|\theta\|^2}{\gamma}$$
$$= \hat{\mu}^\top\theta + \alpha\sqrt{\theta^\top\hat{\Sigma}\theta} + r\sqrt{1+\alpha^2}\,\|\theta\|.$$

Here the second equality is obtained by solving the minimization problem over γ in closed form. We have thus established the desired formula (6.7) for $\hat{\Sigma} \succ 0$.

It remains to be shown that (6.7) remains valid even if $\hat{\Sigma}$ is singular. To this end, use $J(\hat{\Sigma})$ as shorthand for the Gelbrich risk as a function of $\hat{\Sigma}$. By leveraging Berge's maximum theorem (Berge 1963, pp. 115–116) and the continuity of the Gelbrich distance (see the discussion after Proposition 2.2), it is easy to show that $J(\hat{\Sigma})$ is continuous on $\mathbb{S}_+^d$. The claim thus follows by noting that (6.7) holds for all $\hat{\Sigma} \succ 0$, that both sides of (6.7) are continuous in $\hat{\Sigma}$ and that every $\hat{\Sigma} \in \mathbb{S}_+^d$ can be expressed as a limit of positive definite matrices. □

Proposition 6.11 is due to Nguyen *et al.* (2021). It shows that, for a broad class of risk measures, the worst-case risk over a Gelbrich ambiguity set reduces to a Markowitz-type mean–variance risk functional with a 2-norm regularization term. We emphasize that the risk measure ρ enters the resulting optimization model only indirectly through the standard risk coefficient α.

6.9. Worst-case expectations over Kullback–Leibler ambiguity sets

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathrm{KL}(\mathbb{P},\hat{\mathbb{P}}) \leq r\}, \tag{6.11a}$$

which maximizes the expected value of $\ell(Z)$ over the Kullback–Leibler ambiguity set of all distributions supported on $\mathcal{Z}$ whose Kullback–Leibler divergence with respect to $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is at most $r \geq 0$. The Kullback–Leibler ambiguity set is a ϕ-divergence ambiguity set of the form (2.10), where ϕ satisfies $\phi(s) = s\log(s) - s + 1$ for all $s \geq 0$. As $\phi^\infty(1) = +\infty$, we have $\mathrm{KL}(\mathbb{P},\hat{\mathbb{P}}) = \infty$ unless $\mathbb{P} \ll \hat{\mathbb{P}}$. Hence problem (6.11a) maximizes only over distributions $\mathbb{P}$ that are absolutely continuous with respect to $\hat{\mathbb{P}}$. Note that $\phi^*(t) = \mathrm{e}^t - 1$ for all $t \in \mathbb{R}$. By Theorem 4.14 and the definition of the perspective function, the problem dual to (6.11a) is thus given by

$$\inf_{\lambda_0\in\mathbb{R},\,\lambda\in\mathbb{R}_+} \lambda_0 + \lambda(r-1) + \mathbb{E}_{\hat{\mathbb{P}}}\left[\lambda\exp\left(\frac{\ell(Z)-\lambda_0}{\lambda}\right)\right]. \tag{6.11b}$$

The problems (6.11a) and (6.11b) can be solved in closed form if the loss function ℓ is linear and the nominal distribution $\hat{\mathbb{P}}$ is Gaussian.

Proposition 6.12 (Worst-case expectations over KL ambiguity sets). Suppose that $\mathcal{Z} = \mathbb{R}^d$, $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a normal distribution with mean $\hat{\mu} \in \mathbb{R}^d$ and covariance

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

matrix $\hat{\Sigma} \in \mathbb{S}_{++}^d$, and $r > 0$. Suppose also that ℓ is linear, that is, there exists $\theta \in \mathbb{R}^d$ with $\ell(z) = \theta^\top z$ for all $z \in \mathcal{Z}$. Then the primal problem (6.11a) is solved by the normal distribution $\mathbb{P}^\star$ with mean $\hat{\mu} + (2r)^{1/2}\hat{\Sigma}\theta/(\theta^\top\hat{\Sigma}\theta)^{1/2}$ and covariance matrix $\hat{\Sigma}$. The dual problem (6.11b) is solved by $(\lambda_0^\star, \lambda^\star)$, where $\lambda^\star = (\theta^\top\hat{\Sigma}\theta)^{1/2}/(2r)^{1/2}$ and

$$\lambda_0^\star = \lambda^\star \log \mathbb{E}_{\hat{\mathbb{P}}}[\exp(\ell(Z)/\lambda^\star)].$$

The optimal values of (6.11a) and (6.11b) are both equal to $\hat{\mu}^\top\theta + (2r)^{1/2}(\theta^\top\hat{\Sigma}\theta)^{1/2}$.

Proof. Focus first on the dual problem (6.11b), and fix any $\lambda \geq 0$. Then the partial minimization problem over λ_0 is solved by

$$\lambda_0^\star(\lambda) = \lambda \log \mathbb{E}_{\hat{\mathbb{P}}}[\exp(\ell(Z)/\lambda)].$$

Substituting this parametric minimizer back into (6.11b) shows that the optimal value of the dual problem (6.11b) is given by

$$\begin{aligned}\inf_{\lambda\in\mathbb{R}_+} \lambda r + \lambda \log \mathbb{E}_{\hat{\mathbb{P}}}\left[\exp\left(\frac{\ell(Z)}{\lambda}\right)\right] &= \inf_{\lambda\in\mathbb{R}_+} \lambda r + \hat{\mu}^\top\theta + \frac{1}{2\lambda}\theta^\top\hat{\Sigma}\theta \\ &= \hat{\mu}^\top\theta + (2r)^{1/2}(\theta^\top\hat{\Sigma}\theta)^{1/2},\end{aligned}$$

where the first equality exploits the linearity of ℓ, the normality of $\hat{\mathbb{P}}$ and the formula for the expected value of a log-normal distribution. The second equality holds because the minimization problem over $\lambda \geq 0$ is solved by $\lambda^\star = (\theta^\top\hat{\Sigma}\theta)^{1/2}/(2r)^{1/2}$. Next, define $\mathbb{P}^\star \in \mathcal{P}(\mathcal{Z})$ as the normal distribution with mean $\mu^\star = \hat{\mu} + \hat{\Sigma}\theta/\lambda^\star$ and covariance matrix $\Sigma^\star = \hat{\Sigma}$. Comparing the density functions of $\hat{\mathbb{P}}$ and $\mathbb{P}^\star$ shows that

$$\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\hat{\mathbb{P}}}(z) = \exp\left(\frac{\theta^\top(z-\hat{\mu})}{\lambda^\star} - \frac{\theta^\top\hat{\Sigma}\theta}{2(\lambda^\star)^2}\right) \quad \text{for all } z \in \mathcal{Z}.$$

By Definition 2.8, we thus obtain

$$\mathrm{KL}(\mathbb{P}^\star, \hat{\mathbb{P}}) = \int_{\mathcal{Z}} \log\left(\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\hat{\mathbb{P}}}(z)\right) \mathrm{d}\mathbb{P}^\star(z) = \frac{\theta^\top\hat{\Sigma}\theta}{2(\lambda^\star)^2} = r,$$

where the second and third equalities follow readily from our formula for the Radon–Nikodym derivative $\mathrm{d}\mathbb{P}^\star/\mathrm{d}\hat{\mathbb{P}}$ and from basic algebra, respectively. Hence $\mathbb{P}^\star$ is feasible in (6.11b). In addition, its objective function value is given by

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] = \theta^\top\mu^\star = \hat{\mu}^\top\theta + (2r)^{1/2}(\theta^\top\hat{\Sigma}\theta)^{1/2}.$$

As the objective function values of $\mathbb{P}^\star$ and $(\lambda_0^\star, \lambda^\star)$ with $\lambda_0^\star = \lambda_0^\star(\lambda^\star)$ match, weak duality as established in Theorem 4.14 implies that $\mathbb{P}^\star$ is primal optimal and that $(\lambda_0^\star, \lambda^\star)$ is dual optimal. This observation completes the proof. □

Proposition 6.12 is due to Hu and Hong (2013). It is also reminiscent of risk-sensitive control theory (Hansen and Sargent 2008). In this stream of literature, a fictitious adversary may perturb the distribution of the exogenous noise terms of

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

an optimal control problem *arbitrarily* but incurs a penalty equal to the Kullback–Leibler divergence with respect to a Gaussian baseline model.

6.10. *Worst-case expectations over total variation balls*

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathrm{TV}(\mathbb{P},\hat{\mathbb{P}}) \le r\}, \tag{6.12a}$$

which maximizes the expected value of $\ell(Z)$ over a total variation ball of radius $r \in [0,1]$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$. Recall from Section 2.2.3 that the total variation distance is a ϕ-divergence and that the underlying entropy function satisfies $\phi(s) = \frac{1}{2}|s-1|$ for all $s \ge 0$ and $\phi(s) = \infty$ for all $s < 0$. Recall also that the total variation distance between two distributions is bounded above by 1 and that this bound is attained if the two distributions are mutually singular. An elementary calculation reveals that the conjugate entropy function satisfies $\phi^*(t) = \max\{t + \frac{1}{2}, 0\} - \frac{1}{2}$ if $t \le \frac{1}{2}$ and $\phi^*(t) = +\infty$ if $t > \frac{1}{2}$. By Theorem 4.14, the problem dual to (6.12a) is thus given by

$$\begin{array}{ll} \inf\limits_{\lambda_0\in\mathbb{R},\, \lambda\in\mathbb{R}_+} & \lambda_0 + \lambda\left(r - \frac{1}{2}\right) + \mathbb{E}_{\hat{\mathbb{P}}}\left[\max\left\{\ell(Z) - \lambda_0 + \frac{\lambda}{2}, 0\right\}\right] \\ \text{s.t.} & \lambda_0 + \lambda/2 \ge \sup\limits_{z\in\mathcal{Z}} \ell(z). \end{array} \tag{6.12b}$$

The problems (6.12a) and (6.12b) can be solved in closed form if $\mathcal{Z}$ is compact.

Proposition 6.13 (Worst-case expectations over total variation balls). Suppose that $\mathcal{Z} \subseteq \mathbb{R}^d$ is compact, $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ and $r \in (0,1)$, and define $\beta_r = 1 - r$. In addition, assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$ and ℓ is upper semicontinuous. Then the optimal values of (6.12a) and (6.12b) are both equal to

$$(1-\beta_r)\cdot \sup_{z\in\mathcal{Z}} \ell(z) + \beta_r \cdot \beta_r\text{-CVaR}_{\hat{\mathbb{P}}}[\ell(Z)]. \tag{6.13}$$

The proof of Proposition 6.13 will reveal that (6.12a) and (6.12b) are both solvable. Indeed, we will construct optimal solutions $\mathbb{P}^\star$ and $(\lambda_0^\star, \lambda^\star)$ for (6.12a) and (6.12b), respectively. A precise description of these optimizers is cumbersome and thus omitted from the proposition statement. If the loss $\ell(Z)$ has a continuous distribution under $\hat{\mathbb{P}}$, however, then $\mathbb{P}^\star$ admits a simpler and more intuitive description. Indeed, in this case, $\mathbb{P}^\star$ is obtained from $\hat{\mathbb{P}}$ by shifting the probability mass of all outcomes $z \in \mathcal{Z}$ associated with a high loss $\ell(z) \ge \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)]$ to some outcome $z \in \mathcal{Z}$ associated with the highest possible loss $\ell(z) = \max_{z'\in\mathcal{Z}} \ell(z')$.

Proof of Proposition 6.13. For ease of notation, set $\overline{\ell} = \sup_{z\in\mathcal{Z}} \ell(z)$. Focus first on the dual problem (6.12b), and fix any $\lambda \ge 0$. Note that the dual objective function is non-decreasing in λ_0. The partial minimization problem over λ_0 is therefore solved by $\lambda_0^\star(\lambda) = \overline{\ell} - \lambda/2$. Substituting this parametric minimizer back into (6.12b) shows

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

that the optimal value of the dual problem is given by

$$\begin{aligned}&\overline{\ell} + \inf_{\lambda \in \mathbb{R}_+} \ \lambda(r-1) + \mathbb{E}_{\hat{\mathbb{P}}}[\max\{\ell(Z) - \overline{\ell} + \lambda, 0\}] \\ &\quad = r\,\overline{\ell} + (1-r) \inf_{\tau \le \overline{\ell}} \ \tau + (1-r)^{-1} \mathbb{E}_{\hat{\mathbb{P}}}[\max\{\ell(Z) - \tau, 0\}],\end{aligned}$$

where the equality follows from the substitution $\tau \leftarrow \overline{\ell} - \lambda$. By Definition 5.10, the infimum over τ evaluates to β_r-$\mathrm{CVaR}_{\hat{\mathbb{P}}}[\ell(Z)]$ with $\beta_r = 1 - r$. Recall that this infimum is attained by $\tau^\star = \beta_r$-$\mathrm{VaR}_{\hat{\mathbb{P}}}[\ell(Z)]$, which is bounded above by $\overline{\ell}$. In summary, we have thus shown that the optimal value of problem (6.12b) equals

$$(1-\beta_r) \cdot \overline{\ell} + \beta_r \cdot \beta_r\text{-}\mathrm{CVaR}_{\hat{\mathbb{P}}}[\ell(Z)].$$

To construct a primal maximizer, assume first that $\hat{\mathbb{P}}(\ell(Z) < \overline{\ell}) \le r$, which implies that β_r-$\mathrm{CVaR}_{\hat{\mathbb{P}}}[\ell(Z)] = \overline{\ell}$. Thus the optimal value of the dual problem (6.12b) simplifies to $\overline{\ell}$, which is attained by any distribution $\mathbb{P}^\star$ that is obtained from $\hat{\mathbb{P}}$ by moving all probability mass from $\{z \in \mathcal{Z} : \ell(z) < \overline{\ell}\}$ to $\{z \in \mathcal{Z} : \ell(z) = \overline{\ell}\}$.

Next, assume that $\hat{\mathbb{P}}(\ell(Z) < \overline{\ell}) > r$, which implies that β_r-$\mathrm{VaR}_{\hat{\mathbb{P}}}[\ell(Z)] < \overline{\ell}$. In this case, we partition $\mathcal{Z}$ into the following four subsets:

$$\begin{aligned}\mathcal{Z}_1 &= \{z \in \mathcal{Z} : \beta_r\text{-}\mathrm{VaR}_{\hat{\mathbb{P}}}[\ell(Z)] > \ell(z)\}, \\ \mathcal{Z}_2 &= \{z \in \mathcal{Z} : \overline{\ell} > \ell(z) = \beta_r\text{-}\mathrm{VaR}_{\hat{\mathbb{P}}}[\ell(Z)]\}, \\ \mathcal{Z}_3 &= \{z \in \mathcal{Z} : \overline{\ell} > \ell(z) > \beta_r\text{-}\mathrm{VaR}_{\hat{\mathbb{P}}}[\ell(Z)]\}, \\ \mathcal{Z}_4 &= \{z \in \mathcal{Z} : \overline{\ell} = \ell(z)\}.\end{aligned}$$

Note that $\mathcal{Z}_1$ and $\mathcal{Z}_3$ can be empty, whereas $\mathcal{Z}_2$ and $\mathcal{Z}_4$ must be non-empty. We also define $\hat{\mathbb{P}}_i$ as the nominal distribution $\hat{\mathbb{P}}$ conditioned on the event $Z \in \mathcal{Z}_i$ for all $i \in [4]$, and we define $\mathbb{U}_{\mathcal{Z}_4}$ as the uniform distribution on $\mathcal{Z}_4$. Next, we set

$$\begin{aligned}\mathbb{P}^\star = (\beta_r - \hat{\mathbb{P}}(Z \in \mathcal{Z}_3) - \hat{\mathbb{P}}(Z \in \mathcal{Z}_4)) \cdot \hat{\mathbb{P}}_2 \\ + \hat{\mathbb{P}}(Z \in \mathcal{Z}_3) \cdot \hat{\mathbb{P}}_3 + \hat{\mathbb{P}}(Z \in \mathcal{Z}_4) \cdot \hat{\mathbb{P}}_4 + (1-\beta_r) \cdot \mathbb{U}_{\mathcal{Z}_4}.\end{aligned}$$

Thus $\mathbb{P}^\star$ is a mixture of four probability distributions. As the non-negative mixture probabilities sum to 1, $\mathbb{P}^\star$ is a probability distribution. Using $\rho = \hat{\mathbb{P}} + \mathbb{U}_{\mathcal{Z}_4}$ as a dominating measure for $\hat{\mathbb{P}}$ and $\mathbb{P}^\star$ and recalling that $\phi(s) = \frac{1}{2}|s-1|$ if $s \ge 0$, we find

$$\begin{aligned}&\mathrm{TV}(\mathbb{P}^\star, \hat{\mathbb{P}}) \\ &\quad = \mathrm{D}_\phi(\mathbb{P}^\star, \hat{\mathbb{P}}) \\ &\quad = \frac{1}{2} \sum_{i=1}^{4} \int_{\mathcal{Z}_i} \left| \frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\rho}(z) - \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z) \right| \mathrm{d}\rho(z) \\ &\quad = \hat{\mathbb{P}}(Z \in \mathcal{Z}_1) + (\hat{\mathbb{P}}(Z \in \mathcal{Z}_2) + \hat{\mathbb{P}}(Z \in \mathcal{Z}_3) + \hat{\mathbb{P}}(Z \in \mathcal{Z}_4) - \beta_r) + 0 + (1-\beta_r) \\ &\quad = r,\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where the third equality follows from the definition of $\mathbb{P}^\star$ and the relation

$$\hat{\mathbb{P}}(Z \in \mathcal{Z}_2) + \hat{\mathbb{P}}(Z \in \mathcal{Z}_3) + \hat{\mathbb{P}}(Z \in \mathcal{Z}_4) = \hat{\mathbb{P}}(\ell(Z) \geq \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)]) \geq \beta_r,$$

and the last equality follows from the definition of β_r. Thus $\mathbb{P}^\star$ is feasible in (6.12a). In addition, the objective function value of $\mathbb{P}^\star$ in (6.12a) amounts to

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] \\
&\quad = (1-\beta_r)\cdot\overline{\ell} \\
&\qquad + \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z) \mid \ell(Z) > \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)]] \cdot \hat{\mathbb{P}}(\ell(Z) > \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)]) \\
&\qquad + \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z) \mid \ell(Z) = \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)]] \cdot (\beta_r - \hat{\mathbb{P}}(\ell(Z) > \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)])) \\
&\quad = (1-\beta_r)\cdot\overline{\ell} + \beta_r \cdot \beta_r\text{-CVaR}_{\hat{\mathbb{P}}}[\ell(Z)].
\end{aligned}$$

Here the second equality follows from Föllmer and Schied (2008, Theorem 4.47, Remark 4.48). Note that if the marginal distribution of $\ell(Z)$ is continuous under $\hat{\mathbb{P}}$, then the above derivation simplifies. Indeed, in this case we have

$$\hat{\mathbb{P}}(\ell(Z) > \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)]) = \beta_r$$

and

$$\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z) \mid \ell(Z) > \beta_r\text{-VaR}_{\hat{\mathbb{P}}}[\ell(Z)]] = \beta_r\text{-CVaR}_{\hat{\mathbb{P}}}[\ell(Z)].$$

Irrespective of $\hat{\mathbb{P}}$, the objective function value of $\mathbb{P}^\star$ in (6.12a) matches the optimal value of (6.12b). Weak duality as established in Theorem 4.14 thus implies that $\mathbb{P}^\star$ solves the primal problem (6.12a). This observation completes the proof. □

Jiang and Guan (2018) and Shapiro (2017) study a variant of problem (6.12a) that maximizes over a *restricted* total variation ball. Thus they additionally impose $\mathbb{P} \ll \hat{\mathbb{P}}$ in (6.12a). The supremum of the resulting restricted problem amounts to

$$(1-\beta_r)\cdot \operatorname{ess\,sup}_{\hat{\mathbb{P}}}[\ell(Z)] + \beta_r \cdot \beta_r\text{-CVaR}_{\hat{\mathbb{P}}}[\ell(Z)],$$

which may be strictly smaller than (6.13). If additionally $\ell(Z)$ has a continuous marginal distribution under $\hat{\mathbb{P}}$, then the supremum is no longer attained.

6.11. Worst-case expectations over Lévy–Prokhorov balls

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] \colon \mathrm{LP}(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}, \tag{6.14a}$$

which maximizes the expected value of $\ell(Z)$ over a Lévy–Prokhorov ball of radius $r \in [0,1]$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$. We assume here that the Lévy–Prokhorov distance is induced by a norm $\|\cdot\|$ on $\mathbb{R}^d$. By Proposition 2.22, the Lévy–Prokhorov ball of radius $r \in (0,1)$ coincides with the optimal transport ambiguity set

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) \leq r\},$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where the transportation cost function c_r is defined by $c_r(z, \hat{z}) = \mathbb{1}_{\|z-\hat{z}\|>r}$. Theorem 4.18 thus implies that the problem dual to (6.14a) is given by

$$\inf_{\lambda \in \mathbb{R}_+} \left\{ \lambda r + \mathbb{E}_{\hat{\mathbb{P}}} \left[\sup_{z \in \mathcal{Z}} \ell(z) - \lambda c_r(z, \hat{Z}) \right] \right\} \tag{6.14b}$$

whenever ℓ is upper semicontinuous. If $\mathcal{Z}$ is compact, then we can leverage Proposition 6.13 to solve the problems (6.14a) and (6.14b) in closed form.

Proposition 6.14 (Worst-case expectations over Lévy–Prokhorov balls). Suppose that $\mathcal{Z} \subseteq \mathbb{R}^d$ is compact, $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ and $r \in (0, 1)$, and define $\beta_r = 1 - r$. In addition, assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] > -\infty$ and ℓ is upper semicontinuous. Then the optimal values of (6.14a) and (6.14b) are both equal to

$$(1 - \beta_r) \cdot \sup_{z \in \mathcal{Z}} \ell(z) + \beta_r \cdot \beta_r\text{-CVaR}_{\hat{\mathbb{P}}}[\ell_r(\hat{Z})], \tag{6.15}$$

where $\ell_r(\hat{z}) = \sup_{z \in \mathcal{Z}} \{\ell(z) \colon \|z - \hat{z}\| \leq r\}$ is an adversarial loss function that assigns each $\hat{z} \in \mathcal{Z}$ the worst-case loss in the r-neighbourhood of $\hat{z}$.

The proof of Proposition 6.14 will reveal that (6.14a) and (6.14b) are both solvable. However, a precise description of the respective optimizers is cumbersome and thus omitted from the proposition statement. Note that the adversarial loss function ℓ_r inherits upper semicontinuity from ℓ thanks to Berge (1963, Theorem 2, p. 116). The following lemma is needed in the proof of Proposition 6.14.

Lemma 6.15. Assume that $\mathcal{Z} \subseteq \mathbb{R}^d$ is compact, ℓ is upper semicontinuous, $\hat{z} \in \mathcal{Z}$ and $r, \lambda \geq 0$. Then the following identity holds:

$$\sup_{z \in \mathcal{Z}} \{\ell(z) - \lambda \cdot \mathbb{1}_{\|z-\hat{z}\|>r}\} = \sup_{z \in \mathcal{Z}} \{\ell_r(z) - \lambda \cdot \mathbb{1}_{z \neq \hat{z}}\}.$$

Proof. For ease of notation we introduce two auxiliary functions f and g from $\mathcal{Z}$ to $\overline{\mathbb{R}}$, which are defined by $f(z) = \ell(z) - \lambda \cdot \mathbb{1}_{\|z-\hat{z}\|>r}$ and $g(z) = \ell_r(z) - \lambda \cdot \mathbb{1}_{z \neq \hat{z}}$ for all $z \in \mathcal{Z}$. Note that both f and g are upper semicontinuous.

First, select $z^\star \in \arg\max_{z \in \mathcal{Z}} f(z)$, which exists because $\mathcal{Z}$ is compact and f is upper semicontinuous. If $\|z^\star - \hat{z}\| > r$, then the definition of ℓ_r implies that

$$\sup_{z \in \mathcal{Z}} f(z) = f(z^\star) = \ell(z^\star) - \lambda \leq \ell_r(z^\star) - \lambda = g(z^\star) \leq \sup_{z \in \mathcal{Z}} g(z).$$

On the other hand, if $\|z - \hat{z}\| \leq r$, then

$$\sup_{z \in \mathcal{Z}} f(z) = f(z^\star) = \ell(z^\star) \leq \ell_r(\hat{z}) = g(\hat{z}) \leq \sup_{z \in \mathcal{Z}} g(z).$$

Next, select $\tilde{z} \in \arg\max_{z \in \mathcal{Z}} g(z)$. If $\tilde{z} \neq \hat{z}$, then with $z^\star \in \arg\max_{z \in \mathcal{Z}} \ell(z)$ we have

$$\sup_{z \in \mathcal{Z}} g(z) = g(\tilde{z}) = \ell_r(\tilde{z}) - \lambda \leq \ell(z^\star) - \lambda \leq f(z^\star) = \sup_{z \in \mathcal{Z}} f(z),$$

where the inequalities follow from the definition of $z^\star$ and the non-negativity of λ.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Conversely, if $\tilde{z} = \hat{z}$, then with $z_r^\star \in \arg\max_{z' \in \mathcal{Z}} \{\ell(z') \colon \|z' - \hat{z}\| \le r\}$ we have

$$\sup_{z \in \mathcal{Z}} g(z) = g(\tilde{z}) = \ell_r(\hat{z}) = \ell(z_r^\star) = f(z_r^\star) = \sup_{z \in \mathcal{Z}} f(z).$$

Thus the claim follows. □

Proof of Proposition 6.14. Lemma 6.15 allows us to reformulate the dual problem (6.14b) in terms of the adversarial loss function ℓ_r as

$$\inf_{\lambda \in \mathbb{R}_+} \left\{ \lambda r + \mathbb{E}_{\hat{\mathbb{P}}} \left[\sup_{z \in \mathcal{Z}} \ell_r(z) - \lambda \cdot \mathbb{1}_{z \neq \hat{z}} \right] \right\}. \tag{6.16}$$

As $r > 0$, $\mathcal{Z}$ is compact and ℓ_r is upper semicontinuous, Theorem 4.18 implies that (6.16) is the strong dual of a problem that maximizes the expected value of the adversarial loss function ℓ_r over an optimal transport ambiguity set corresponding to the transportation cost function $c_0(z, \hat{z}) = \mathbb{1}_{z \neq \hat{z}}$. Its optimal value thus matches

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell_r(Z)] \colon \mathrm{OT}_{c_0}(\mathbb{P}, \hat{\mathbb{P}}) \le r\} = \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell_r(Z)] \colon \mathrm{TV}(\mathbb{P}, \hat{\mathbb{P}}) \le r\},$$

where the equality holds because $\mathrm{TV} = \mathrm{OT}_{c_0}$ as shown in Proposition 2.24. Since $\sup_{z \in \mathcal{Z}} \ell_r(z) = \sup_{z \in \mathcal{Z}} \ell(z) = \overline{\ell}$, Proposition 6.13 readily implies that the supremum of the resulting maximization problem over a total variation ball is given by

$$(1 - \beta_r) \cdot \overline{\ell} + \beta_r \cdot \beta_r\text{-CVaR}_{\hat{\mathbb{P}}}[\ell_r(\hat{Z})],$$

Assume now that $\psi \colon \mathcal{Z} \to \mathcal{Z}$ is a Borel-measurable function satisfying

$$\psi(\hat{z}) \in \arg\max_{z \in \mathcal{Z}} \{\ell(z) \colon \|z - \hat{z}\| \le r\} \quad \text{for all } \hat{z} \in \mathcal{Z},$$

which exists thanks to Rockafellar and Wets (2009, Corollary 14.6, Theorem 14.37), and define $\hat{\mathbb{P}}_\psi = \hat{\mathbb{P}} \circ \psi^{-1}$ as the pushforward distribution of $\hat{\mathbb{P}}$ under ψ. Next, we construct a primal maximizer under the assumption that $\hat{\mathbb{P}}_\psi(\ell(Z) < \overline{\ell}) > r$. To this end, we partition $\mathcal{Z}$ into the following four subsets:

$$\begin{aligned}
\mathcal{Z}_1 &= \{z \in \mathcal{Z} \colon \beta_r\text{-VaR}_{\hat{\mathbb{P}}_\psi}[\ell(\hat{Z})] > \ell(z)\},\\
\mathcal{Z}_2 &= \{z \in \mathcal{Z} \colon \overline{\ell} > \ell(z) = \beta_r\text{-VaR}_{\hat{\mathbb{P}}_\psi}[\ell(\hat{Z})]\},\\
\mathcal{Z}_3 &= \{z \in \mathcal{Z} \colon \overline{\ell} > \ell(z) > \beta_r\text{-VaR}_{\hat{\mathbb{P}}_\psi}[\ell(\hat{Z})]\},\\
\mathcal{Z}_4 &= \{z \in \mathcal{Z} \colon \overline{\ell} = \ell(z)\}.
\end{aligned}$$

We also define $\hat{\mathbb{P}}_i$ as the distribution $\hat{\mathbb{P}}_\psi$ conditioned on the event $\hat{Z} \in \mathcal{Z}_i$ for all $i \in [4]$, and we define $\mathbb{U}_{\mathcal{Z}_4}$ as the uniform distribution on $\mathcal{Z}_4$. Next, we set

$$\begin{aligned}
\mathbb{P}^\star = {} & (\beta_r - \hat{\mathbb{P}}_\psi(\hat{Z} \in \mathcal{Z}_3) - \hat{\mathbb{P}}_\psi(\hat{Z} \in \mathcal{Z}_4)) \cdot \hat{\mathbb{P}}_2\\
& + \hat{\mathbb{P}}_\psi(\hat{Z} \in \mathcal{Z}_3) \cdot \hat{\mathbb{P}}_3 + \hat{\mathbb{P}}_\psi(\hat{Z} \in \mathcal{Z}_4) \cdot \hat{\mathbb{P}}_4 + (1 - \beta_r) \cdot \mathbb{U}_{\mathcal{Z}_4}.
\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Note that $\mathbb{P}^\star$ is constructed as in the proof of Proposition 6.13, the only difference being that $\hat{\mathbb{P}}$ is now replaced with its pushforward distribution $\hat{\mathbb{P}}_\psi$. We then find

$$\begin{aligned}\mathrm{LP}(\mathbb{P}^\star, \hat{\mathbb{P}}) &\leq \max\{\mathrm{OT}_{c_r}(\mathbb{P}^\star, \hat{\mathbb{P}}), r\} \\ &\leq \max\{\mathrm{OT}_{c_r}(\mathbb{P}^\star, \hat{\mathbb{P}}_\psi) + \mathrm{OT}_{c_r}(\hat{\mathbb{P}}_\psi, \hat{\mathbb{P}}), r\} \\ &\leq \max\{\mathrm{TV}(\mathbb{P}^\star, \hat{\mathbb{P}}_\psi), r\} \\ &= r,\end{aligned}$$

where the first inequality follows from Proposition 2.22, and the second inequality holds because c_r is a pseudo-metric on $\mathcal{Z}$, which implies that OT_{c_r} is a pseudo-metric on $\mathcal{P}(\mathcal{Z})$ and thus satisfies the triangle inequality. The third inequality holds because $\mathrm{OT}_{c_r}(\hat{\mathbb{P}}_\psi, \hat{\mathbb{P}}) = 0$ and because $c_0(z, \hat{z}) \geq c_r(z, \hat{z})$ for all $z, \hat{z} \in \mathcal{Z}$, which implies that $\mathrm{OT}_{c_r}(\mathbb{P}^\star, \hat{\mathbb{P}}_\psi) \leq \mathrm{TV}(\mathbb{P}^\star, \hat{\mathbb{P}}_\psi)$. Finally, the equality follows from the proof of Proposition 6.13, which ensures that $\mathrm{TV}(\mathbb{P}^\star, \hat{\mathbb{P}}_\psi) = r$. We also have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] &= (1 - \beta_r) \cdot \overline{\ell} + \beta_r \cdot \beta_r\text{-CVaR}_{\hat{\mathbb{P}}_\psi}[\ell(\hat{Z})] \\ &= (1 - \beta_r) \cdot \overline{\ell} + \beta_r \cdot \beta_r\text{-CVaR}_{\hat{\mathbb{P}}}[\ell(\psi(\hat{Z}))].\end{aligned}$$

where the two equalities again follow from the proof of Proposition 6.13 and from the measure-theoretic change of variables formula, respectively. As $\ell(\psi(\hat{z})) = \ell_r(\hat{z})$ for every $\hat{z} \in \mathcal{Z}$, the objective function value of $\mathbb{P}^\star$ in (6.14a) matches the optimal value of the dual problem (6.14b). Weak duality as established in Theorem 4.18 thus implies that $\mathbb{P}^\star$ solves the primal problem (6.14a). If $\hat{\mathbb{P}}_\psi(\ell(Z) < \overline{\ell}) \leq r$, the construction of a primal maximizer is simpler, and is thus omitted for brevity. □

The results of this section were first obtained by Bennouna and Van Parys (2023) under the assumption that the nominal distribution $\hat{\mathbb{P}}$ is discrete.

6.12. Worst-case expectations over ∞-Wasserstein balls

Consider the worst-case expectation problem

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathrm{W}_\infty(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}, \tag{6.17a}$$

which maximizes the expected value of $\ell(Z)$ over an ∞-Wasserstein ball of radius $r \in \mathbb{R}_+$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$. We assume here that the ∞-Wasserstein distance is induced by a given norm $\|\cdot\|$ on $\mathbb{R}^d$. Recall from Proposition 2.27 that the ∞-Wasserstein ambiguity set coincides with the optimal transport ambiguity set

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) : \mathrm{OT}_{c_r}(\mathbb{P}, \hat{\mathbb{P}}) \leq 0\},$$

where the transportation cost function c_r is defined by $c_r(z, \hat{z}) = \mathbb{1}_{\|z - \hat{z}\| > r}$. We emphasize that, while the radius of the ∞-Wasserstein ball under consideration is r, the radius of the corresponding optimal transport ambiguity set $\mathcal{P}$ is 0. Theorem 4.18

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

thus implies that the problem dual to (6.17a) is given by

$$\inf_{\lambda\in\mathbb{R}_+} \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c_r(z,\hat{Z})\right] \tag{6.17b}$$

whenever ℓ is upper semicontinuous. If $\mathcal{Z}$ is compact, then the problems (6.17a) and (6.17b) can be solved in closed form.

Proposition 6.16 (Worst-case expectations over ∞-Wasserstein balls). Suppose that $\mathcal{Z} \subseteq \mathbb{R}^d$ is compact, $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$, $r \in \mathbb{R}_+$, $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$ and ℓ is upper semicontinuous. Define the adversarial loss function

$$\ell_r(\hat{z}) = \sup_{z\in\mathcal{Z}}\{\ell(z)\colon \|z-\hat{z}\| \le r\}$$

as in Proposition 6.14, and let $\psi\colon \mathcal{Z} \to \mathcal{Z}$ be a Borel function that satisfies

$$\psi(\hat{z}) \in \arg\max_{z\in\mathcal{Z}}\{\ell(z)\colon \|z-\hat{z}\| \le r\} \quad \text{for all } \hat{z}\in\mathcal{Z}.$$

Then the primal problem (6.17a) is solved by $\mathbb{P}^\star = \hat{\mathbb{P}}\circ\psi^{-1}$. In addition, the optimal values of (6.17a) and (6.17b) are both equal to $\mathbb{E}_{\hat{\mathbb{P}}}[\ell_r(\hat{Z})]$.

Proof. Note that the Borel function ψ exists thanks to Rockafellar and Wets (2009, Corollary 14.6, Theorem 14.37). This ensures that the pushforward distribution $\mathbb{P}^\star = \hat{\mathbb{P}}\circ\psi^{-1}$ is well-defined. Note also that $\mathbb{P}^\star$ is feasible in (6.17a) because

$$\mathrm{W}_\infty(\mathbb{P}^\star,\hat{\mathbb{P}}) = \inf\{r'\ge 0\colon \mathrm{OT}_{c_{r'}}(\mathbb{P}^\star,\hat{\mathbb{P}}) \le 0\} \le r,$$

where the equality follows from Proposition 2.27 with $d(z,\hat{z}) = \|z-\hat{z}\|$, and the inequality holds because $\mathrm{OT}_{c_r}(\mathbb{P}^\star,\hat{\mathbb{P}}) = 0$. We also have

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\psi(Z))] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell_r(Z)].$$

Next, note that $\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c_r(z,\hat{z})$ is non-increasing in λ for any fixed $\hat{z}\in\mathcal{Z}$. Also, it is uniformly bounded above by $\sup_{z\in\mathcal{Z}}\ell(z)$, which is a finite constant thanks to the compactness of $\mathcal{Z}$ and the upper semicontinuity of ℓ. By the monotone convergence theorem, the optimal value of the dual problem (6.17b) thus satisfies

$$\inf_{\lambda\in\mathbb{R}_+} \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c_r(z,\hat{Z})\right] = \mathbb{E}_{\hat{\mathbb{P}}}\left[\inf_{\lambda\in\mathbb{R}_+}\sup_{z\in\mathcal{Z}} \ell(z) - \lambda c_r(z,\hat{Z})\right] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell_r(\hat{Z})],$$

where the second equality holds because $\mathcal{Z}$ is compact. Weak duality as established in Theorem 4.18 thus implies that $\mathbb{P}^\star$ solves the primal problem (6.17a). □

Proposition 6.16 shows that the worst-case expectation of the original loss $\ell(Z)$ with respect to an ∞-Wasserstein ball coincides with the crisp expectation of the adversarial loss $\ell_r(\hat{Z})$ with respect to the nominal distribution $\hat{\mathbb{P}}$. This result was first discovered by Gao *et al.* (2017) for discrete nominal distributions and later extended by Gao *et al.* (2024*b*) to general nominal distributions. The loss function ℓ_r is routinely used in machine learning for the adversarial training of neural networks (Szegedy *et al.* 2014, Goodfellow *et al.* 2015). Proposition 6.16 thus reveals

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

an intimate connection between adversarial training and distributionally robust optimization with respect to an ∞-Wasserstein ambiguity set. This connection has been further explored in the context of adversarial classification by García Trillos and García Trillos (2022), García Trillos and Murray (2022), García Trillos and Jacobs (2023), Bungert *et al.* (2023, 2024), Pydi and Jog (2024) and Frank and Niles-Weed (2024*a*,*b*).

6.13. Worst-case expectations over 1*-Wasserstein balls*

Consider the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{\mathbb{P}}[\ell(Z)] : \mathrm{W}_1(\mathbb{P},\hat{\mathbb{P}}) \le r\}, \tag{6.18a}$$

which maximizes the expected value of $\ell(Z)$ over a 1-Wasserstein ball of radius $r \in \mathbb{R}_+$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$. We assume here that the 1-Wasserstein distance is induced by a given norm $\|\cdot\|$ on $\mathbb{R}^d$. Thus the 1-Wasserstein ambiguity set coincides with the optimal transport ambiguity set $\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) : \mathrm{OT}_c(\mathbb{P},\hat{\mathbb{P}}) \le r\}$ corresponding to the transportation cost function c is defined by $c(z,\hat{z}) = \|z - \hat{z}\|$. Theorem 4.18 thus implies that the problem dual to (6.18a) is given by

$$\inf_{\lambda\ge 0} \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda\|z - \hat{Z}\|\right] \tag{6.18b}$$

whenever ℓ is upper semicontinuous. If $\mathcal{Z} = \mathbb{R}^d$ and ℓ is convex and Lipschitz-continuous, then the problems (6.18a) and (6.18b) can be solved in closed form.

Proposition 6.17 (Worst-case expectations over 1-Wasserstein balls). Suppose that $\mathcal{Z} = \mathbb{R}^d$, $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ and $r \in \mathbb{R}_+$. If $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] > -\infty$ and ℓ is convex and Lipschitz-continuous, then the optimal values of (6.18a) and (6.18b) are equal to

$$\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + r\,\mathrm{lip}(\ell).$$

Under the conditions of Proposition 6.17, the supremum of the primal problem (6.18a) is usually *not* attained. The proof constructs a sequence of distributions that attain the supremum asymptotically. These distributions move an increasingly small portion of $\hat{\mathbb{P}}$ increasingly far along the direction of steepest increase of ℓ. Intuitively, the amount of probability mass transported over a distance Δ must decay as $O(r/\Delta)$ as Δ grows. The dual problem (6.18b) is solved by $\lambda^\star = \mathrm{lip}(\ell)$.

Proof of Proposition 6.17. As the convex function ℓ is Lipschitz-continuous, it is in particular proper and closed. By the Fenchel–Moreau theorem (Lemma 4.2), ℓ thus admits the dual representation

$$\ell(z) = \sup_{y\in\mathrm{dom}(\ell^*)} z^\top y - \ell^*(y),$$

where ℓ^* denotes the convex conjugate of ℓ. Put differently, ℓ coincides with the pointwise supremum of the affine functions $f_y(z) = y^\top z - \ell^*(y)$ parametrized by

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$y \in \mathrm{dom}(\ell^*)$. Hölder's inequality then implies that

$$|f_y(z) - f_y(\hat{z})| = |y^\top (z - \hat{z})| \leq \|y\|_* \|z - \hat{z}\|,$$

where $\|\cdot\|_*$ denotes the norm dual to $\|\cdot\|$. As Hölder's inequality is tight, f_y is Lipschitz-continuous with Lipschitz modulus $\mathrm{lip}(f_y) = \|y\|_*$. In addition, as the Lipschitz modulus of a supremum of affine functions coincides with the supremum of the corresponding Lipschitz moduli, the Lipschitz modulus of ℓ is given by

$$\mathrm{lip}(\ell) = \sup_{y \in \mathrm{dom}(\ell^*)} \|y\|_* = \max_{y \in \mathrm{cl}(\mathrm{dom}(\ell^*))} \|y\|_*.$$

The maximum in the last expression is attained by some $y^\star \in \mathbb{R}^d$ because $\mathrm{lip}(\ell) < \infty$ by assumption. Next, define $z^\star$ as any optimal solution of $\max_{\|z\| \leq 1} (y^\star)^\top z$. By construction, we thus have $(y^\star)^\top z^\star = \|y^\star\|_*$. We also introduce a sequence $\{y_i\}_{i \in \mathbb{N}}$ in $\mathrm{dom}(\ell^*)$ that converges to $y^\star$, and we set $q_i = i^{-1}(1 + |\ell^*(y_i)|)^{-1}$ for every $i \in \mathbb{N}$. In addition, we define $f_i \colon \mathbb{R}^d \to \mathbb{R}^d$ through $f_i(z) = z + r z^\star / q_i$ for any $i \in \mathbb{N}$. Thus f_i represents the translation that shifts each point in $\mathbb{R}^d$ along the direction $z^\star$ by a distance equal to r/q_i. We further define

$$\mathbb{P}_i = (1 - q_i)\, \hat{\mathbb{P}} + q_i\, \hat{\mathbb{P}} \circ f_i^{-1},$$

where $\hat{\mathbb{P}} \circ f_i^{-1}$ stands for the pushforward distribution of $\hat{\mathbb{P}}$ under f_i. Intuitively, $\mathbb{P}_i$ is obtained by decomposing $\hat{\mathbb{P}}$ into two parts $(1 - q_i)\hat{\mathbb{P}}$ and $q_i \hat{\mathbb{P}}$ and then translating the second part by $r z^\star / q_i$. By construction, we thus have $\mathrm{OT}_c(\mathbb{P}_i, \hat{\mathbb{P}}) \leq r$ and

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_i}[\ell(Z)] &= (1 - q_i)\, \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + q_i\, \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z + r z^\star / q_i)] \\ &\geq (1 - q_i)\, \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + q_i\, \mathbb{E}_{\hat{\mathbb{P}}}[(y_i)^\top (Z + r z^\star / q_i) - \ell^*(y_i)]. \end{aligned}$$

Here the inequality follows from the representation of ℓ in terms of its conjugate ℓ^*. As i tends to infinity, q_i as well as $q_i \ell^*(y_i)$ converge to 0, and y_i converges to $y^\star$. Recall also that $(y^\star)^\top z^\star = \|y^\star\|_* = \mathrm{lip}(\ell)$. This shows that the supremum of the worst-case expectation problem (6.18a) is bounded below by $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r\, \mathrm{lip}(\ell)$.

Next, define $\lambda^\star = \mathrm{lip}(\ell)$, and note that

$$\ell(\hat{z}) \leq \sup_{z \in \mathcal{Z}} \ell(z) - \lambda^\star \|z - \hat{z}\| \leq \sup_{z \in \mathcal{Z}} \ell(\hat{z}) + \mathrm{lip}(\ell)\|z - \hat{z}\| - \lambda^\star \|z - \hat{z}\| = \ell(\hat{z})$$

for all $\hat{z} \in \mathcal{Z}$, where the second inequality follows from the Lipschitz continuity of ℓ, and the equality holds thanks to the definition of $\lambda^\star$. Thus the objective function value of $\lambda^\star$ in the dual problem (6.18b) is given by

$$\lambda^\star r + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z \in \mathcal{Z}} \ell(z) - \lambda \|z - \hat{Z}\|\right] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + r\, \mathrm{lip}(\ell).$$

In summary, we have shown that – asymptotically for large i – the objective function value of $\mathbb{P}_i$ in (6.18a) matches that of $\lambda^\star$ in (6.18b). By weak duality as established in Theorem 4.18, the supremum of the primal problem (6.18a) thus coincides with the Lipschitz-regularized nominal loss $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r\, \mathrm{lip}(\ell)$ and is asymptotically

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

attained by the distribution $\mathbb{P}_i$, which moves a fraction q_i of the total probability mass by a distance r/q_i along the direction $z^\star$. □

The connection between robustification and Lipschitz regularization was discovered by Mohajerin Esfahani and Kuhn (2018). It offers a probabilistic interpretation for regularization techniques commonly used in statistics and machine learning (Shafieezadeh-Abadeh *et al.* 2015, 2019). Further extensions to non-convex loss functions have been established by Blanchet *et al.* (2019*a*), Ho-Nguyen and Wright (2023), Shafiee, Aolaritei, Dörfler and Kuhn (2023), Gao *et al.* (2024*b*) and Zhang *et al.* (2024*a*).

6.14. 1-*Wasserstein risk*

Consider a law-invariant risk measure ϱ that can be expressed as a superposition of CVaRs with different risk levels $\beta \in [0, 1]$. Specifically, assume that

$$\varrho_{\mathbb{P}}[\ell(Z)] = \int_0^1 \beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)]\, \mathrm{d}\sigma(\beta) \tag{6.19}$$

for all $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$, where σ is a probability distribution on $[0, 1]$ with $\int_0^1 \beta^{-1}\mathrm{d}\sigma(\beta) < \infty$. Any ϱ with these properties is called a *spectral* risk measure (Acerbi 2002), and (6.19) is termed a Kusuoka representation of ϱ (Kusuoka 2001, Shapiro 2013).

If the distribution of Z is only known to be close to $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$, then it is natural to quantify the riskiness of an uncertain loss $\ell(Z)$ under a spectral risk measure ϱ by the 1-*Wasserstein risk*, that is, the supremum of $\varrho_{\mathbb{P}}[\ell(Z)]$ over all distributions $\mathbb{P}$ in a 1-Wasserstein ball around $\hat{\mathbb{P}}$. The 1-Wasserstein risk is available in closed form whenever $\mathcal{Z} = \mathbb{R}^d$ and ℓ is convex and Lipschitz-continuous.

Proposition 6.18 (1-Wasserstein risk). Let ϱ be a spectral risk measure satisfying (6.19) with $\int_0^1 \beta^{-1}\mathrm{d}\sigma(\beta) < \infty$. Assume that $\hat{\mathbb{P}} \in \mathcal{P}(\mathbb{R}^d)$ with $\mathbb{E}_{\hat{\mathbb{P}}}[\|Z\|] < \infty$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$. Define $\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)\colon \mathrm{W}_1(\mathbb{P}, \hat{\mathbb{P}}) \le r\}$, where $r \ge 0$ and W_1 is the 1-Wasserstein distance with transportation cost function $c(z, \hat{z}) = \|z - \hat{z}\|$. If ℓ is convex and Lipschitz-continuous with $\mathrm{lip}(\ell) < \infty$, then we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] = \varrho_{\hat{\mathbb{P}}}[\ell(Z)] + r\,\mathrm{lip}(\ell) \int_0^1 \beta^{-1}\mathrm{d}\sigma(\beta).$$

Proof. The assumption $\int_0^1 \beta^{-1}\mathrm{d}\sigma(\beta) < \infty$ ensures that $\sigma(\{0\}) = 0$, and the assumption $\mathbb{E}_{\hat{\mathbb{P}}}[\|Z\|] < \infty$ ensures via the Lipschitz continuity of ℓ that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]$ is finite. We first bound the worst-case risk from above. To this end, note that

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] &\le \int_0^1 \sup_{\mathbb{P}\in\mathcal{P}} \beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)]\, \mathrm{d}\sigma(\beta) \\ &\le \int_0^1 \inf_{\tau\in\mathbb{R}} \tau + \frac{1}{\beta} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\max\{\ell(Z) - \tau, 0\}]\, \mathrm{d}\sigma(\beta)\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$$
\begin{aligned}
&= \int_0^1 \inf_{\tau\in\mathbb{R}} \tau + \frac{1}{\beta}(\mathbb{E}_{\hat{\mathbb{P}}}[\max\{\ell(Z)-\tau, 0\}] + r\,\mathrm{lip}(\ell))\,\mathrm{d}\sigma(\beta) \\
&= \varrho_{\hat{\mathbb{P}}}[\ell(Z)] + r\,\mathrm{lip}(\ell)\int_0^1 \beta^{-1}\mathrm{d}\sigma(\beta) \\
&< +\infty,
\end{aligned}
$$

where the first inequality holds because $\mathbb{P}$ may adapt to β when the supremum is evaluated inside the integral, and the second inequality follows from the standard max-min inequality. The first equality follows from the results on worst-case expectations over 1-Wasserstein balls in Section 6.13.

To derive the converse inequality, we assume first that $\sigma(\{1\}) = 0$. The general case will be addressed later. Note that $\mu = \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)]$ is finite because ℓ is Lipschitz-continuous and because $\mathbb{E}_{\hat{\mathbb{P}}}[\|Z\|] < \infty$, which implies via the proof of Theorem 3.19 that all distributions in $\mathcal{P}$ have uniformly bounded first moment. We may assume without loss of generality that $\mu \geq 0$. Otherwise, we may replace $\ell(z)$ with $\ell(z) - \mu$, which simply increases the worst-case risk by $-\mu$ because any spectral risk measure is translation-invariant. The assumption that $\mu \geq 0$ then implies that

$$
\beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)] \geq \mathbb{E}_{\mathbb{P}}[\ell(Z)] \geq 0 \quad \text{for all } \beta \in [0,1],\ \mathbb{P}\in\mathcal{P}.
$$

Thus we have

$$
\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] &= \sup_{\mathbb{P}\in\mathcal{P}} \sup_{\delta>0} \int_\delta^{1-\delta} \beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)]\,\mathrm{d}\sigma(\beta) \\
&= \sup_{\delta>0} \sup_{\mathbb{P}\in\mathcal{P}} \int_\delta^{1-\delta} \beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)]\,\mathrm{d}\sigma(\beta),
\end{aligned}
$$

where the first equality follows from the monotone convergence theorem and the assumption that $\sigma(\{0\}) = \sigma(\{1\}) = 0$. Hence, for any $\varepsilon > 0$ there is $\delta > 0$ with

$$
\left| \sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] - \sup_{\mathbb{P}\in\mathcal{P}} \int_\delta^{1-\delta} \beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)]\,\mathrm{d}\sigma(\beta) \right| \leq \varepsilon \tag{6.20a}
$$

and

$$
\left| \int_0^1 \beta^{-1}\mathrm{d}\sigma(\beta) - \int_\delta^{1-\delta} \beta^{-1}\mathrm{d}\sigma(\beta) \right| \leq \varepsilon. \tag{6.20b}
$$

Recall now from Theorem 3.19 that $\mathcal{P}$ is weakly compact and thus tight. Hence there exists a compact set $\mathcal{C} \subseteq \mathbb{R}^d$ with $\mathbb{P}(Z \notin \mathcal{C}) \leq \delta/2$ for every $\mathbb{P} \in \mathcal{P}$. As $\mathcal{C}$ is compact, $\underline{\tau} = \min_{z\in\mathcal{C}} \ell(z)$ and $\overline{\tau} = \max_{z\in\mathcal{C}} \ell(z)$ are both finite. Using the trivial bounds $\mathbb{P}(\ell(Z) \geq \underline{\tau}) \geq \mathbb{P}(Z \in \mathcal{C})$ and $\mathbb{P}(\ell(Z) \leq \overline{\tau}) \geq \mathbb{P}(Z \in \mathcal{C})$ and noting that $\mathbb{P}(Z \in \mathcal{C}) \geq 1 - \delta/2$ for every $\mathbb{P} \in \mathcal{P}$, one can then readily show that

$$
\underline{\tau} \leq (1-\delta)\text{-VaR}_{\mathbb{P}}[\ell(Z)] \leq \beta\text{-VaR}_{\mathbb{P}}[\ell(Z)] \leq \delta\text{-VaR}_{\mathbb{P}}[\ell(Z)] \leq \overline{\tau}
$$

for all $\beta \in [\delta, 1-\delta]$ and for all $\mathbb{P} \in \mathcal{P}$. Next, define $y_i \in \mathrm{dom}(\ell^*)$, $q_i \in [0,1]$, the

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

function $f_i\colon \mathbb{R}^d \to \mathbb{R}^d$ and the distribution $\mathbb{P}_i = (1-q_i)\,\hat{\mathbb{P}} + q_i\,\hat{\mathbb{P}} \circ f_i^{-1}$ for $i \in \mathbb{N}$ as in Section 6.13. We then obtain

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] &\geq \int_{\delta}^{1-\delta} \inf_{\tau\in\mathbb{R}} \tau + \frac{1}{\beta}\mathbb{E}_{\mathbb{P}_i}[\max\{\ell(Z)-\tau,0\}]\,\mathrm{d}\sigma(\beta)\\ &= \int_{\delta}^{1-\delta} \inf_{\tau\in[\underline{\tau},\overline{\tau}]} \tau + \frac{1-q_i}{\beta}\mathbb{E}_{\hat{\mathbb{P}}}[\max\{\ell(Z)-\tau,0\}]\\ &\quad + \frac{q_i}{\beta}\mathbb{E}_{\hat{\mathbb{P}}}[\max\{\ell(Z+rz^\star/q_i)-\tau,0\}]\,\mathrm{d}\sigma(\beta). \qquad (6.21)\end{aligned}$$

The inequality in (6.21) holds because β-$\mathrm{CVaR}_{\mathbb{P}}[\ell(Z)] \geq 0$ for all $\beta \in [0,1]$ by assumption and because $\mathbb{P}_i \in \mathcal{P}$ as shown in Section 6.13. The equality follows from the definition of $\mathbb{P}_i$ and from Rockafellar and Uryasev (2002, Theorem 10), which ensures that the minimization problem over τ is solved by β-$\mathrm{VaR}_{\mathbb{P}}[\ell(Z)] \in [\underline{\tau},\overline{\tau}]$. As ℓ is proper, convex and lower semicontinuous, and as y_i belongs to the domain of ℓ^*, the Fenchel–Moreau theorem further implies that

$$\ell(z+rz^\star/q_i) = \sup_{y\in\mathrm{dom}(\ell^*)} (z+rz^\star/q_i)^\top y - \ell^*(y) \geq (z+rz^\star/q_i)^\top y_i - \ell^*(y_i).$$

The last expectation in (6.21) thus admits the lower bound

$$\begin{aligned}\mathbb{E}_{\hat{\mathbb{P}}}[\max\{\ell(Z+rz^\star/q_i)-\tau,0\}] &\geq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z+rz^\star/q_i)-\tau]\\ &\geq \mathbb{E}_{\hat{\mathbb{P}}}[y_i^\top Z] + ry_i^\top z^\star/q_i - \ell^*(y_i) - \overline{\tau}.\end{aligned}$$

Substituting this estimate into (6.21) and letting i tend to infinity yields

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] &\geq \lim_{i\to\infty} \int_{\delta}^{1-\delta} \inf_{\tau\in[\underline{\tau},\overline{\tau}]} \tau + \frac{1-q_i}{\beta}\mathbb{E}_{\hat{\mathbb{P}}}[\max\{\ell(Z)-\tau,0\}]\,\mathrm{d}\sigma(\beta)\\ &\quad + r\,\mathrm{lip}(\ell)\int_{\delta}^{1-\delta} \beta^{-1}\,\mathrm{d}\sigma(\beta)\\ &= \int_{\delta}^{1-\delta} \beta\text{-}\mathrm{CVaR}_{\mathbb{P}}[\ell(Z)]\,\mathrm{d}\sigma(\beta) + r\,\mathrm{lip}(\ell)\int_{\delta}^{1-\delta} \beta^{-1}\,\mathrm{d}\sigma(\beta),\end{aligned}$$

where we have used that q_i as well as $q_i\ell^*(y_i)$ converge to 0 and that $y_i^\top z^\star$ converges to $(y^\star)^\top z^\star = \mathrm{lip}(\ell)$ as i tends to infinity; see also Section 6.13. The equality follows from the monotone convergence theorem, which applies because q_i is monotonically decreasing with i. Letting ε tend to 0 thus implies via (6.20) that

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] \geq \int_0^1 \beta\text{-}\mathrm{CVaR}_{\mathbb{P}}[\ell(Z)]\,\mathrm{d}\sigma(\beta) + r\,\mathrm{lip}(\ell)\int_0^1 \beta^{-1}\,\mathrm{d}\sigma(\beta).$$

This lower bound matches the upper bound derived in the first part of the proof, and thus the claim follows, provided that $\sigma(\{1\}) = 0$. If the probability distribution

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

σ has an atom at 1, then it can be decomposed as $\sigma = \hat{\sigma} + \sigma(\{1\}) \cdot \delta_1$, where $\hat{\sigma}$ is a non-negative measure on $(0, 1)$. We can thus decompose the risk under $\mathbb{P}$ as

$$\varrho_{\mathbb{P}}[\ell(Z)] = \int_0^1 \beta\text{-CVaR}_{\mathbb{P}}[\ell(Z)]\, \mathrm{d}\hat{\sigma}(\beta) + \sigma(\{1\}) \cdot \mathbb{E}_{\mathbb{P}}[\ell(Z)].$$

The first term in this decomposition can then be handled as above, and the second term can be handled as in Section 6.13. Details are omitted for brevity. □

Proposition 6.18 shows that the 1-Wasserstein risk of a Lipschitz-continuous convex loss function coincides with the sum of the nominal risk and a Lipschitz regularization term. It is asymptotically attained by the distribution $\mathbb{P}_i$, which moves a fraction q_i of the total probability mass by a distance r/q_i along the direction $z^\star$. Proposition 6.17 emerges as a special case of Proposition 6.18 when $\sigma = \delta_1$. The worst-case risk over p-Wasserstein balls for $p \geq 1$ was first studied by Pflug *et al.* (2012), and a result akin to Proposition 6.18 was obtained for linear loss functions. Extensions to more general risk measures were studied by Pichler (2013) and Wozabal (2014). The extension to convex loss functions is new.

6.15. p-Wasserstein risk

We now show that if the loss function $\ell(z)$ is linear, then the worst-case risk over a p-Wasserstein ball may be available in closed form even if $p \in (1, \infty)$. The results of this section depend on the following lemma, which characterizes the conjugates of powers of norms; see also Zhen *et al.* (2023, Lemma C.9).

Lemma 6.19 (Conjugates of powers of norms). Assume that $\|\cdot\|$ and $\|\cdot\|_*$ are mutually dual norms on $\mathbb{R}^d$ and that $p, q \in (1, \infty)$ are conjugate exponents with $\frac{1}{p} + \frac{1}{q} = 1$. Define $\varphi(q) = (q-1)^{(q-1)}/q^q$. Then the following statements hold.

(i) If $f(z) = \frac{1}{p}\|z\|^p$, then $f^*(y) = \frac{1}{q}\|y\|_*^q$.

(ii) If $g(z) = \|z - \hat{z}\|^p$, then $g^*(y) = y^\top \hat{z} + \varphi(q)\|y\|_*^q$.

Proof. As for assertion (i), fix any $z, y \in \mathbb{R}^d$. We then have

$$z^\top y - \frac{1}{p}\|z\|^p \leq \|z\|\|y\|_* - \frac{1}{p}\|z\|^p \leq \max_{t \geq 0}\, t\|y\|_* - \frac{1}{p}t^p = \frac{1}{q}\|y\|_*^q,$$

where the first inequality follows from the construction of the dual norm, and the second inequality is obtained by maximizing over $t = \|z\|$. The equality holds because the maximization problem is solved by $\tau = \|y\|_*^{1/(p-1)}$. Both inequalities collapse to equalities if $z \in \arg\max_{\|z\|=\tau} z^\top y$. This allows us to conclude that

$$f^*(y) = \sup_{z \in \mathbb{R}^d} z^\top y - \frac{1}{p}\|z\|^p = \frac{1}{q}\|y\|_*^q.$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

As for assertion (ii), note that

$$\begin{aligned} g^*(y) &= \sup_{z\in\mathbb{R}^d} y^\top z - \|z-\hat{z}\|^p \\ &= y^\top \hat{z} + p \cdot \sup_{z\in\mathbb{R}^d} (y/p)^\top z - \frac{1}{p}\|z\|^p \\ &= y^\top \hat{z} + \frac{p}{q}\|y/p\|_*^q \\ &= y^\top \hat{z} + \varphi(q)\|y\|_*^q, \end{aligned}$$

where the last two equalities exploit assertion (i) and the definition of $\varphi(q)$. □

We now show that the worst-case CVaR of a linear loss function $\ell(z) = \theta^\top z$ over a p-Wasserstein ball of radius r around $\hat{\mathbb{P}}$ equals the sum of the nominal CVaR under $\hat{\mathbb{P}}$ and a regularization term that scales with the norm of θ and with r.

Proposition 6.20 (p-Wasserstein risk). Assume that $\hat{\mathbb{P}} \in \mathcal{P}(\mathbb{R}^d)$ with $\mathbb{E}_{\hat{\mathbb{P}}}[\|Z\|^p] < \infty$ for some $p \in (1,\infty)$ and for some norm $\|\cdot\|$ on $\mathbb{R}^d$. Define $\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)\colon \mathrm{W}_p(\mathbb{P},\hat{\mathbb{P}}) \le r\}$, where $r \ge 0$ and W_p is the p-Wasserstein distance with transportation cost function $c(z,\hat{z}) = \|z-\hat{z}\|^p$. If $\theta \in \mathbb{R}^d$ and $\beta \in (0,1)$, then

$$\sup_{\mathbb{P}\in\mathcal{P}} \beta\text{-CVaR}_{\mathbb{P}}[\theta^\top Z] = \beta\text{-CVaR}_{\hat{\mathbb{P}}}[\theta^\top Z] + r\beta^{-1/p}\|\theta\|_*.$$

Proof. By the definition of the CVaR by Rockafellar and Uryasev (2000), we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \beta\text{-CVaR}_{\mathbb{P}}[\theta^\top Z] \le \inf_{\tau\in\mathbb{R}} \tau + \frac{1}{\beta}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\max\{\theta^\top Z - \tau, 0\}], \tag{6.22}$$

where the inequality is obtained by interchanging the supremum over $\mathbb{P}$ and the infimum over τ. The underlying worst-case expectation problem satisfies

$$\begin{aligned} &\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\max\{\theta^\top Z - \tau, 0\}] \\ &\quad\le \inf_{\lambda\ge 0} \lambda r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathbb{R}^d} \max\{\theta^\top z - \tau, 0\} - \lambda\|z-\hat{Z}\|^p\right] \\ &\quad= \inf_{\lambda\ge 0} \lambda r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\max\left\{\sup_{z\in\mathbb{R}^d} \theta^\top z - \tau - \lambda\|z-\hat{Z}\|^p, \sup_{z\in\mathbb{R}^d} -\lambda\|z-\hat{Z}\|^p\right\}\right] \\ &\quad= \inf_{\lambda\ge 0} \lambda r^p + \mathbb{E}_{\hat{\mathbb{P}}}[\max\{\theta^\top \hat{Z} - \tau + \varphi(q)\lambda\|\theta/\lambda\|_*^q, 0\}], \end{aligned}$$

where the inequality exploits weak duality, and the first equality is obtained by interchanging the order of the two maximization operations. The second equality follows from Lemma 6.19(ii). Substituting the resulting formula into (6.22) and

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

interchanging the infimum over τ with the infimum over λ then yields

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}} \beta\text{-CVaR}_{\mathbb{P}}[\theta^\top Z] \\
&\quad\le \inf_{\lambda\ge 0} \frac{\lambda r^p}{\beta} + \inf_{\tau\in\mathbb{R}} \tau + \frac{1}{\beta}\mathbb{E}_{\hat{\mathbb{P}}}[\max\{\theta^\top \hat{Z} - \tau + \varphi(q)\lambda\|\theta/\lambda\|_*^q, 0\}] \\
&\quad= \inf_{\lambda\ge 0} \frac{\lambda r^p}{\beta} + \beta\text{-CVaR}_{\hat{\mathbb{P}}}[\theta^\top \hat{Z} + \varphi(q)\lambda\|\theta/\lambda\|_*^q] \\
&\quad= \beta\text{-CVaR}_{\hat{\mathbb{P}}}[\theta^\top \hat{Z}] + \inf_{\lambda\ge 0} \frac{\lambda r^p}{\beta} + \varphi(q)\lambda\|\theta/\lambda\|_*^q,
\end{aligned}$$

where the equalities follow from the definition and the translation-invariance of the CVaR, respectively. Solving the minimization problem over λ analytically yields

$$\sup_{\mathbb{P}\in\mathcal{P}} \beta\text{-CVaR}_{\mathbb{P}}[\theta^\top Z] \le \beta\text{-CVaR}_{\hat{\mathbb{P}}}[\theta^\top \hat{Z}] + r\beta^{-1/p}\|\theta\|_*.$$

To derive the converse inequality, we use τ_β as shorthand for $\beta\text{-VaR}_{\hat{\mathbb{P}}}[\theta^\top \hat{Z}]$, which is finite because $\beta \in (0,1)$, and we select any $z^\star \in \arg\max_{\|z\|=1} \theta^\top z$. In addition, we decompose the nominal distribution as $\hat{\mathbb{P}} = \beta\,\hat{\mathbb{P}}_+ + (1-\beta)\,\hat{\mathbb{P}}_-$, where $\hat{\mathbb{P}}_+$ and $\hat{\mathbb{P}}_-$ are probability distributions supported on $\mathcal{Z}_+ = \{z \in \mathbb{R}^d : \theta^\top z \ge \tau_\beta\}$ and $\mathcal{Z}_- = \{z \in \mathbb{R}^d : \theta^\top z \le \tau_\beta\}$, respectively. Such a decomposition always exists thanks to the definition of τ_β. For example, if $\hat{\mathbb{P}}(\theta^\top Z = \tau_\beta) = 0$, as would be the case if $\hat{\mathbb{P}}$ were absolutely continuous with respect to Lebesgue measure, then $\hat{\mathbb{P}}_-$ and $\hat{\mathbb{P}}_+$ can simply be obtained by conditioning $\hat{\mathbb{P}}$ on $\mathcal{Z}_-$ and $\mathcal{Z}_+$, respectively. We also define $f : \mathbb{R}^d \to \mathbb{R}^d$ through $f(z) = z + rz^\star/\beta^{1/p}$. Thus f shifts all points in $\mathbb{R}^d$ along the direction $z^\star$ by a distance equal to $r/\beta^{1/p}$. Finally, we set $\mathbb{P}^\star = \beta\,\hat{\mathbb{P}}_+ \circ f^{-1} + (1-\beta)\,\hat{\mathbb{P}}_-$. Hence $\mathbb{P}^\star$ is obtained by decomposing $\hat{\mathbb{P}}$ into two parts $\beta\,\hat{\mathbb{P}}_+$ and $(1-\beta)\,\hat{\mathbb{P}}_-$ and then translating the first part by $rz^\star/\beta^{1/p}$. We thus have $\mathrm{W}_p(\mathbb{P}^\star, \hat{\mathbb{P}}) \le r$, and $\beta\text{-VaR}_{\mathbb{P}}[\theta^\top Z] = \tau_\beta$. This in turn implies that

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}} \beta\text{-CVaR}_{\mathbb{P}}[\theta^\top Z] \\
&\quad\ge \beta\text{-CVaR}_{\mathbb{P}^\star}[\theta^\top Z] \\
&\quad= \tau_\beta + \frac{1}{\beta}\mathbb{E}_{\mathbb{P}^\star}[\max\{\theta^\top Z - \tau_\beta, 0\}] \\
&\quad= \tau_\beta + \mathbb{E}_{\hat{\mathbb{P}}_+}[\max\{\theta^\top f(Z) - \tau_\beta, 0\}] + \frac{1-\beta}{\beta}\mathbb{E}_{\hat{\mathbb{P}}_-}[\max\{\theta^\top Z - \tau_\beta, 0\}] \\
&\quad= \mathbb{E}_{\hat{\mathbb{P}}_+}[\theta^\top Z] + r\beta^{-1/p}\|\theta\|_* \\
&\quad= \beta\text{-CVaR}_{\hat{\mathbb{P}}}[\theta^\top \hat{Z}] + r\beta^{-1/p}\|\theta\|_*.
\end{aligned}$$

Here the first equality follows from the definition of the CVaR and from Rockafellar and Uryasev (2002, Theorem 10), which ensures τ matches $\beta\text{-VaR}_{\mathbb{P}^\star}[\ell(Z)] = \tau_\beta$ at optimality. The second equality exploits the definition of $\mathbb{P}^\star$, and the third equality

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

holds because $\theta^\top z^\star = \|\theta\|_*$ and because $\theta^\top z \geq \tau_\beta$ for all $z \in \mathcal{Z}_+$ and $\theta^\top z \leq \tau_\beta$ for all $z \in \mathcal{Z}_-$. Finally, the fourth equality follows from the construction of $\hat{\mathbb{P}}_+$ and from Rockafellar and Uryasev (2002, Proposition 5). This completes the proof. □

7. Finite convex reformulations of nature's subproblem

Although nature's subproblem admits analytical solutions in important special cases (see Section 6), it can usually only be solved numerically. Sometimes, nature's subproblem can be reformulated as an equivalent convex optimization problem. In these cases, it can be addressed with off-the-shelf solvers. In other cases, however, it may be necessary or preferable to develop customized solution algorithms.

This section focuses on finite convex reductions. That is, we will describe conditions under which the dual worst-case expectation problems derived in Section 4 can be reformulated as finite convex *minimization* problems. These finite reformulations are significant because they can be combined with the outer minimization problem over $x \in \mathcal{X}$ to construct a reformulation of the overall DRO problem (1.2) as a classical minimization problem amenable to standard optimization software. We subsequently dualize the finite convex reformulations of nature's subproblem to obtain equivalent finite convex *maximization* problems. These finite bi-dual maximization problems are significant because their optimal solutions allow us to construct worst-case distributions that (asymptotically) attain the supremum of nature's subproblem (4.1). Even though we only address worst-case expectations, all results of this section readily extend to worst-case optimized certainty equivalents thanks to Theorem 5.18. For the sake of brevity, however, we will not elaborate on these extensions. To simplify notation, we will always suppress the dependence of the loss function ℓ on the decision variables x.

The remainder of this section develops as follows. In Section 7.1, we first outline a general strategy for deriving finite convex dual and bi-dual reformulations of nature's subproblem (4.1). We subsequently exemplify this strategy for worst-case expectation problems over Chebyshev ambiguity sets (Section 7.2), ϕ-divergence ambiguity sets (Section 7.3) and optimal transport ambiguity sets (Section 7.4).

7.1. *General proof strategy*

The worst-case expectation problem (4.1) constitutes a semi-infinite program that involves infinitely many decision variables (because it optimizes over a subset of an infinite-dimensional measure space) but only finitely many constraints (e.g. moment conditions and/or bounds on the divergence or discrepancy to a reference distribution). The duality results of Section 4 enable us to recast this semi-infinite maximization problem as a semi-infinite minimization problem with finitely many variables and infinitely many constraints. We then leverage reformulation techniques from robust optimization to recast the dual semi-infinite program as a finite-dimensional convex minimization problem. These techniques exploit

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

standard results from convex analysis as well as the $\mathcal{S}$-Lemma, which we review next. Throughout this discussion we adopt the convention that $0 \cdot \infty = \infty$.

We first show that scaling and perspectivication constitute dual operations.

Lemma 7.1 (Duality of scaling and perspectivication). If $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is a proper, closed and convex function and $\alpha \in \mathbb{R}_+$ a fixed constant, then the following hold.

(i) If $g(z) = \alpha f(z)$, then $g^*(y) = (f^*)^\pi(y, \alpha)$ for all $y \in \mathbb{R}^d$.

(ii) If $g(z) = f^\pi(z, \alpha)$, then $g^*(y) = \mathrm{cl}(\alpha f^*)(y)$ for all $y \in \mathbb{R}^d$.

Proof. We prove assertion (i) by case distinction. First, if $\alpha > 0$, then we have

$$\begin{aligned} g^*(y) &= \sup_{z \in \mathbb{R}^d} y^\top z - \alpha f(z) \\ &= \alpha \sup_{z \in \mathbb{R}^d} (y/\alpha)^\top z - f(z) \\ &= \alpha f^*(y/\alpha) \\ &= (f^*)^\pi(y, \alpha). \end{aligned}$$

If $\alpha = 0$, on the other hand, then similar reasoning shows that

$$\begin{aligned} g^*(y) &= \sup_{z \in \mathbb{R}^d} y^\top z - \delta_{\mathrm{dom}(f)}(z) \\ &= \delta^*_{\mathrm{dom}(f)}(y) \\ &= \delta^*_{\mathrm{dom}(f^{**})}(y) \\ &= (f^*)^\infty(y) \\ &= (f^*)^\pi(y, \alpha), \end{aligned}$$

where the first equality follows from our convention that $0 \cdot \infty = \infty$, which implies that $0 f(z) = \delta_{\mathrm{dom}(f)}(z)$. The second equality follows from the definition of the support function, and the third equality holds because f is convex and closed, which implies via Lemma 4.2 that $f = f^{**}$. Finally, the fourth equality follows from Rockafellar (1970, Theorem 13.3), and the last equality exploits the definition of the perspective function for $\alpha = 0$. This completes the proof of assertion (i).

As for assertion (ii), assume first that $\alpha > 0$, and note that

$$g^*(y) = \sup_{z \in \mathbb{R}^d} y^\top z - f^\pi(z, \alpha) = \alpha \sup_{z \in \mathbb{R}^d} y^\top (z/\alpha) - f(z/\alpha) = \alpha f^*(y) = \mathrm{cl}(\alpha f^*)(y),$$

where the last equality holds because f^* is closed. If $\alpha = 0$, then we have

$$g^*(y) = \sup_{z \in \mathbb{R}^d} y^\top z - f^\infty(z) = \sup_{z \in \mathbb{R}^d} y^\top (z) - \delta^*_{\mathrm{dom}(f^*)} = \delta_{\mathrm{cl}(\mathrm{dom}(f^*))}(y) = \mathrm{cl}(\alpha f^*)(y).$$

Here the first equality exploits the definition of the perspective. The second and third equalities follow from Rockafellar (1970, Theorem 13.3) and Rockafellar (1970, Theorem 13.2), respectively. The last equality, finally, holds because $0 f^* = \delta_{\mathrm{dom}(f^*)}$ by our conventions of extended arithmetic. This proves assertion (ii). □

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The following lemma derives a formula for the conjugate of a sum of functions.

Lemma 7.2 (Conjugates of sums). If $f_k : \mathbb{R}^d \to \overline{\mathbb{R}}$, $k \in [K]$, are proper, convex and closed functions, then the conjugate of $f = \sum_{k\in[K]} f_k$ satisfies

$$f^*(y) \leq \inf_{y_1,\ldots,y_K\in\mathbb{R}^d} \left\{ \sum_{k\in[K]} f_k^*(y_k) \colon \sum_{k\in[K]} y_k = y \right\} \quad \text{for all } y \in \mathbb{R}^d. \tag{7.1}$$

If there exists $\bar{z} \in \cap_{k\in[K]} \operatorname{rint}(\operatorname{dom}(f_k))$, then the inequality in the above expression reduces to an equality, and the minimum is attained for every $y \in \mathbb{R}^d$.

The infimum on the right-hand side of (7.1) defines a function of y. This function is called the *infimal convolution* of the functions f_k^*, $k \in [K]$. Thus Lemma 7.2 asserts that, under a mild Slater-type condition, the conjugate of a sum of functions coincides with the infimal convolution of the conjugates of these functions.

Proof of Lemma 7.2. By using a standard variable splitting trick and the max-min inequality, one can show that the conjugate of f admits the following upper bound:

$$\begin{aligned}
f^*(y) &= \sup_{z,z_1,\ldots,z_K\in\mathbb{R}^d} \left\{ y^\top z - \sum_{k\in[K]} f(z_k) \colon z_k = z \ \ \forall k \in [K] \right\} \\
&= \sup_{z,z_1,\ldots,z_K\in\mathbb{R}^d} \inf_{y_1,\ldots,y_K\in\mathbb{R}^d} y^\top z - \sum_{k\in[K]} f(z_k) - y_k^\top (z - z_k) \\
&\leq \inf_{y_1,\ldots,y_K\in\mathbb{R}^d} \sup_{z,z_1,\ldots,z_K\in\mathbb{R}^d} y^\top z - \sum_{k\in[K]} f(z_k) - y_k^\top (z - z_k) \\
&= \inf_{y_1,\ldots,y_K\in\mathbb{R}^d} \sup_{z\in\mathbb{R}^d} \left\{ y^\top z - \sum_{k\in[K]} y_k^\top z \right\} + \sum_{k\in[K]} \sup_{z_k\in\mathbb{R}^d} \left\{ y_k^\top z_k - f(z_k) \right\}.
\end{aligned}$$

The supremum over z in the resulting expression evaluates to 0 if $\sum_{k\in[K]} y_k = y$ and to ∞ otherwise. In addition, the supremum over z_k evaluates to $f_k^*(y_k)$ for every $k \in [K]$. Substituting these analytical formulas into the last expression yields

$$f^*(y) \leq \inf_{y_1,\ldots,y_K\in\mathbb{R}^d} \left\{ \sum_{k\in[K]} f^*(y_k) \colon \sum_{k\in[K]} y_k = y \right\}.$$

If $\cap_{k\in[K]} \operatorname{rint}(\operatorname{dom}(f_k))$ is non-empty, then the above inequality becomes an equality, and the infimum is attained thanks to Rockafellar (1970, Theorem 16.4). □

Now consider a classical optimization problem

$$\inf_{z\in\mathbb{R}^d} \{ f(z) \colon g_k(z) \leq 0 \ \ \forall k \in [K] \} \tag{P}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

with objective function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ and constraint functions $g_k\colon \mathbb{R}^d \to \overline{\mathbb{R}}$, $k \in [K]$. Below we will show that the problem dual to (P) is given by

$$\sup_{\substack{\alpha_1,\ldots,\alpha_K \in \mathbb{R}_+ \\ \beta_0,\ldots,\beta_K \in \mathbb{R}^d}} \left\{ -f^*(\beta_0) - \sum_{k=1}^K (g_k^*)^\pi(\beta_k, \alpha_k)\colon \sum_{k=0}^K \beta_k = 0 \right\}. \tag{D}$$

To this end, we adopt the following definition of a Slater point.

Definition 7.3 (Slater point). A Slater point of the set

$$\mathcal{Z} = \{z \in \mathbb{R}^d\colon g_k(z) \leq 0 \ \ \forall k \in [K]\}$$

is any vector $\bar{z} \in \mathcal{Z}$ with $\bar{z} \in \operatorname{rint}(\operatorname{dom}(g_k))$ for all $k \in [K]$ and $g_k(\bar{z}) < 0$ for all $k \in [K]$ such that g_k is nonlinear. A Slater point $\bar{z}$ of the set $\mathcal{Z}$ is a Slater point of the minimization problem $\inf\{f(z)\colon z \in \mathcal{Z}\}$ if $\bar{z} \in \operatorname{rint}(\operatorname{dom}(f))$.

Slater points of maximization problems are defined in the obvious way. We simply replace the requirement $\bar{z} \in \operatorname{rint}(\operatorname{dom}(f))$ with $\bar{z} \in \operatorname{rint}(\operatorname{dom}(-f))$. Using Lemmas 7.1 and 7.2, we can now prove that (P) and (D) are indeed duals.

Theorem 7.4 (Convex duality). Assume that the functions f and g_k, $k \in [K]$, are proper, closed and convex. Then the infimum of (P) is larger than or equal to the supremum of (D). In addition, the following strong duality relations hold.

(i) If (P) or (D) admits a Slater point, then the infimum of (P) matches the supremum of (D), and (D) or (P) is solvable, respectively.

(ii) If the feasible set of (P) or (D) is non-empty and bounded, then the infimum of (P) matches the supremum of (D), and (P) or (D) is solvable, respectively.

Proof. The max-min inequality readily implies that the infimum of (P) is bounded below by the optimal value of its Lagrangian dual, that is, we have

$$\begin{aligned}
\inf(\mathrm{P}) &= \inf_{z\in\mathbb{R}^d} \sup_{\alpha\in\mathbb{R}_+^K} f(z) + \sum_{k\in[K]} \alpha_k g_k(z) \\
&\geq \sup_{\alpha\in\mathbb{R}_+^K} \inf_{z\in\mathbb{R}^d} f(z) + \sum_{k\in[K]} \alpha_k g_k(z) \\
&= \sup_{\alpha\in\mathbb{R}_+^K} -\sup_{z\in\mathbb{R}^d} 0^\top z - f(z) - \sum_{k\in[K]} \alpha_k g_k(z) \\
&= \sup_{\alpha\in\mathbb{R}_+^K} -\left(f + \sum_{k\in[K]} \alpha_k g_k\right)^*(0).
\end{aligned}$$

The resulting lower bound involves the conjugate of a sum of several functions. By Lemma 7.2, the conjugate of this sum is bounded below by the infimal convolution

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

of the conjugates of all functions in the sum. Consequently, we obtain

$$\inf(\mathrm{P}) \geq \sup_{\substack{\alpha_1,\dots,\alpha_K \in \mathbb{R}_+ \\ \beta_0,\dots,\beta_K \in \mathbb{R}^d}} \left\{ -f^*(\beta_0) - \sum_{k=1}^K (\alpha_k g_k)^*(\beta_k) \colon \sum_{k=0}^K \beta_k = 0 \right\}. \tag{7.2}$$

By Lemma 7.1 (i), we further have $(\alpha_k g_k)^*(\beta_k) = (g_k^*)^\pi(\beta_k, \alpha_k)$ for all $\beta_k \in \mathbb{R}^d$ and $\alpha_k \in \mathbb{R}_+$. Thus the lower bound in (7.2) matches the supremum of (D). This proves weak duality. For a proof of strong duality and solvability under the conditions (i) and (ii), we refer to Zhen *et al.* (2023, Theorem 2). □

Armed with Theorem 7.4, we can now show that the semi-infinite constraints appearing in the dual worst-case expectation problems derived in Section 4 can be systematically reformulated in terms of finitely many convex constraints.

Proposition 7.5 (Semi-infinite constraints I). Assume that the functions $f \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ and $g_k \colon \mathbb{R}^d \to \overline{\mathbb{R}}$, $k \in [K]$, are proper, closed and convex, and that there is $\bar{z} \in \mathbb{R}^d$ with $\bar{z} \in \operatorname{rint}(\operatorname{dom}(g_k))$, $k \in [K]$, $\bar{z} \in \operatorname{rint}(\operatorname{dom}(f))$ and $g_k(\bar{z}) < 0$ for all $k \in [K]$ such that g_k is nonlinear. Then the semi-infinite constraint

$$f(z) \geq 0 \quad \forall z \in \mathbb{R}^d \colon g_k(z) \leq 0 \ \ \forall k \in [K]$$

holds if and only if there exist $\alpha_1, \dots, \alpha_K \in \mathbb{R}_+$ and $\beta_0, \dots, \beta_K \in \mathbb{R}^d$ with

$$f^*(\beta_0) + \sum_{k=1}^K (g_k^*)^\pi(\beta_k, \alpha_k) \leq 0 \quad \text{and} \quad \sum_{k=0}^K \beta_k = 0.$$

Proof. The semi-infinite constraint in the statement of the proposition is satisfied if and only if the infimum of (P) is non-negative. Under the stated assumptions, Theorem 7.4 implies that this is the case precisely when the supremum of (D) is non-negative. Since (P) admits a Slater point, the supremum of (D) is attained. Thus the supremum of (D) is non-negative if and only if there are $\alpha_1, \dots, \alpha_K \in \mathbb{R}_+$ and $\beta_0, \dots, \beta_K \in \mathbb{R}^d$ satisfying the constraints in the statement of the proposition. □

Proposition 7.5 enables us to derive finite convex reformulations of the semi-infinite constraints that appear in the dual of the worst-case expectation problem (4.1) whenever the relevant objective and constraint functions are convex in z.

Another similar reformulation technique relies on the $\mathcal{S}$-lemma (see e.g. Pólik and Terlaky 2007), which we present without a proof.

Lemma 7.6 ($\mathcal{S}$-lemma (Yakubovich 1971)). Assume that $f \colon \mathbb{R}^d \to \mathbb{R}$ and $g \colon \mathbb{R}^d \to \mathbb{R}$ are quadratic functions. If there exists a Slater point $\bar{z} \in \mathbb{R}^d$ such that $g(\bar{z}) < 0$, then the following two statements are equivalent.

(i) There is no $z \in \mathbb{R}^d$ such that $f(z) < 0$ and $g(z) \leq 0$.

(ii) There exists $\alpha \in \mathbb{R}_+$ such that $f(z) + \alpha g(z) \geq 0$ for all $z \in \mathbb{R}^d$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The $\mathcal{S}$-lemma allows us to derive a finite convex reformulations of semi-infinite constraints that require a (possibly indefinite) quadratic function to be non-negative over the feasible set of a single quadratic constraint. Note in particular that the involved functions f and g are *not* required to be convex in z.

Proposition 7.7 (Semi-infinite constraints II). Assume that $Q_0, Q_1 \in \mathbb{S}^d$, $q_0, q_1 \in \mathbb{R}^d$, and $r_0, r_1 \in \mathbb{R}$. In addition, assume that there exists a Slater point $\bar{z} \in \mathbb{R}^d$ such that $\bar{z}^\top Q_0 \bar{z} + 2q_0^\top \bar{z} + r_0 < 0$. Then the semi-infinite constraint

$$z^\top Q_1 z + 2q_1^\top z + r_1 \geq 0 \quad \forall z \in \mathbb{R}^d : \; z^\top Q_0 z + 2q_0^\top z + r_0 \leq 0$$

holds if and only if there exists $\alpha \in \mathbb{R}_+$ with

$$\begin{bmatrix} Q_1 + \alpha Q_0 & q_1 + \alpha q_0 \\ q_1^\top + \alpha q_0^\top & r_1 + \alpha r_0 \end{bmatrix} \succeq 0.$$

Proof. We observe that

$$\begin{aligned} & z^\top Q_1 z + 2q_1^\top z + r_1 \geq 0 \quad \forall z \in \mathbb{R}^d : \; z^\top Q_0 z + 2q_0^\top z + r_0 \leq 0 \\ \iff & \exists \alpha \in \mathbb{R}_+ \text{ with } z^\top (Q_1 + \alpha Q_0)\, z + 2(q_1 + \alpha q_0)^\top z + r_1 + \alpha r_0 \geq 0 \quad \forall z \in \mathbb{R}^d \\ \iff & \exists \alpha \in \mathbb{R}_+ \text{ with } \begin{bmatrix} z \\ 1 \end{bmatrix}^\top \begin{bmatrix} Q_1 + \alpha Q_0 & q_1 + \alpha q_0 \\ q_1^\top + \alpha q_0^\top & r_1 + \alpha r_0 \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} \geq 0 \quad \forall z \in \mathbb{R}^d, \end{aligned}$$

where the first equivalence applies Lemma 7.6 to $f(z) = z^\top Q_1 z + 2q_1^\top z + r_1$ and $g(z) = z^\top Q_0 z + 2q_0^\top z + r_0$. As quadratic forms are homogeneous of degree 2 as well as continuous, the last statement is equivalent to the desired positive semi-definiteness condition. This observation concludes the proof. □

Proposition 7.7 is particularly useful for deriving finite convex reformulations of the dual worst-case expectation problems over Chebyshev or Gelbrich ambiguity sets; see (4.5) and (4.6). As we will see, the corresponding semi-infinite constraints fail to be convex in z, which implies that Proposition 7.5 is not applicable.

Finite convex reformulations of the dual worst-case expectation problem (4.1) are key to solving the DRO problem (1.2). They allow us to combine the outer minimization over $x \in \mathcal{X}$ with the inner minimization over the auxiliary decision variables of the dual worst-case expectation problem to obtain a finite convex reformulation of (1.2). However, the finite dual reformulations of (4.1) do not allow us to readily identify worst-case distributions that (asymptotically) attain the supremum of (4.1). Such worst-case distributions enable decision-makers to evaluate how a given candidate decision performs under the most challenging conditions, which is the essence of stress testing and contamination experiments; see e.g. Dupačová (2006). They also play a pivotal role in optimal uncertainty quantification, where they are used to determine the sharpest possible probabilistic bounds on quantities of interest, given limited information about the underlying probability distributions. We direct the readers to Owhadi *et al.* (2013) and Ghanem, Higdon and Owhadi (2017) for more details.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

To identify a worst-case distribution that attains the supremum of (4.1), or to identify a sequence of distributions that attain this supremum asymptotically, we consider the *bi-dual* reformulation of the worst-case expectation problem (4.1) that results from dualizing the finite convex dual of (4.1). The bi-dual can often be interpreted as a restriction of the worst-case expectation problem (4.1) to a subset of distributions $\mathbb{P} \in \mathcal{P}$ that are parametrized by finitely many decision variables. Strong duality between problem (4.1), its dual and its bi-dual then allows us to conclude that any optimal solution to this bi-dual problem represents a (sequence of) distribution(s) that attains the supremum of (4.1) (asymptotically).

The idea of extracting worst-case distributions from the finite bi-dual of problem (4.1) was formalized by Delage and Ye (2010, § 4.2) for Chebyshev ambiguity sets and later extended to optimal transport ambiguity sets by Mohajerin Esfahani and Kuhn (2018). In Section 7.2 we will see that, for the Chebyshev ambiguity set (2.4) with uncertain moments, the worst-case distributions constitute mixtures of distributions with first and second moments that are determined by the optimal solution of the finite bi-dual problem. For ϕ-divergence ambiguity sets centred at a discrete distribution $\hat{\mathbb{P}}$, Section 7.3 will show that the worst-case distributions are supported on the atoms of $\hat{\mathbb{P}}$ and (if ϕ grows at most linearly) on $\arg\max_{z \in \mathcal{Z}} \ell(z)$ with probability weights determined by the optimal solution to the finite bi-dual problem. Similarly, for the optimal transport ambiguity set (2.27) centred at a discrete distribution $\hat{\mathbb{P}}$, Section 7.4 will show that the worst-case distributions constitute mixtures of discrete distributions, with the locations and probability weights of their atoms determined by the optimal solution to the finite bi-dual problem.

7.2. Chebyshev ambiguity sets with uncertain moments

Recall that the Chebyshev ambiguity set (2.4) with uncertain moments is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}_2(\mathcal{Z}) \colon \mathbb{E}_{\mathbb{P}}[Z] = \mu, \;\; \mathbb{E}_{\mathbb{P}}[ZZ^\top] = M \;\; \forall (\mu, M) \in \mathcal{F}\},$$

where $\mathcal{F} \subseteq \mathbb{R}^d \times \mathbb{S}_+^d$ represents a closed moment uncertainty set and $\mathcal{P}_2(\mathcal{Z})$ stands for the family of all probability distributions on $\mathcal{Z}$ with finite second moments. This section combines the duality result for Chebyshev ambiguity sets (see Theorem 4.6) with the finite dual reformulation of the ensuing semi-infinite program (see Proposition 7.7) to derive an equivalent reformulation of nature's subproblem (4.1) as a finite-dimensional minimization problem. We also show how the corresponding bi-dual allows us to extract worst-case distributions $\mathbb{P}^\star \in \mathcal{P}$ that attain the optimal value of (4.1). Since the support-only ambiguity sets (see Section 2.1.1), the Markov ambiguity sets (see Section 2.1.2), the Chebyshev ambiguity sets with known moments (see Section 2.1.3) and the mean-dispersion ambiguity sets (see Section 2.1.5) can all be viewed as special instances of the Chebyshev ambiguity set with uncertain moments, our results immediately extend to those ambiguity sets as well, and we do not re-derive the corresponding statements for the sake of brevity. Due to its recent applications in statistics (Nguyen, Kuhn and Mohajerin Esfahani

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

2022), signal processing (Nguyen *et al.* 2023*b*) and control (Taşkesen *et al.* 2024), however, we report the finite dual and bi-dual reformulations of the Gelbrich ambiguity set with moment uncertainty set (2.8). All reformulations derived in this section leverage Lemma 7.6. Thus they require quadratic representations of the loss function ℓ and the support set $\mathcal{Z}$, as detailed in the following assumption.

Assumption 7.8 (Regularity conditions for Chebyshev ambiguity sets).

(i) The loss function ℓ is a pointwise maximum of quadratic functions,

$$\ell(z) = \max_{j\in[J]} \ell_j(z) \quad \text{with} \quad \ell_j(z) = z^\top Q_j z + 2q_j^\top z + q_j^0, \tag{7.3}$$

where $J \in \mathbb{N}$, $Q_j \in \mathbb{S}^d$, $q_j \in \mathbb{R}^d$, and $q_j^0 \in \mathbb{R}$ for all $j \in [J]$.

(ii) The support set $\mathcal{Z}$ is an ellipsoid of the form

$$\mathcal{Z} = \{z \in \mathbb{R}^d : (z - z_0)^\top Q_0 (z - z_0) \le 1\}, \tag{7.4}$$

where $Q_0 \in \mathbb{S}_+^d$ and $z_0 \in \mathbb{R}^d$.

Note that Assumption 7.8 does *not* impose any convexity conditions on the quadratic component functions $z^\top Q_j z + 2q_j^\top z + q_j^0$ that make up the loss function ℓ.

Theorem 7.9 (Finite dual reformulation for Chebyshev ambiguity sets). If $\mathcal{P}$ is the Chebyshev ambiguity set (2.4) with any $\mathcal{F} \subseteq \mathbb{R}^d \times \mathbb{S}_+^d$ and Assumption 7.8 holds, then the worst-case expectation problem (4.1) satisfies the weak duality relation

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \le \begin{cases} \inf & \lambda_0 + \delta^*_{\mathcal{F}}(\lambda, \Lambda) \\ \text{s.t.} & \lambda_0 \in \mathbb{R},\ \lambda \in \mathbb{R}^d,\ \Lambda \in \mathbb{S}^d,\ \alpha \in \mathbb{R}^J_+ \\ & \begin{bmatrix} \Lambda - Q_j + \alpha_j Q_0 & \frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0 \\ \left(\frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0\right)^\top & \lambda_0 - q_j^0 + \alpha_j (z_0^\top Q_0 z_0 - 1) \end{bmatrix} \succeq 0 \end{cases}$$

for all $j \in [J]$. If $\mathcal{F}$ is a convex and compact set with $M \succ \mu\mu^\top$ for all $(\mu, M) \in \operatorname{rint}(\mathcal{F})$, then strong duality holds, that is, the above inequality becomes an equality.

Proof. Weak duality follows from Theorem 4.6 and from the following equivalent reformulation of the semi-infinite constraint in the dual problem (4.5):

$$\begin{aligned} & \lambda_0 + \lambda^\top z + z^\top \Lambda z \ge \ell(z) \quad \forall z \in \mathcal{Z} \\ \iff\ & \lambda_0 + \lambda^\top z + z^\top \Lambda z \ge z^\top Q_j z + 2q_j^\top z + q_j^0 \quad \forall z \in \mathcal{Z}, \forall j \in [J] \\ \iff\ & \exists \alpha \in \mathbb{R}^J_+ \text{ with} \\ & \begin{bmatrix} \Lambda - Q_j + \alpha_j Q_0 & \frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0 \\ \left(\frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0\right)^\top & \lambda_0 - q_j^0 + \alpha_j (z_0^\top Q_0 z_0 - 1) \end{bmatrix} \succeq 0 \ \ \forall j \in [J] \end{aligned}$$

Here the first equivalence holds thanks to Assumption 7.8 (i), and the second equivalence follows from Proposition 7.7, which applies because $z_0 \in \operatorname{rint}(\mathcal{Z})$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

constitutes a Slater point thanks to Assumption 7.8 (ii). In addition, as the loss function is quadratic, strong duality follows readily from Theorem 4.6. □

Recall next that the Gelbrich ambiguity set (2.8) is defined in as an instance of the Chebyshev ambiguity set (2.4) with moment uncertainty set

$$\mathcal{F} = \left\{ (\mu, M) \in \mathbb{R}^d \times \mathbb{S}_+^d : \begin{array}{l} \exists \Sigma \in \mathbb{S}_+^d \text{ with } M = \Sigma + \mu\mu^\top, \\ \mathrm{G}((\mu, \Sigma), (\hat{\mu}, \hat{\Sigma})) \leq r \end{array} \right\}.$$

Here $(\hat{\mu}, \hat{\Sigma})$ is a nominal mean–covariance pair, and $r \geq 0$ is a size parameter. The next result follows directly from Theorems 4.9 and 7.9. We thus omit its proof.

Theorem 7.10 (Finite dual reformulation for Gelbrich ambiguity sets). If $\mathcal{P}$ is the Chebyshev ambiguity set (2.4) with $\mathcal{F}$ given by (2.8) and Assumption 7.8 holds, then the worst-case expectation problem (4.1) satisfies the weak duality relation

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \left\{ \begin{array}{ll} \inf & \lambda_0 + \gamma(r^2 - \|\hat{\mu}\|^2 - \mathrm{Tr}(\hat{\Sigma})) + \mathrm{Tr}(A_0) + \alpha_0 \\ \text{s.t.} & \lambda_0 \in \mathbb{R},\ \alpha_0, \gamma \in \mathbb{R}_+,\ \alpha \in \mathbb{R}_+^J,\ \lambda \in \mathbb{R}^d,\ \Lambda \in \mathbb{S}^d,\ A_0 \in \mathbb{S}_+^d \\ & \begin{bmatrix} \Lambda - Q_j + \alpha_j Q_0 & \frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0 \\ \left(\frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0\right)^\top & \lambda_0 - q_j^0 + \alpha_j (z_0^\top Q_0 z_0 - 1) \end{bmatrix} \geq 0 \ \ \forall j \in [J] \\ & \begin{bmatrix} \gamma I_d - \Lambda & \gamma \hat{\Sigma}^{1/2} \\ \gamma \hat{\Sigma}^{1/2} & A_0 \end{bmatrix} \geq 0, \ \begin{bmatrix} \gamma I_d - \Lambda & \gamma\hat{\mu} + \lambda/2 \\ (\gamma\hat{\mu} + \lambda/2)^\top & \alpha_0 \end{bmatrix} \geq 0. \end{array} \right.$$

If $r > 0$, then strong duality holds, that is, the above inequality becomes an equality.

In order to characterize the extremal distributions that attain the supremum in the worst-case expectation problem (4.1) over Chebyshev and Gelbrich ambiguity sets, we first derive the corresponding bi-duals of (4.1).

Theorem 7.11 (Finite bi-dual reformulation for Chebyshev ambiguity sets). If $\mathcal{P}$ is the Chebyshev ambiguity set (2.4) with $\mathcal{F} \subseteq \mathbb{R}^d \times \mathbb{S}_+^d$ and Assumption 7.8 holds, then the worst-case expectation problem (4.1) satisfies the weak duality relation

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \left\{ \begin{array}{ll} \sup & \sum_{j \in [J]} \mathrm{Tr}(Q_j \Theta_j) + 2 q_j^\top \theta_j + q_j^0 p_j \\ \text{s.t.} & \mu \in \mathbb{R}^d,\ M \in \mathbb{S}_+^d,\ p_j \in \mathbb{R}_+,\ \theta_j \in \mathbb{R}^d,\ \Theta_j \in \mathbb{S}_+^d \ \ \forall j \in [J] \\ & \begin{bmatrix} \Theta_j & \theta_j \\ \theta_j^\top & p_j \end{bmatrix} \geq 0,\ \mathrm{Tr}(Q_0 \Theta_j) - 2 z_0^\top Q_0 \theta_j + z_0^\top Q_0 z_0 p_j \leq p_j \ \ \forall j \in [J] \\ & \sum_{j \in [J]} p_j = 1,\ \mu = \sum_{j \in [J]} \theta_j,\ M = \sum_{j \in [J]} \Theta_j,\ (\mu, M) \in \mathcal{F}. \end{array} \right. \tag{7.5}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

If $\mathcal{F}$ is a convex and compact set with $M \succ \mu\mu^\top$ for all $(\mu, M) \in \mathrm{rint}(\mathcal{F})$, then strong duality holds, that is, the inequality (7.5) becomes an equality.

Proof. By decomposing the Gelbrich ambiguity set into Chebyshev ambiguity sets of the form $\mathcal{P}(\mu, M) = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{E}_\mathbb{P}[Z] = \mu, \mathbb{E}_\mathbb{P}[ZZ^\top] = M\}$, we obtain

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_\mathbb{P}[\ell(Z)] = \sup_{(\mu,M)\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}(\mu,M)} \mathbb{E}_\mathbb{P}[\ell(Z)]. \tag{7.6}$$

The inner maximization problem on the right-hand side of (7.6) represents a worst-case expectation problem over an instance of the ambiguity set (2.4) with the moment uncertainty set being the singleton $\{(\mu, M)\}$. The support function of this singleton is given by $\delta^*_{\{(\mu,M)\}}(\lambda, \Lambda) = \lambda^\top\mu + \mathrm{Tr}(\Lambda M)$. Thus Theorem 7.9 implies that the inner supremum on the right-hand side of (7.6) is bounded above by

$$\begin{array}{ll} \inf & \lambda_0 + \lambda^\top\mu + \mathrm{Tr}(\Lambda M) \\ \text{s.t.} & \lambda_0 \in \mathbb{R},\ \lambda \in \mathbb{R}^d,\ \Lambda \in \mathbb{S}^d,\ \alpha \in \mathbb{R}^J_+ \\ & \begin{bmatrix} \Lambda - Q_j + \alpha_j Q_0 & \frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0 \\ \left(\frac{1}{2}\lambda - q_j - \alpha_j Q_0 z_0\right)^\top & \lambda_0 - q_j^0 + \alpha_j(z_0^\top Q_0 z_0 - 1) \end{bmatrix} \succeq 0 \quad \forall j \in [J]. \end{array}$$

The dual of this semidefinite program can be represented as

$$\begin{array}{lll} \sup & \displaystyle\sum_{j\in[J]} \mathrm{Tr}(Q_j\Theta_j) + 2q_j^\top\theta_j + q_j^0 p_j & \\ \text{s.t.} & p_j \in \mathbb{R}_+,\ \theta_j \in \mathbb{R}^d,\ \Theta_j \in \mathbb{S}^d_+ & \forall j \in [J] \\ & \begin{bmatrix} \Theta_j & \theta_j \\ \theta_j^\top & p_j \end{bmatrix} \succeq 0,\ \mathrm{Tr}(Q_0\Theta_j) - 2z_0^\top Q_0\theta_j + z_0^\top Q_0 z_0 p_j \le p_j & \forall j \in [J] \\ & \displaystyle\sum_{j\in[J]} p_j = 1,\ \sum_{j\in[J]} \theta_j = \mu,\ \sum_{j\in[J]} \Theta_j = M. & \end{array}$$

Strong duality holds because the primal minimization problem admits a Slater point. Indeed, by defining $\Lambda = \lambda_0 I_d$ and setting λ_0 to a large value, one can ensure that the linear matrix inequality in the primal problem holds strictly. Replacing the inner supremum on the right-hand side of (7.6) with the above dual semidefinite program yields the upper bound in (7.5). If $\mathcal{F}$ is convex and compact with $M \succ \mu\mu^\top$ for all $(\mu, M) \in \mathrm{rint}(\mathcal{F})$, then (7.5) becomes an equality thanks to Theorem 7.9. □

Note that the bi-dual reformulation in (7.5) is solvable whenever $\mathcal{F}$ is compact. Indeed, its objective function is ostensibly continuous. In addition, it is easy to verify that its feasible region is compact provided that $\mathcal{F}$ is compact.

Theorem 7.12 (Finite bi-dual reformulation for Gelbrich ambiguity sets). If $\mathcal{P}$ is the Chebyshev ambiguity set (2.4) with $\mathcal{F}$ given by (2.8) and Assumption 7.8 holds, then the worst-case expectation problem (4.1) satisfies the weak duality

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

relation

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \max & \displaystyle\sum_{j\in[J]} \mathrm{Tr}(Q_j\Theta_j) + 2q_j^\top\theta_j + q_j^0 p_j \\ \text{s.t.} & \mu\in\mathbb{R}^d,\ M,U\in\mathbb{S}_+^d,\ C\in\mathbb{R}^{d\times d} \\ & p_j\in\mathbb{R}_+,\ \theta_j\in\mathbb{R}^d,\ \Theta_j\in\mathbb{S}_+^d & \forall j\in[J] \\ & \begin{bmatrix} M-U & C \\ C^\top & \hat{\Sigma}\end{bmatrix}\succeq 0,\ \begin{bmatrix} U & \mu \\ \mu^\top & 1\end{bmatrix}\succeq 0,\ \begin{bmatrix}\Theta_j & \theta_j \\ \theta_j^\top & p_j\end{bmatrix}\succeq 0 & \forall j\in[J] \\ & \mathrm{Tr}(Q_0\Theta_j) - 2z_0^\top Q_0\theta_j + z_0^\top Q_0 z_0 p_j \leq p_j & \forall j\in[J] \\ & \displaystyle\sum_{j\in[J]} p_j = 1,\ \mu = \sum_{j\in[J]}\theta_j,\ M = \sum_{j\in[J]}\Theta_j \\ & \|\hat{\mu}\|_2^2 - 2\mu^\top\hat{\mu} + \mathrm{Tr}(M+\hat{\Sigma}-2C) \leq r^2. \end{cases} \tag{7.7}$$

If $r > 0$, then strong duality holds, that is, the above inequality becomes an equality.

The proof of Theorem 7.12 follows from Proposition 2.3 and Theorem 7.11 and is thus omitted. We are now ready to construct extremal distributions $\mathbb{P}^\star \in \mathcal{P}(\mathcal{Z})$ that attain the supremum of the worst-case expectation problem (4.1) over the Chebyshev ambiguity set (2.4). To this end, fix any maximizer $(\mu^\star, M^\star, p^\star, \theta^\star, \Theta^\star)$ of the bi-dual problem (7.5), which exists if $\mathcal{F}$ is compact. Next, define the index sets

$$\mathcal{J}^\infty = \{j\in[J]\colon p_j^\star = 0,\ \Theta_j^\star \neq 0\} \quad \text{and} \quad \mathcal{J}^+ = \{j\in[J]\colon p_j^\star > 0\},$$

and define $\mathcal{J} = \mathcal{J}^+ \cup \mathcal{J}^\infty$. The extremal distributions $\mathbb{P}^\star$ will be constructed as mixtures of constituent distributions $\mathbb{P}_j$, $j\in\mathcal{J}$, corresponding to different pieces of the loss function ℓ. In the following, we use $\mathbb{P}\sim(\mu, M)$ to indicate that the distribution $\mathbb{P}$ has mean μ and second-order moment matrix M. Note that if $\mathcal{Z} = \{z\in\mathbb{R}^d\colon (z-z_0)^\top Q_0(z-z_0)\leq 1\}$ is the ellipsoid from Assumption 7.8 (ii) and $\mathbb{P}\in\mathcal{P}(\mathcal{Z})$ is a distribution supported on $\mathcal{Z}$ with $\mathbb{P}\sim(\mu, M)$, then we have

$$1 \geq \mathbb{E}_{\mathbb{P}}[(Z-z_0)^\top Q_0(Z-z_0)] = \mathrm{Tr}(Q_0M) + 2z_0^\top\mu + z_0^\top Q_0 z_0.$$

The inequality in the above expression holds because $\mathbb{P}\in\mathcal{P}(\mathcal{Z})$, and the equality holds because $\mathbb{P}\sim(\mu, M)$. The following lemma by Hanasusanto *et al.* (2015*a*, Proposition 6.1) shows the reverse implication. That is, if μ and M satisfy the above inequality, then there is a (discrete) distribution $\mathbb{P}\sim(\mu, M)$ supported on $\mathcal{Z}$.

Lemma 7.13 (Distributions on ellipsoids with given moments). If $\mathcal{Z}$ is the ellipsoid from Assumption 7.8 (ii), and if $\mathrm{Tr}(Q_0M) + 2z_0^\top\mu + z_0^\top Q_0 z_0 \leq 1$ for some $M\in\mathbb{S}_+^d$ and $\mu\in\mathbb{R}^d$ with $M\succeq\mu\mu^\top$, then there exists a discrete distribution $\mathbb{P}\in\mathcal{P}(\mathcal{Z})$ with at most $2d$ atoms that satisfies $\mathbb{P}\sim(\mu, M)$.

The proof of Lemma 7.13 is simple but tedious and is thus omitted.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Theorem 7.14 (Extremal distributions of Chebyshev ambiguity sets). If all conditions of Theorem 7.11 for weak as well as strong duality are satisfied and $(\mu^\star, M^\star, p^\star, \theta^\star, \Theta^\star)$ solves (7.5), then the following hold.

(i) If $\mathcal{J}^\infty = \emptyset$, then there exist discrete distributions $\mathbb{P}_j^\star \sim (\theta_j^\star/p_j^\star, \Theta_j^\star/p_j^\star)$ supported on $\mathcal{Z}$ for all $j \in \mathcal{J}^+$, and (4.1) is solved by $\mathbb{P}^\star = \sum_{j\in\mathcal{J}^+} p_j^\star \mathbb{P}_j^\star$. In addition, we have $\mathbb{P}^\star \sim (\mu^\star, M^\star)$, and $\mathbb{P}^\star$ is supported on $\mathcal{Z}$.

(ii) If $\mathcal{J}^\infty \neq \emptyset$, then there exist discrete distributions $\mathbb{P}_j^m \sim (\theta_j^\star/p_j^m, \Theta_j^\star/p_j^m)$ supported on $\mathcal{Z}$ for all $j \in \mathcal{J}$, where $p_j^m = (1 - |\mathcal{J}^\infty|/m)p_j^\star$ for $j \in \mathcal{J}^+$ and $p_j^m = 1/m$ for $j \in \mathcal{J}^\infty$, and where m is any integer with $m \geq |\mathcal{J}^\infty|$. In addition, (4.1) is asymptotically solved by $\mathbb{P}^m = \sum_{j\in\mathcal{J}} p_j^m \mathbb{P}_j^m$ as m grows.

Proof. As for assertion (i), the constraints of problem (7.7) imply that

$$\mathrm{Tr}(Q_0\Theta_j^\star/p_j^\star) - 2z_0^\top Q_0\theta_j^\star/p_j^\star + z_0^\top Q_0 z_0 \leq 1$$

and

$$\begin{bmatrix} \Theta_j^\star & \theta_j^\star \\ (\theta_j^\star)^\top & p_j^\star \end{bmatrix} \succeq 0 \quad \Longleftrightarrow \quad \Theta_j^\star/p_j^\star \succeq (\theta_j^\star/p_j^\star)(\theta_j^\star/p_j^\star)^\top$$

for all $j \in \mathcal{J}^+$. Lemma 7.13 thus guarantees that there exist discrete distributions $\mathbb{P}_j^\star \sim (\theta_j^\star/p_j^\star, \Theta_j^\star/p_j^\star)$, $j \in \mathcal{J}^+$, all of which are supported on $\mathcal{Z}$. Consequently, $\mathbb{P}^\star = \sum_{j\in\mathcal{J}^+} p_j^\star \mathbb{P}_j^\star$ is also supported on $\mathcal{Z}$. In addition, we have

$$\mathbb{E}_{\mathbb{P}^\star}[Z] = \sum_{j\in\mathcal{J}^+} p_j^\star \cdot \mathbb{E}_{\mathbb{P}_j^\star}[Z] = \sum_{j\in\mathcal{J}^+} p_j^\star \cdot \theta_j^\star/p_j^\star = \sum_{j\in\mathcal{J}^+} \theta_j^\star = \mu^\star$$

and

$$\mathbb{E}_{\mathbb{P}^\star}[ZZ^\top] = \sum_{j\in\mathcal{J}^+} p_j^\star \cdot \mathbb{E}_{\mathbb{P}_j^\star}[ZZ^\top] = \sum_{j\in\mathcal{J}^+} p_j^\star \cdot \Theta_j^\star/p_j^\star = \sum_{j\in\mathcal{J}^+} \Theta_j^\star = M^\star,$$

that is, $\mathbb{P}^\star \sim (\mu^\star, M^\star)$. As $(\mu^\star, M^\star) \in \mathcal{F}$, it is now clear that $\mathbb{P}^\star \in \mathcal{P}$ and that

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] \leq \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \sum_{j\in[J]} \mathrm{Tr}(Q_j\Theta_j^\star) + 2q_j^\top\theta_j^\star + q_j^0 p_j^\star,$$

where the equality follows from strong duality as established in Theorem 7.11. At the same time, the definition of $\mathbb{P}^\star$ as a mixture distribution and the definition of ℓ in (7.3) as a pointwise maximum of quadratic component functions implies that

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] \geq \sum_{j\in\mathcal{J}^+} p_j^\star \cdot \mathbb{E}_{\mathbb{P}_j^\star}[\ell_j(Z)] = \sum_{j\in[J]} \mathrm{Tr}(Q_j\Theta_j^\star) + 2q_j^\top\theta_j^\star + q_j^0 p_j^\star.$$

Specifically, the inequality holds because $\ell \geq \ell_j$ for every $j \in [J]$, and the equality holds because $\theta_j^\star = 0$ and $\Theta_j^\star = 0$ whenever $p_j^\star = 0$. Indeed, if $p_j^\star = 0$, then $\Theta_j^\star = 0$ because the index set $\mathcal{J}^\infty$ is empty, and the linear matrix inequality in (7.5) implies that $\theta_j^\star = 0$ whenever $\Theta_j^\star = 0$. The above inequalities thus ensure that $\mathbb{P}^\star$ solves the worst-case expectation problem (4.1). This completes the proof of assertion (i).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Next, we address assertion (ii). Similar arguments as in the proof of assertion (i) can be used to show that $\mathbb{P}^m \in \mathcal{P}$ for every $m \geq |\mathcal{J}^\infty|$. This implies that $\mathbb{E}_{\mathbb{P}^m}[\ell(Z)] \leq \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)]$ whenever $m \geq |\mathcal{J}^\infty|$. In addition, we observe that

$$\begin{aligned}\lim_{m\to\infty} \mathbb{E}_{\mathbb{P}^m}[\ell(Z)] &\geq \lim_{m\to\infty} \sum_{j\in\mathcal{J}} p_j^m \cdot \mathbb{E}_{\mathbb{P}^m}[\ell_j(Z)] \\ &= \sum_{j\in\mathcal{J}} \lim_{m\to\infty} p_j^m \cdot \mathbb{E}_{\mathbb{P}^m}[\ell_j(Z)] \\ &= \sum_{j\in[J]} \mathrm{Tr}(Q_j\Theta_j^\star) + 2q_j^\top\theta_j^\star + q_j^0 p_j^\star \\ &= \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)],\end{aligned}$$

where the second equality exploits the definition of $\mathbb{P}^m$ and the third equality follows from strong duality as established in Theorem 7.11. This completes the proof. □

Theorem 7.14 also applies to the Gelbrich ambiguity set, which constitutes a Chebyshev ambiguity set of the form (2.4) with $\mathcal{F}$ given by (2.8). The extremal distribution $\mathbb{P}^\star$ identified in Theorem 7.14 (i) constitutes a mixture of different distributions $\mathbb{P}_j^\star$, each of which corresponds to a component ℓ_j of the loss function ℓ; see Assumption 7.8 (i). The mixture components $\mathbb{P}_j^\star$ may be set to *any* distributions on $\mathcal{Z}$ that satisfy the prescribed moment conditions. Note that *discrete* distributions consistent with these requirements are guaranteed to exist thanks to Lemma 7.13. However, if $\mathcal{Z} = \mathbb{R}^d$, say, then one could also set $\mathbb{P}_j^\star$ to the Gaussian distribution with the given first and second moments. From the proof of Theorem 7.14 it becomes clear that $\mathbb{P}_j^\star$ must be supported on $\{z \in \mathcal{Z} : \ell_j(z) \geq \ell_{j'}(z)\ \forall j' \neq j\}$, which is generically non-convex. Therefore Kuhn *et al.* (2019, § 2.2) conjectured that the construction of $\mathbb{P}_j^\star$ is NP-hard. From the proof of Lemma 7.13 in Hanasusanto *et al.* (2015*a*, § 6) it becomes clear, however, that $\mathbb{P}_j^\star$ can be constructed efficiently. Similar comments are in order for the distributions $\mathbb{P}_j^m$ appearing in Theorem 7.14 (ii).

If $\mathcal{J}^\infty \neq \emptyset$, then the extremal distributions constructed in Theorem 7.14 contain diverging mixture components whose covariance matrices explode along certain recession directions of the support set $\mathcal{Z}$ (i.e. along the eigenvectors of $\Theta_j^\star$, $j \in \mathcal{J}^\infty$, corresponding to non-zero eigenvalues). However, these diverging mixture components are assigned weights that decay with their variances such that the covariance matrix of the entire mixture distribution remains bounded.

The following lemma establishes a sufficient condition for $\mathcal{J}^\infty$ to be empty, which ensures via Theorem 7.14 (i) that problem (4.1) is solvable.

Lemma 7.15. If all conditions of Theorem 7.14 are satisfied and the support set $\mathcal{Z}$ defined in (7.4) is compact, then $\mathcal{J}^\infty = \emptyset$, and thus problem (4.1) is solvable.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Proof. If $p_j^\star = 0$ for some $j \in [J]$, then the linear matrix inequality in (7.5) implies that $\theta_j^\star = 0$. Consequently, the jth trace inequality simplifies to $\mathrm{Tr}(Q_0\Theta_j^\star) \leq 0$. As $Q_0 \succ 0$ because $\mathcal{Z}$ is compact, we thus find that $\Theta_j^\star = 0$. In summary, we have shown that $p_j^\star = 0$ implies $\Theta_j^\star = 0$, and therefore $\mathcal{J}^\infty$ is empty as desired. □

We conclude this section with some remarks on worst-case expectation problems with more generic moment ambiguity sets. Translated into our terminology, Richter (1957) and Rogosinski (1958) show that if $\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathbb{E}_\mathbb{P}[f(Z)] = \mu\}$ for some $f\colon \mathcal{Z} \to \mathbb{R}^m$ and $\mu \in \mathbb{R}^m$, and if (4.1) is solvable, then the supremum in (4.1) is attained by a *discrete* distribution with at most $m + 2$ atoms. See Shapiro *et al.* (2009, Theorem 7.32) for modern proof of this result. Note also that, under the given assumptions, the worst-case expectation problem (4.1) can be recast as

$$\sup_{\varrho \in \mathcal{M}_+(\mathcal{Z})} \left\{ \int_\mathcal{Z} \ell(z)\,\mathrm{d}\rho(z)\colon \int_\mathcal{Z} \mathrm{d}\rho(z) = 1, \quad \int_\mathcal{Z} f(z)\,\mathrm{d}\rho(z) = \mu \right\}. \tag{7.8}$$

Problem (7.8) constitutes an *infinite*-dimensional linear program over the non-negative Borel measures on $\mathcal{Z}$ with $m + 1$ linear equality constraints. Every *finite*-dimensional linear program with non-negative variables and $m + 1$ equality constraints is known to admit an optimal basic feasible solution with at most $m + 1$ non-zero entries. The infinite-dimensional analogue of a basic feasible solution is a discrete measure with at most $m + 1$ atoms. Accordingly, one can prove that if (7.8) is solvable, then its supremum is attained by a measure with at most $m + 1$ atoms (Pinelis 2016, Corollary 5 and Proposition 6(v)). This result strengthens the Richter–Rogosinski theorem. However, the minimum number of atoms required for an optimal measure cannot be reduced beyond $m+1$ without additional assumptions.

The above reasoning implies that the worst-case expectation problem (4.1) and its reformulation (7.8) as a semi-infinite linear program can be reduced to a finite-dimensional optimization problem over the locations and probabilities of the $m + 1$ atoms of a discrete measure. Finite reductions of this type are routinely studied in optimal uncertainty quantification (Owhadi *et al.* 2013). However, they generically represent *non-convex* optimization problems. Indeed, even the integral of a linear function with respect to a discrete measure involves products of the probabilities and the coordinates of the measure's atoms. If (7.8) is solvable and ℓ is representable as a pointwise maximum of J concave functions, then the $m + 1$ atoms of an extremal measure can be further condensed. That is, using an induction argument and an iterative application of Jensen's inequality, one can show that (7.8) is solved by a discrete measure with at most J atoms (Han *et al.* 2015, Lemma 3.1). This result is significant even though J is not necessarily smaller than $m + 1$. It implies that (7.8) admits a finite reduction that optimizes over discrete measures with J atoms. And this (non-convex) finite reduction is intimately related to the dual problem (4.4) derived in Theorem 4.5 through a 'primal-worst-equals-dual-best' duality scheme for robust optimization problems (Beck and Ben-Tal 2009). Specifically, (4.4) can be viewed as a 'primal-worst' robust optimization problem,

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

and the finite reduction corresponding to discrete measures with J atoms can be viewed as the corresponding 'dual-best' optimization problem (Zhen *et al.* 2023). These problems share the same optimal value under mild regularity conditions. In addition, the (dual best) finite reduction can be convexified by applying a variable transformation and a perspectification trick (Han *et al.* 2015, Theorem 1.1). The same convex reformulation can also be obtained by dualizing the finite dual reformulation of the (primal worst) problem (4.4) as outlined in Section 7.1. For further details we refer to Zhen *et al.* (2023).

7.3. ϕ-divergence ambiguity sets

Recall that the ϕ-divergence ambiguity set (2.10) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq r\},$$

where r is a size parameter, ϕ is an entropy function in the sense of Definition 2.4, D_ϕ is the corresponding ϕ-divergence in the sense of Definition 2.5, and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. In the following, we first demonstrate that the worst-case expectation problem (4.1) over a ϕ-divergence ambiguity sets can be reformulated as a finite convex program whenever $\hat{\mathbb{P}}$ is discrete and ℓ is real-valued.

Assumption 7.16 (Discrete reference distribution). We have $\hat{\mathbb{P}} = \sum_{i\in[N]} \hat{p}_i \delta_{\hat{z}_i}$ for some $N \in \mathbb{N}$, where the probabilities $\hat{p}_i$, $i \in [N]$, are strictly positive and sum to 1, and where $\hat{z}_i \in \mathcal{Z}$ for every $i \in [N]$. In addition, $\ell(z) \in \mathbb{R}$ for all $z \in \mathcal{Z}$.

The requirement that $\hat{p}_i$ be positive for every $i \in [N]$ is non-restrictive because atoms with zero probability can simply be eliminated without changing $\hat{\mathbb{P}}$.

Theorem 7.17 (Finite dual reformulation for ϕ-divergence ambiguity sets). If $\mathcal{P}$ is the ϕ-divergence ambiguity set (2.10) and Assumption 7.16 holds, then the worst-case expectation problem (4.1) satisfies the weak duality relation

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf\limits_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} & \lambda_0 + \lambda r + \sum\limits_{i\in[N]} \hat{p}_i \cdot (\phi^*)^\pi(\ell(\hat{z}_i) - \lambda_0, \lambda) \\ \text{s.t.} & \lambda_0 + \lambda\, \phi^\infty(1) \geq \sup\limits_{z\in\mathcal{Z}} \ell(z), \end{cases} \tag{7.9}$$

where the product $\lambda\, \phi^\infty(1)$ is assumed to evaluate to ∞ if $\lambda = 0$ and $\phi^\infty(1) = \infty$. If $r > 0$ and ϕ is continuous at 1, then strong duality holds, that is, the above inequality becomes an equality.

Theorem 7.17 is an immediate corollary of Theorem 4.11. Indeed, problem (7.9) is obtained from (4.11) by re-expressing the integral with respect to the discrete reference distribution $\hat{\mathbb{P}}$ as a weighted sum. Thus no proof is required. Recall now that the *restricted* ϕ-divergence ambiguity set is defined as the set of all distributions $\mathbb{P} \in \mathcal{P}$ with $\mathbb{P} \ll \hat{\mathbb{P}}$. It is straightforward to verify that if $\mathcal{P}$ is discrete, then the corresponding worst-case expectation problem (4.1) admits a finite convex reformulation that is given by a relaxation of (7.9) without constraints. Details

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

are omitted for brevity. Next, we derive a finite convex program dual to (7.9) that allows us to construct an extremal distribution.

Theorem 7.18 (Finite bi-dual reformulations for ϕ-divergence ambiguity sets). If $\mathcal{P}$ is the ϕ-divergence ambiguity set (2.10), Assumption 7.16 holds, $r > 0$ and ϕ is continuous at 1, then problem (4.1) satisfies the strong duality relation

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \begin{cases} \max\limits_{p_0,\ldots,p_N\in\mathbb{R}_+} & p_0\overline{\ell} + \sum\limits_{i\in[N]} p_i \ell(\hat{z}_i) \\ \text{s.t.} & p_0 + \sum\limits_{i\in[N]} p_i = 1 \\ & p_0\phi^\infty(1) + \sum\limits_{i\in[N]} \hat{p}_i \phi\left(\dfrac{p_i}{\hat{p}_i}\right) \le r, \end{cases} \tag{7.10}$$

where $\overline{\ell}$ is shorthand for $\sup_{z\in\mathcal{Z}} \ell(z)$. The product $p_0\phi^\infty(1)$ is assumed to equal 0 if $p_0 = 0$ and $\phi^\infty(1) = \infty$. Similarly, $p_0\overline{\ell}$ is assumed to equal 0 if $p_0 = 0$ and $\overline{\ell} = \infty$.

The finite bi-dual reformulation (7.10) can readily be derived from the primal worst-case expectation problem (4.1) or from its finite dual reformulation (7.9). We find it insightful to derive (7.10) from (7.9). This is also more consistent with the general proof strategy outlined in Section 7.1. We will briefly touch on the derivation of (7.10) from the primal problem (4.1) after the proof.

Proof of Theorem 7.18. Assume first that $\phi^\infty(1) < \infty$. Under the assumptions stated in the theorem, the worst-case expectation problem (4.1) and its dual (7.9) share the same optimal value thanks to Theorem 7.17. By dualizing the single explicit constraint in (4.11) and using Lemma 7.1 (i), we thus find

$$\begin{aligned} &\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \\ &\quad = \inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \lambda_0 + \lambda r + \sum_{i\in[N]} \hat{p}_i \left(\sup_{y_i\in\mathbb{R}_+} y_i(\ell(\hat{z}_i) - \lambda_0) - \lambda\phi(y_i) \right) \\ &\qquad + \sup_{p_0\in\mathbb{R}_+} (\overline{\ell} - \lambda_0 - \lambda\phi^\infty(1))\, p_0. \end{aligned}$$

Interchanging the infima and suprema and rearranging terms further yields

$$\begin{aligned} &\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \\ &\quad = \sup_{p_0,y_1,\ldots,y_N\in\mathbb{R}_+} p_0\overline{\ell} + \sum_{i\in[N]} \hat{p}_i y_i \ell(\hat{z}_i) + \inf_{\lambda_0\in\mathbb{R}} \left(1 - p_0 - \sum_{i\in[N]} \hat{p}_i y_i\right)\lambda_0 \\ &\qquad + \inf_{\lambda\in\mathbb{R}_+} \left(r - p_0\phi^\infty(1) - \sum_{i\in[N]} \phi(y_i)\right)\lambda \end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$$= \begin{cases} \sup\limits_{p_0, y_0 \ldots, y_N \in \mathbb{R}_+} & p_0 \overline{\ell} + \sum\limits_{i \in [N]} \hat{p}_i y_i \ell(\hat{z}_i) \\ \text{s.t.} & p_0 + \sum\limits_{i \in [N]} \hat{p}_i y_i = 1, \;\; p_0 \phi^\infty(1) + \sum\limits_{i \in [N]} \hat{p}_i \phi(y_i) \leq r. \end{cases}$$

The first equality in the above expression follows from strong duality, which holds because $r > 0$ and ϕ is continuous at 1. Indeed, these conditions ensure that the resulting maximization problem admits a Slater point with $p_0 = 0$ and $y_i = 1$ for all $i \in [N]$. The substitution $p_i \leftarrow \hat{p}_i y_i$, $i \in [N]$, finally shows that the obtained problem is equivalent to (7.10). This proves the claim for $\phi^\infty(1) < \infty$.

Suppose next that $\phi^\infty(1) = \infty$, in which case $0\, \phi^\infty(1)$ evaluates to ∞. Hence the constraint in (4.11) is satisfied for any $(\lambda_0, \lambda) \in \mathbb{R} \times \mathbb{R}_+$ and is thus redundant. Repeating the steps from the first part of the proof, with obvious minor modifications, shows that (7.10) still holds if we assume that $p_0 \phi^\infty(1)$ and $p_0 \overline{\ell}$ evaluate to 0 when $p_0 = 0$. Indeed, this means that $p_0 = 0$ is the only feasible solution in (7.10), and problem (7.10) can be simplified by eliminating p_0 altogether. □

The finite bi-dual reformulation on the right-hand side of (7.10) has a linear objective function and a compact convex feasible region. Therefore it is solvable thanks to Weierstrass's maximum theorem. In particular, note that the feasible region is a subset of the probability simplex in $\mathbb{R}^{N+1}$. If there exists a worst-case scenario $\hat{z}_0 \in \arg\max_{z \in \mathcal{Z}} \ell(\hat{z})$ (which must satisfy $\ell(z_0) = \overline{\ell}$), then any maximizer $p^\star$ of the bi-dual can be used to construct an extremal distribution $\mathbb{P}^\star = \sum_{i=0}^N p_i^\star \delta_{\hat{z}_i}$ for the worst-case expectation problem (4.1). Indeed, the constraints of problem (7.10) ensure that $p_0^\star, \ldots, p_N^\star$ are non-negative probabilities that sum to 1. Thus $\mathbb{P}^\star$ is a valid distribution supported on $\mathcal{Z}$. Setting $\rho = \sum_{i=0}^N \delta_{\hat{z}_i}$, we also find

$$\begin{aligned} \mathrm{D}_\phi(\mathbb{P}^\star, \hat{\mathbb{P}}) &= \int_{\mathcal{Z}} \phi^\pi \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(z), \frac{\mathrm{d}\hat{\mathbb{P}}}{\mathrm{d}\rho}(z) \right) \mathrm{d}\rho(z) \\ &= \phi^\pi(p_0^\star, 0) + \sum_{i \in [N]} \phi^\pi(p_i^\star, \hat{p}_i) \\ &\leq r, \end{aligned}$$

where the first equality exploits the definition of D_ϕ, and the second equality exploits our choice of the reference distribution ρ. In addition, the inequality follows from the constraints of problem (7.10) and the observation that

$$\phi^\pi(p_0^\star, 0) = \phi^\infty(p_0^\star) = p_0^\star \phi^\infty(1).$$

This confirms that $\mathbb{P}^\star$ is feasible in (4.1). Also, its objective function value equals

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] = \sum_{i=0}^N p_i^\star \ell(\hat{z}_i).$$

As $\ell(\hat{z}_0) = \overline{\ell}$, we may conclude that $\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)]$ coincides with the maximum of the

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

bi-dual reformulation in (7.10), which in turn matches the supremum of (4.1) by virtue of Theorem 7.18. Hence $\mathbb{P}^\star$ is indeed a maximizer of problem (4.1).

Recall that if $\phi^\infty(1) = \infty$, then $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) = \infty$ unless $\mathbb{P} \ll \hat{\mathbb{P}}$. Therefore every distribution $\mathbb{P}$ in a ϕ-divergence ambiguity set around $\hat{\mathbb{P}}$ must be absolutely continuous with respect to $\hat{\mathbb{P}}$. If $\phi^\infty(1) < \infty$, on the other hand, then $\mathbb{P}$ can assign a positive probability to points in $\mathcal{Z}$ that have zero probability under $\hat{\mathbb{P}}$. Note that $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}})$ only depends on *how much* probability mass $\mathbb{P}$ removes from the support of $\hat{\mathbb{P}}$, but it does not depend on *where* that probability mass is moved. As nature aims to maximize the expected loss, it will move all of this probability mass to a point with maximal loss within $\mathcal{Z}$ (i.e. to some point $\hat{z}_0 \in \arg\max_{z \in \mathcal{Z}} \ell(\hat{z})$).

If $\mathcal{P}$ is the *restricted* ϕ-divergence ambiguity set (2.11), Assumption 7.16 holds, $r > 0$ and ϕ is continuous at 1, then Theorem 7.18 remains valid with a minor modification. That is, one must append the constraint $p_0 = 0$ to the finite bi-dual reformulation on the right-hand side of (7.10). Details are omitted for brevity.

7.4. Optimal transport ambiguity sets

Recall that the optimal transport ambiguity set (2.27) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{OT}_c(\mathbb{P}, \hat{\mathbb{P}}) \leq r\},$$

where $r \geq 0$ is a size parameter, c is a transportation cost function in the sense of Definition 2.14, OT_c is the corresponding optimal transport discrepancy in the sense of Definition 2.15, and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. We will first show that the worst-case expectation problem (4.1) over an optimal transport ambiguity set can often be reformulated as a finite convex minimization problem. To this end, we restrict attention to discrete reference distributions as in Assumption 7.16, and we impose convexity conditions on the transportation cost function, the loss function and the support set $\mathcal{Z}$. In addition, we impose a mild technical condition on the support points of the discrete reference distribution $\hat{\mathbb{P}}$.

Assumption 7.19 (Regularity conditions for optimal transport ambiguity sets).

(i) The loss function ℓ is a pointwise maximum of $J \in \mathbb{N}$ concave functions, that is, $\ell(z) = \max_{j \in [J]} \ell_j(z)$, where $-\ell_j \colon \mathcal{Z} \to \mathbb{R}$ is proper, convex and closed.

(ii) The support set is representable as $\mathcal{Z} = \{z \in \mathbb{R}^d \colon g_k(z) \leq 0 \ \ \forall k \in [K]\}$ for some $K \in \mathbb{N}$, where $g_k \colon \mathcal{Z} \to \overline{\mathbb{R}}$ is proper, convex and closed.

(iii) The transportation cost function $c(z, \hat{z})$ is convex in z for every fixed $\hat{z} \in \mathcal{Z}$.

(iv) The support point $\hat{z}_i$ belongs to $\mathrm{rint}(\mathrm{dom}(c(\cdot, \hat{z}_i)))$ and constitutes a Slater point for $\mathcal{Z}$ in the sense of Definition 7.3 for every $i \in [N]$.

Assumption 7.19 (i) is non-restrictive because any continuous function ℓ on a compact set $\mathcal{Z}$ can be uniformly approximated by a pointwise maximum of finitely many concave functions ℓ_j, $j \in [J]$, albeit maybe at the expense of requiring large numbers J of pieces. Assumptions 7.19 (ii) and (iii) are restrictive but

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

satisfied by support sets and transportation cost functions commonly encountered in applications. Finally, Assumption 7.19 (iv) is of a purely technical nature and can always be enforced by slightly perturbing the problem data.

Theorem 7.20 (Finite dual reformulation for optimal transport ambiguity sets). If $\mathcal{P}$ is the optimal transport ambiguity set (2.27) and Assumptions 7.16 and 7.19 hold, then the worst-case expectation problem (4.1) obeys the weak duality relation

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf & \lambda r + \sum_{i\in[N]} \hat{p}_i s_i \\ \text{s.t.} & \lambda \in \mathbb{R}_+,\ s_i \in \mathbb{R} & \forall i \in [N], j \in [J], k \in [K] \\ & \zeta_{ij}^{\ell}, \zeta_{ij}^{c}, \zeta_{ijk}^{g} \in \mathbb{R}^d & \forall i \in [N], j \in [J], k \in [K] \\ & (-\ell_j)^*(\zeta_{ij}^{\ell}) + (c_i^*)^{\pi}(\zeta_{ij}^{c}, \lambda) & \\ & \quad + \sum_{k\in[K]} (g_k^*)^{\pi}(\zeta_{ijk}^{g}, \alpha_{ijk}) \leq s_i & \forall i \in [N], j \in [J] \\ & \zeta_{ij}^{\ell} + \zeta_{ij}^{c} + \sum_{k\in[K]} \zeta_{ijk}^{g} = 0 & \forall i \in [N], j \in [J], \end{cases} \tag{7.11}$$

where $c_i \colon \mathcal{Z} \to \overline{\mathbb{R}}$ is defined by $c_i(z) = c(z, \hat{z}_i)$ for every $i \in [N]$. If $r > 0$, then strong duality holds, that is, the above inequality becomes an equality.

The dual minimization problem of Theorem 7.20 constitutes a finite convex program because the conjugates $(-\ell_j)^*$, c_i^* and g_k^* and their perspectives are convex functions. It accommodates $O(NJK)$ decision variables and $O(NJ)$ constraints.

Proof of Theorem 7.20. By Theorem 4.18, we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \begin{cases} \inf & \lambda r + \sum_{i\in[N]} \hat{p}_i s_i \\ \text{s.t.} & \lambda \in \mathbb{R}_+,\ s_i \in \mathbb{R} & \forall i \in [N] \\ & \sup_{z\in\mathcal{Z}} \ell(z) - \lambda c(z, \hat{z}_i) \leq s_i & \forall i \in [N], \end{cases}$$

where s_i represents an auxiliary epigraphical decision variable for any $i \in [N]$. By Assumption 7.19 (i) and the definition of the functions c_i, $i \in [N]$, the above minimization problem is equivalent to the following robust convex program:

$$\begin{array}{ll} \inf & \lambda r + \sum_{i\in[N]} \hat{p}_i s_i \\ \text{s.t.} & \lambda \in \mathbb{R}_+,\ s_i \in \mathbb{R} \\ & \sup_{z\in\mathcal{Z}} \ell_j(z) - \lambda c_i(z) \leq s_i \ \ \forall i \in [N], j \in [J]. \end{array} \tag{7.12}$$

For any fixed $i \in [N]$ and $j \in [J]$, Assumptions 7.19 (i) and 7.19 (ii) imply that the embedded maximization problem over z constitutes a convex program. In addition,

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

this problem admits a Slater point $\hat{z}_i$ thanks to Assumptions 7.19 (i) and 7.19 (iv). In order to dualize this convex program, we first recall from Lemma 7.2 that the conjugate of $f(z) = -\ell_j(z) + \lambda c_i(z)$ at $\zeta \in \mathbb{R}^d$ can be represented as

$$f^*(\zeta) = \min_{\zeta_{ij}^\ell, \zeta_{ij}^c \in \mathbb{R}^d} \left\{ (-\ell_j)^*\left(\zeta_{ij}^\ell\right) + (c_i^*)^\pi\left(\zeta_{ij}^c, \lambda\right) : \zeta_{ij}^\ell + \zeta_{ij}^c = \zeta \right\}.$$

By Theorem 7.4, we thus obtain

$$\sup_{z \in \mathcal{Z}} \ell_j(z) - \lambda c_i(z) = \begin{cases} \min & (-\ell_j)^*(\zeta_{ij}^\ell) + (c_i^*)^\pi\left(\zeta_{ij}^c, \lambda\right) + \displaystyle\sum_{k \in [K]} (g_k^*)^\pi\left(\zeta_{ijk}^g, \alpha_{ijk}\right) \\ \text{s.t.} & \alpha_{ijk} \in \mathbb{R}_+, \ \zeta_{ij}^\ell, \zeta_{ij}^c, \zeta_{ijk}^g \in \mathbb{R}^d \ \ \forall k \in [K] \\ & \zeta_{ij}^\ell + \zeta_{ij}^c + \displaystyle\sum_{k \in [K]} \zeta_{ijk}^g = 0. \end{cases}$$

Next, we replace each embedded maximization problem in (7.12) with its equivalent dual minimization problem, and we eliminate the corresponding minimization operators, which is allowed because all minima are attained. This yields the desired finite convex reformulation of the problem dual to (4.1), and it establishes weak duality. If $r > 0$, then strong duality follows from Theorem 4.18. □

The finite convex reformulation of Theorem 7.20 was first derived under the more restrictive assumption that $c(z, \hat{z}) = \|z - \hat{z}\|$ by Mohajerin Esfahani and Kuhn (2018, Theorem 4.2) and later generalized to arbitrary convex transportation cost functions by Zhen *et al.* (2023, § 6). We next derive a finite convex bi-dual for the worst-case expectation problem (4.1) over the optimal transport ambiguity set (2.27), which forms the basis for identifying extremal distributions that (asymptotically) attain the supremum in (4.1). Our derivation will rely on the following two lemmas.

First, we derive a formula for the conjugate of a scaled perspective function.

Lemma 7.21 (Conjugates of scaled perspectives I). If $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is proper, convex and closed, and if $\alpha \in \mathbb{R}_+$, then, for all $y \in \mathbb{R}^d$ and $y_0 \in \mathbb{R}$, we have

$$(\alpha f^\pi)^*(y, y_0) = \begin{cases} 0 & \text{if } (f^*)^\pi(y, \alpha) \leq -y_0, \\ \infty & \text{otherwise.} \end{cases}$$

Proof. Assume first that $\alpha > 0$. If $\lambda > 0$, then we have

$$\alpha f^\pi(z, \lambda) = \alpha\lambda f(z/\lambda) = (\alpha f)^\pi(z, \lambda) \quad \text{for all } z \in \mathbb{R}^d.$$

Similarly, if $\lambda = 0$, then $\alpha f^\pi(z, \lambda) = \alpha f^\infty(z) = (\alpha f)^\infty(z) = (\alpha f)^\pi(z, \lambda)$ for all $z \in \mathbb{R}^d$. We have thus shown that $\alpha f^\pi = (\alpha f)^\pi$. Next, define the set

$$\begin{aligned} \mathcal{C} &= \{(y, y_0) \in \mathbb{R}^d \times \mathbb{R} : (\alpha f)^*(y) \leq -y_0\} \\ &= \{(y, y_0) \in \mathbb{R}^d \times \mathbb{R} : (f^*)^\pi(y, \alpha) \leq -y_0\}, \end{aligned}$$

where the second equality follows from the definition of the perspective function.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

By Rockafellar (1970, Corollary 13.5.1), we have $(\alpha f)^\pi = \delta^*_{\mathcal{C}}$. As $\mathcal{C}$ is closed, this implies that $(f^\pi)^* = \delta^{**}_{\mathcal{C}} = \delta_{\mathcal{C}}$, and thus the claim follows for $\alpha > 0$.

Assume next that $\alpha = 0$. In this case we have $\alpha f^\pi = \delta_{\mathrm{dom}(f^\pi)}$ thanks to our rules of extended arithmetic. This observation implies that

$$\begin{aligned}
(\alpha f^\pi)^*(y, y_0) &= \delta^*_{\mathrm{dom}(f^\pi)}(y, y_0) \\
&= \sup_{\lambda \in \mathbb{R}_{++}} \sup_{z \in \mathbb{R}^d} \{y^\top z + y_0 \lambda : (z, \lambda) \in \mathrm{dom}(f^\pi)\} \\
&= \sup_{\lambda \in \mathbb{R}_{++}} \lambda \sup_{z \in \mathbb{R}^d} \{y^\top (z/\lambda) : z/\lambda \in \mathrm{dom}(f)\} + y_0 \lambda \\
&= \sup_{\lambda \in \mathbb{R}_{++}} \lambda \delta^*_{\mathrm{dom}(f)}(y) + \lambda y_0 \\
&= \begin{cases} 0 & \text{if } \delta^*_{\mathrm{dom}(f)}(y) + y_0 \leq 0, \\ \infty & \text{otherwise.} \end{cases}
\end{aligned}$$

Note that it is sufficient to optimize only over $\lambda > 0$ because $\mathrm{dom}(f^\pi) \subseteq \mathbb{R}^d \times \mathbb{R}_+$. As f is convex and closed, we have $f = f^{**}$ thanks to Lemma 4.2, and thus we find

$$\delta^*_{\mathrm{dom}(f)}(y) = \delta^*_{\mathrm{dom}(f^{**})}(y) = (f^*)^\infty(y) = (f^*)^\pi(y, 0),$$

where the second and third equalities follow from Rockafellar (1970, Theorem 13.3) and from the definition of the perspective, respectively. Combining the above observations proves the claim for $\alpha = 0$. □

The next lemma derives a formula for the conjugate of a sum of scaled perspectives. It thus generalizes Lemma 7.21, which addresses only one single scaled perspective, and it is also related to Lemma 7.2, which characterizes the conjugate of a sum of arbitrary convex functions – not necessarily scaled perspectives.

Lemma 7.22 (Conjugates of perspective functions II). Suppose that $f_i : \mathbb{R}^d \to \overline{\mathbb{R}}$, $i \in [m]$, are proper, convex and closed and that there is $\bar{z} \in \cap_{i \in [m]} \mathrm{rint}(\mathrm{dom}(f_i))$. Let $f(z_1, \ldots, z_m, \lambda) = \sum_{i \in [m]} \alpha_i f_i^\pi(z_i, \lambda)$ be a weighted sum of the corresponding perspective functions with weight vector $\alpha \in \mathbb{R}^m_+$. Then the conjugate of f satisfies

$$f^*(y_1, \ldots, y_m, y_0) = \begin{cases} 0 & \begin{cases} \text{if } \exists \beta \in \mathbb{R}^m \text{ with } \sum_{i \in [m]} \beta_i = y_0 \text{ and} \\ (f_i^*)^\pi(y_i, \alpha_i) \leq -\beta_i \ \ \forall i \in [m], \end{cases} \\ \infty & \text{otherwise.} \end{cases}$$

Proof. By using a variable splitting trick as in the proof of Lemma 7.2, we find

$$\begin{aligned}
f^*(y_1, \ldots, y_m, y_0) &= \sup_{z_1, \ldots, z_m \in \mathbb{R}^d} \sup_{\lambda \in \mathbb{R}_+} y_0 \lambda + \sum_{i \in [m]} y_i^\top z_i - \sum_{i \in [m]} \alpha_i f_i^\pi(z_i, \lambda) \\
&= \begin{cases} \sup & y_0 \lambda + \sum_{i \in [m]} y_i^\top z_i - \alpha_i f_i^\pi(z_i, \lambda_i) \\ z_1, \ldots, z_m \in \mathbb{R}^d & \\ \lambda \in \mathbb{R}, \lambda_1, \ldots, \lambda_m \in \mathbb{R}_+ & \\ \text{s.t.} & \lambda_i = \lambda \quad i \in [m]. \end{cases}
\end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The resulting convex maximization problem admits a Slater point. To see this, recall that there exists $\bar{z} \in \cap_{i\in[m]} \operatorname{rint}(\operatorname{dom}(f_i))$. As $\operatorname{dom}(f^\pi)$ is contained in the cone generated by $\operatorname{dom}(f) \times \{1\}$, we may thus conclude that the solution with $\lambda = 1$, $\lambda_i = 1$ and $z_i = \bar{z}$ for all $i \in [m]$ constitutes a Slater point. Therefore the above maximization problem admits a strong Lagrangian dual, that is, we have

$$\begin{aligned} & f^*(y_1, \ldots, y_m, y_0) \\ &\quad = \min_{\beta_1,\ldots,\beta_m\in\mathbb{R}} \sup_{\substack{z_1,\ldots,z_m\in\mathbb{R}^d \\ \lambda\in\mathbb{R},\, \lambda_1,\ldots,\lambda_m\in\mathbb{R}_+}} y_0\lambda + \sum_{i\in[m]} y_i^\top z_i - \alpha_i f_i^\pi(z_i, \lambda_i) + \beta_i(\lambda_i - \lambda) \\ &\quad = \min_{\beta_1,\ldots,\beta_m\in\mathbb{R}} \left\{ \sum_{i\in[m]} (\alpha_i f_i^\pi)^*(y_i, \beta_i) \colon \sum_{i\in[m]} \beta_i = y_0 \right\}; \end{aligned}$$

see also Theorem 7.4. By Lemma 7.21, we further have $(\alpha_i f_i^\pi)^* = \delta_{\mathcal{C}_i}$, where

$$\mathcal{C}_i = \{(y, y_0) \in \mathbb{R}^d \times \mathbb{R} \colon (f_i^*)^\pi(y, \alpha_i) \le -y_0\}$$

for all $i \in [m]$. Substituting this alternative expression for $(\alpha_i f_i^\pi)^*$ into the above dual problem yields the desired formula. Thus the claim follows. □

We emphasize that Lemmas 7.21 and 7.22 are complementary to Lemma 4.11. Indeed, while Lemma 4.11 evaluates the conjugate only with respect to the first argument of a perspective function, Lemmas 7.21 and 7.22 do so with respect to *both* arguments. We are now ready to derive a finite bi-dual reformulation of the worst-case expectation problem over an optimal transport ambiguity set.

Theorem 7.23 (Finite bi-dual reformulation for optimal transport ambiguity sets). If $\mathcal{P}$ is the optimal transport ambiguity set (2.27) and Assumptions 7.16 and 7.19 hold, then the worst-case expectation problem (4.1) satisfies the weak duality relation

$$\begin{aligned} & \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \\ & \le \begin{cases} \sup & \displaystyle\sum_{i\in[N]} \sum_{j\in[J]} -(-\ell_j)^\pi(p_{ij}\hat{z}_i + z_{ij}, p_{ij}) & \\ \text{s.t.} & p_{ij} \in \mathbb{R}_+,\ z_{ij} \in \mathbb{R}^d & \forall i \in [N],\ j \in [J] \\ & g_k^\pi(p_{ij}\hat{z}_i + z_{ij}, p_{ij}) \le 0 & \forall i \in [N],\ j \in [J],\ k \in [K] \\ & \displaystyle\sum_{j\in[J]} p_{ij} = \hat{p}_i & \forall i \in [N] \\ & \displaystyle\sum_{i\in[N]} \sum_{j\in[J]} c_i^\pi(p_{ij}\hat{z}_i + z_{ij}, p_{ij}) \le r, & \end{cases} \end{aligned} \tag{7.13}$$

where $c_i \colon \mathcal{Z} \to \overline{\mathbb{R}}$ is defined by $c_i(z) = c(z, \hat{z}_i)$ for every $i \in [N]$. If $r > 0$, then strong duality holds, that is, the above inequality becomes an equality.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Proof. We will show that (7.13) is obtained by dualizing the finite dual reformulation (7.11) of problem (4.1). To see this, we assign Lagrange multipliers $p_{ij} \in \mathbb{R}_+$ and $z_{ij} \in \mathbb{R}^d$, $i \in [N]$, $j \in [J]$, to the first and second constraint groups in (7.11), respectively. The Lagrangian dual of (7.11) can then be represented compactly as

$$\sup_{p \geq 0, z} \inf_{\substack{\lambda \geq 0, \alpha \geq 0 \\ s, \zeta^\ell, \zeta^c, \zeta^g}} L_1(s; p, z) + L_2(\zeta^\ell; p, z) + L_3(\lambda, \zeta^c; p, z, \lambda) + L_4(\alpha, \zeta^g; p, z),$$

where the Lagrangian is additively separable with respect to four disjoint groups of primal decision variables, namely s, ζ^ℓ, (λ, ζ^c) and (α, ζ^g). The corresponding partial Lagrangians are defined as follows:

$$\begin{aligned}
L_1(s; p, z) &= \sum_{i \in [N]} \hat{p}_i s_i - \sum_{i \in [N]} \sum_{j \in [J]} p_{ij} s_i, \\
L_2(\zeta^\ell; p, z) &= \sum_{i \in [N]} \sum_{j \in [J]} p_{ij} \cdot (-\ell_j)^*(\zeta^\ell_{ij}) - z_{ij}^\top \zeta^\ell_{ij}, \\
L_3(\lambda, \zeta^c; p, z) &= \lambda r + \sum_{i \in [N]} \sum_{j \in [J]} p_{ij} \cdot (c_i^*)^\pi(\zeta^c_{ij}, \lambda) - z_{ij}^\top \zeta^c_{ij}, \\
L_4(\alpha, \zeta^g; p, z) &= \sum_{i \in [N]} \sum_{j \in [J]} \sum_{k \in [K]} p_{ij} \cdot (g_k^*)^\pi(\zeta^g_{ijk}, \alpha_{ijk}) - z_{ij}^\top \zeta^g_{ijk}.
\end{aligned}$$

These partial Lagrangians can be minimized separately with respect to the primal decision variables. For example, an elementary calculation shows that

$$\inf_s L_1(s; p, z) = \begin{cases} 0 & \text{if } \sum_{j \in [J]} p_{ij} = \hat{p}_i \ \ \forall i \in [N], \\ -\infty & \text{otherwise.} \end{cases}$$

Recall now that $-\ell_j$ is proper, convex and closed, which implies via Lemma 4.2 that $(-\ell_j)^{**} = -\ell_j$. Note also that minimizing $L_2(\zeta^\ell; p, z)$ with respect to ζ^ℓ amounts to evaluating the conjugate of a sum of conjugates with mutually different arguments. By using Lemma 7.1 (i) and applying a few elementary manipulations, we thus find

$$\inf_{\zeta^\ell} L_2(\zeta^\ell; p, z) = \sum_{i \in [N]} \sum_{j \in [J]} -(-\ell_j)^\pi(z_{ij}, p_{ij}).$$

Similarly, recall that c_i is proper, convex and closed such that $c_i^{**} = c_i$. Note also that minimizing $L_3(\lambda, \zeta^c; p, z)$ with respect to λ and ζ^c amounts to evaluating the conjugate of a sum of perspective functions with one common argument. By using Lemma 7.22 and applying a few elementary manipulations, we thus find

$$\inf_{\lambda \geq 0, \zeta^c} L_3(\lambda, \zeta^c; p, z) = \begin{cases} 0 & \begin{cases} \text{if } \exists \beta_{ij} \in \mathbb{R}^m \text{ with } \sum_{i \in [N]} \sum_{j \in [J]} \beta_{ij} = r \text{ and} \\ c_i^\pi(z_{ij}, p_{ij}) \leq \beta_{ij} \ \ \forall i \in [N], \ j \in [J], \end{cases} \\ -\infty & \text{otherwise.} \end{cases}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Finally, recall that g_k is proper, convex and closed such that $g_k^{**} = g_k$. Note also that minimizing $L_4(\alpha, \zeta^g; p, z)$ with respect to α and ζ^g amounts to evaluating the conjugate of a sum of perspective functions with mutually different arguments. By using Lemma 7.21 and applying a few elementary manipulations, we thus find

$$\inf_{\alpha \geq 0, \zeta^g} L_4(\alpha, \zeta^g; p, z) = \begin{cases} 0 & \text{if } g_k^\pi(z_{ij}, p_{ij}) \leq 0 \ \ \forall i \in [N],\ j \in [J],\ k \in [K], \\ -\infty & \text{otherwise.} \end{cases}$$

Substituting the infima of the partial Lagrangians into the dual objective yields the following equivalent reformulation for the problem dual to (7.11):

$$\begin{array}{lll} \sup & \displaystyle\sum_{i\in[N]} \sum_{j\in[J]} -(-\ell_j)^\pi(z_{ij}, p_{ij}) & \\ \text{s.t.} & p_{ij} \in \mathbb{R}_+,\ \beta_{ij} \in \mathbb{R},\ z_{ij} \in \mathbb{R}^d & \forall i \in [N],\ j \in [J] \\ & g_k^\pi(z_{ij}, p_{ij}) \leq 0 & \forall i \in [N],\ j \in [J],\ k \in [K] \\ & \displaystyle\sum_{j\in[J]} p_{ij} = \hat{p}_i & \forall i \in [N] \\ & c_i^\pi(z_{ij}, p_{ij}) \leq \beta_{ij} & \forall i \in [N],\ j \in [J] \\ & \displaystyle\sum_{i\in[N]} \sum_{j\in[J]} \beta_{ij} = r & \end{array} \tag{7.14}$$

Note that if the finite dual reformulation (7.11) of the worst-case expectation problem is viewed as an instance of the primal convex program (P), then problem (7.14) represents the corresponding instance of the dual convex program (D). By Assumptions 7.16 and 7.19, problem (7.14) admits a Slater point with $p_{ij} = \hat{p}_i/J$ and $z_{ij} = \hat{z}_i$ for all $i \in [N]$ and $j \in [J]$. Thus strong duality holds thanks to Theorem 7.4 (i). It remains to be shown that (7.14) is equivalent to (7.13). To this end, note first that the last constraint in (7.14) can be relaxed to a less-than-or-equal-to inequality without increasing the problem's supremum such that $\beta_{ij} = c_i^\pi(z_{ij}, p_{ij})$ at optimality. This allows us to eliminate the β_{ij} variables from (7.14). Problem (7.13) is then obtained by applying the substitution $z_{ij} \leftarrow z_{ij} - p_{ij}\hat{z}_i$. □

The finite bi-dual reformulation (7.13) is guaranteed to be solvable provided that the transportation cost function satisfies the following additional assumption.

Assumption 7.24 (Identity of indiscernibles). The transportation cost function is real-valued and satisfies $c(z, \hat{z}) = 0$ if and only if $z = \hat{z}$.

Lemma 7.25 (Solvability of the finite bi-dual reformulation). Suppose that Assumptions 7.16, 7.19 and 7.24 hold. Then problem (7.13) is solvable.

Proof. Under the stated assumptions, problem (7.13) maximizes an upper semicontinuous function over a compact feasible region, and thus the claim follows from Weierstrass's maximum theorem. To see that the objective function of (7.13) is upper semicontinuous, note that the functions $-\ell_j$ are proper, convex and closed for all $j \in [J]$ thanks to Assumption 7.19 (i). By Rockafellar (1970, pp. 35, 67),

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

their perspectives are proper, convex and closed too; see also Zhen *et al.* (2023, Proposition C.2). Thus the *negative* perspective functions appearing in the objective function of problem (7.13) are indeed upper semicontinuous. Similarly, one can show that the feasible region of problem (7.13) is closed. Indeed, g_k and c_i are proper, convex and closed for all $k \in [K]$ and $i \in [N]$ thanks to Assumption 7.19 and Definition 2.14. This readily implies that their perspectives are lower semicontinuous, and thus the feasible region of (7.13) is indeed closed. To see that the feasible region is also bounded, note first that $p_{ij} \in [0, 1]$ for all $i \in [N]$ and $j \in [J]$. Indeed, these variables must be non-negative and compatible with the probabilities $\hat{p}_i$, $i \in [N]$, of the discrete reference distribution. Next, we show that the variables z_{ij} for $i \in [N]$ and $j \in [J]$ are restricted to a bounded set as well. Indeed, by Zhen *et al.* (2023, Lemma C.10), which applies thanks to Assumption 7.24 and Definition 2.14, there exists $\delta > 0$ such that $c_i(\hat{z}_i + z) \geq \delta\|z\|_2 - 1$ for all $z \in \mathbb{R}^d$ and $i \in [N]$. The last constraint of problem (7.13) therefore implies that

$$\sum_{i\in[N]} \sum_{j\in[J]} c_i^\pi(p_{ij}\hat{z}_i + z_{ij}, p_{ij}) \leq r \quad \Longrightarrow \quad \sum_{i\in[N]} \sum_{j\in[J]} \|z_{ij}\|_2 \leq \frac{1+r}{\delta},$$

where we used the identity

$$\sum_{i\in[N]} \sum_{j\in[J]} p_{ij} = \sum_{i\in[N]} \hat{p}_i = 1$$

and the definition of the perspective function. Thus the feasible region of (7.13) is indeed bounded. □

We are now ready to construct extremal distributions $\mathbb{P}^\star \in \mathcal{P}(\mathcal{Z})$ that attain the supremum of the worst-case expectation problem (4.1) over the optimal transport ambiguity set (2.27). To this end, fix any maximizer $(p^\star, z^\star)$ of the bi-dual problem (7.13), which exists thanks to Lemma 7.25. Next, define the index sets

$$\mathcal{J}_i^\infty = \{j \in [J] : p_{ij}^\star = 0,\ z_{ij}^\star \neq 0\} \quad \text{and} \quad \mathcal{J}_i^+ = \{j \in [J] : p_{ij}^\star > 0\},$$

and define $\mathcal{J}_i = \mathcal{J}_i^+ \cup \mathcal{J}_i^\infty$ for any $i \in [N]$. The following theorem uses the maximizer $(p^\star, z^\star)$ and the corresponding index sets to construct $\mathbb{P}^\star$.

Theorem 7.26 (Extremal distributions of optimal transport ambiguity sets). Suppose that all conditions of Theorem 7.23 for weak and strong duality are satisfied, Assumption 7.24 holds, and $(p^\star, z^\star)$ solves (7.13). Then the following hold.

(i) If $\mathcal{J}_i^\infty = \emptyset$ for all $i \in [N]$, then problem (4.1) is solved by

$$\mathbb{P}^\star = \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^+} p_{ij}^\star\, \delta_{\hat{z}_i + z_{ij}^\star / p_{ij}^\star}.$$

(ii) If $\mathcal{J}_i^\infty \neq \emptyset$ for some $i \in [N]$, then problem (4.1) is asymptotically solved by

$$\mathbb{P}^m = \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i} p_{ij}^m\, \delta_{z_{ij}^m}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

as $m \in \mathbb{N}$, $m \geq \max_{i\in[N]} |\mathcal{J}_i^\infty|$, grows, where

$$p_{ij}^m = \begin{cases} \left(1 - \dfrac{|\mathcal{J}_i^\infty|}{m}\right) p_{ij}^\star & \text{if } j \in \mathcal{J}_i^+, \\ \dfrac{\hat{p}_i}{m} & \text{if } j \in \mathcal{J}_i^\infty, \end{cases} \quad \text{and} \quad z_{ij}^m = \begin{cases} \hat{z}_i + \dfrac{z_{ij}^\star}{p_{ij}^\star} & \text{if } j \in \mathcal{J}_i^+, \\ \hat{z}_i + \dfrac{z_{ij}^\star}{p_{ij}^m} & \text{if } j \in \mathcal{J}_i^\infty. \end{cases}$$

Proof. In view of assertion (i), we first show that $\mathbb{P}^\star$ defined in the statement of the theorem is feasible in the worst-case expectation problem (4.1). To this end, observe first that feasibility of $(p^\star, z^\star)$ in (7.13) implies that $p_{ij}^\star \geq 0$ for all $i \in [N]$ and $j \in [J]$, and that $\sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^+} p_{ij}^\star = 1$. Note also that $\hat{z}_i + z_{ij}^\star / p_{ij}^\star \in \mathcal{Z}$ for all $i \in [N]$ and $j \in \mathcal{J}_i^+$ due to the second constraint in (7.13). This confirms that $\mathbb{P}^\star \in \mathcal{P}(\mathcal{Z})$. The penultimate constraint group of problem (7.13) also implies that

$$\sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^+} p_{ij}^\star \, \delta_{(\hat{z}_i + z_{ij}^\star / p_{ij}^\star, \hat{z}_i)} \in \Gamma(\mathbb{P}^\star, \hat{\mathbb{P}})$$

constitutes a valid transportation plan for morphing $\hat{\mathbb{P}}$ into $\mathbb{P}^\star$. Thus we find

$$\begin{aligned} \mathrm{OT}_c(\mathbb{P}^\star, \hat{\mathbb{P}}) &\leq \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^+} p_{ij}^\star \cdot c\left(\hat{z}_i + z_{ij}^\star / p_{ij}^\star, \hat{z}_i\right) \\ &= \sum_{i\in[N]} \sum_{j\in[J]} c_i^\pi \left(p_{ij}^\star \hat{z}_i + z_{ij}^\star, p_{ij}^\star\right) \\ &\leq r. \end{aligned}$$

Here the equality holds because all terms corresponding to $i \in [N]$ and $j \notin \mathcal{J}_i^+$ vanish. Indeed, if $j \notin \mathcal{J}_i^+$, then $p_{ij}^\star = 0$. As $\mathcal{J}_i^\infty = \emptyset$, this implies that $z_{ij}^\star = 0$. Thus we have $c_i^\pi(p_{ij}^\star \hat{z}_i + z_{ij}^\star, p_{ij}^\star) = c_i^\pi(0,0) = c_i^\infty(0) = 0$ by the definitions of the perspective and the recession function. The second inequality in the above expression follows from the last constraint in (7.13). In summary, we have shown that $\mathbb{P}^\star$ is feasible in (4.1). As for the objective function value of $\mathbb{P}^\star$, note that

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] \leq \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \sum_{i\in[N]} \sum_{j\in[J]} -(-\ell_j)^\pi \left(p_{ij}^\star \hat{z}_i + z_{ij}^\star, p_{ij}^\star\right),$$

where the second inequality follows from the weak duality relation established in Theorem 7.23. At the same time, however, the expected loss under $\mathbb{P}^\star$ satisfies

$$\begin{aligned} \mathbb{E}_{\mathbb{P}^\star}[\ell(Z)] &= \sum_{i\in[N]} \sum_{j\in\mathcal{J}^+} \max_{j'\in[J]} p_{ij}^\star \ell_{j'} \left(\hat{z}_i + \frac{z_{ij}^\star}{p_{ij}^\star}\right) \\ &\geq \sum_{i\in[N]} \sum_{j\in\mathcal{J}^+} -(-\ell_j)^\pi \left(p_{ij}^\star \hat{z}_i + z_{ij}^\star, p_{ij}^\star\right) \\ &= \sum_{i\in[N]} \sum_{j\in[J]} -(-\ell_j)^\pi \left(p_{ij}^\star \hat{z}_i + z_{ij}^\star, p_{ij}^\star\right), \end{aligned}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where the inequality uses the definition of the perspective function and the trivial observation that $j \in \mathcal{J}^+$ is a feasible choice for $j' \in [J]$. The last equality holds once more because $p^\star_{ij} = 0$ implies $z^\star_{ij} = 0$ and $(-\ell_j)^\pi(0,0) = (-\ell_j)^\infty(0) = 0$ by the definition of the perspective and the recession function. In summary, the above inequalities imply that $\mathbb{P}^\star$ is optimal in (4.1). Hence assertion (i) follows.

As for assertion (ii), we first show that $\mathbb{P}^m \in \mathcal{P}$ for any fixed $m \geq \max_{i\in[N]} |\mathcal{J}_i^\infty|$. The constraints of problem (7.13) imply that $p^m_{ij} \geq 0$ for all $j \in \mathcal{J}_i$ and $i \in [N]$ and that $\sum_{i\in[N]} \sum_{j\in\mathcal{J}} p^m_{ij} = 1$. They also imply that $z^m_{ij} \in \mathcal{Z}$ for every $j \in \mathcal{J}_i$ and $i \in [N]$. This is easy to see if $j \in \mathcal{J}_i^+$. If $j \in \mathcal{J}_i^\infty$, on the other hand, then $p^\star_{ij} = 0$, $z^\star_{ij} \neq 0$ and $g^\pi_k(z^\star_{ij}, 0) \leq 0$ for all $k \in [K]$, which implies via Rockafellar (1970, Theorem 8.6) that $z^\star_{ij}$ is a recession direction of $\mathcal{Z}$. Geometrically, this means that the ray emanating from any point in $\mathcal{Z}$ along the direction $z^\star_{ij}$ never leaves $\mathcal{Z}$. Thus $z^m_{ij} = \hat{z}_i + m\, z^\star_{ij}/\hat{p}_i \in \mathcal{Z}$ for all $i \in [N]$ and $j \in \mathcal{J}_i^\infty$. In addition, one verifies that

$$\sum_{i\in[N]} \sum_{j\in\mathcal{J}_i} p^m_{ij}\, \delta_{(z^m_{ij}, \hat{z}_i)} \in \Gamma(\mathbb{P}^\star, \hat{\mathbb{P}})$$

constitutes a valid transportation plan for morphing $\hat{\mathbb{P}}$ into $\mathbb{P}^m$. Thus we find

$$\begin{aligned}
&\mathrm{OT}_c(\mathbb{P}^m, \hat{\mathbb{P}}) \\
&\leq \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i} p^m_{ij}\, c(z^m_{ij}, \hat{z}_i) \\
&= \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^+} p^\star_{ij} \left(1 - \frac{|\mathcal{J}_i^\infty|}{m}\right) c\left(\hat{z}_i + \frac{z^\star_{ij}}{p^\star_{ij}}, \hat{z}_i\right) + \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^\infty} \frac{\hat{p}_i}{m}\, c\left(\hat{z}_i + m\frac{z^\star_{ij}}{\hat{p}_i}, \hat{z}_i\right) \\
&\leq \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^+} p^\star_{ij}\, c\left(\hat{z}_i + \frac{z^\star_{ij}}{p^\star_{ij}}, \hat{z}_i\right) + \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^\infty} \lim_{m\to\infty} \frac{\hat{p}_i}{m}\, c\left(\hat{z}_i + m\frac{z^\star_{ij}}{\hat{p}_i}, \hat{z}_i\right) \\
&= \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^+} p^\star_{ij}\, c\left(\hat{z}_i + \frac{z^\star_{ij}}{p^\star_{ij}}, \hat{z}_i\right) + \sum_{i\in[N]} \sum_{j\in\mathcal{J}_i^\infty} \lim_{m\to\infty} \frac{\hat{p}_i}{m}\, c\left(m\frac{z^\star_{ij}}{\hat{p}_i}, \hat{z}_i\right) \\
&= \sum_{i\in[N]} \sum_{j\in[J]} c_i^\pi\left(p^\star_{ij}\hat{z}_i + z^\star_{ij}, p^\star_{ij}\right) \\
&\leq r,
\end{aligned}$$

where the first equality follows from the definitions of p^m_{ij} and z^m_{ij}. The second inequality holds because the transportation cost function $c(z, \hat{z})$ is non-negative and convex in z, which implies that both terms in the third line are non-decreasing in m. The second equality follows from Assumption 7.24, which ensures that $c(z, \hat{z})$ is real-valued such that the reference point in the definition of the recession function of $c(\cdot, \hat{z}_i)$ can be chosen freely. The third equality exploits the definition of the perspective function c_i^π and the observation that $c_i^\pi(0,0) = c_i^\infty(0) = 0$. Finally, the last inequality follows from the last constraint of problem (7.13). We have thus

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

shown that $\mathbb{P}^m$ is feasible in (4.1). In analogy to analysis for $\mathbb{P}^\star$, one can show that the asymptotic expected loss $\lim_{m\to\infty} \mathbb{E}_{\mathbb{P}^m}[\ell(Z)]$ is at least as large as the optimal value $\sum_{i\in[N]} \sum_{j\in[J]} -(-\ell_j)^\pi(p^\star_{ij}\hat{z}_i + z^\star_{ij}, p^\star_{ij})$ of the finite bi-dual reformulation (7.13). However, as the suprema of (4.1) and (7.13) match, it is clear that the distributions $\mathbb{P}^m$, $m \in \mathbb{N}$, must be asymptotically optimal in (4.1). □

If $\mathcal{J}_i^\infty \neq \emptyset$ for some $i \in [N]$, then the extremal distributions constructed in Theorem 7.26 send atoms with decaying probabilities to infinity along specific recession directions $z^\star_{ij}$, $j \in \mathcal{J}_i^\infty$, of the support set $\mathcal{Z}$. Moving atoms to infinity is possible even when only a finite transportation budget r is available, provided that the probability mass transported scales inversely with the transportation cost. The following lemma establishes sufficient conditions for $\mathcal{J}_i^\infty$ to be empty for every $i \in [N]$, which ensures via Theorem 7.26 (i) that problem (4.1) is solvable.

Lemma 7.27. If all assumptions of Theorem 7.26 are satisfied and either of the following conditions holds, then $\mathcal{J}_i^\infty = \emptyset$ for every $i \in [N]$, and (4.1) is solvable.

(i) The transportation cost function grows superlinearly in its first argument. By this we mean that $c_i^\infty(z) = \infty$ for any $z \neq 0$ and for any $i \in [N]$.

(ii) The support set $\mathcal{Z}$ is bounded.

Proof. As usual, let $(p^\star, z^\star)$ be a maximizer of problem (7.13), which exists thanks to Lemma 7.25. As for assertion (i), assume that the transportation cost function grows superlinearly. For the sake of argument, assume also that there exists $i \in [N]$ with $\mathcal{J}_i^\infty \neq \emptyset$. For every $j \in \mathcal{J}_i^\infty$ we thus have $p^\star_{ij} = 0$ and $z^\star_{ij} \neq 0$. Hence we find

$$c_i^\pi\left(p^\star_{ij}\hat{z}_i + z^\star_{ij}, p^\star_{ij}\right) = c_i^\infty(z^\star_{ij}) = \infty,$$

where the first equality uses the definition of the perspective function, and the second equality holds because the transportation cost function grows superlinearly. Thus $(p^\star, z^\star)$ violates the last constraint of problem (7.13), which contradicts its assumed feasibility. We may thus conclude that $\mathcal{J}_i^\infty = \emptyset$ and that (4.1) is solvable.

As for assertion (ii), assume now that $\mathcal{Z}$ is bounded. Without loss of generality, we may also assume that $p^\star_{ij} = 0$ for some $i \in [N]$ and $j \in [J]$, for otherwise $\mathcal{J}_i^\infty$ is trivially empty. The constraints of problem (7.13) then ensure that $g_k^\pi(z^\star_{ij}, 0) \leq 0$ for all $k \in [K]$, which implies via Rockafellar (1970, Theorem 8.6) that $z^\star_{ij}$ is a recession direction of $\mathcal{Z}$. As $\mathcal{Z}$ is compact, however, this implies that $z^\star_{ij} = 0$. We may thus again conclude that $\mathcal{J}_i^\infty = \emptyset$ and that (4.1) is solvable. □

Condition (i) of Lemma 7.27 is satisfied whenever $\mathcal{P}$ is a p-Wasserstein ball and the transportation cost function is of the form $c(z, \hat{z}) = \|z - \hat{z}\|^p$ for some $p > 1$.

The structural properties of the distributions that solve the worst-case expectation problem (4.1) over an optimal transport ambiguity set, as well as necessary and sufficient conditions for their existence, were studied by Wozabal (2012), Owhadi and Scovel (2017), Yue *et al.* (2022) and Gao and Kleywegt (2023). In particular,

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

significant efforts were spent on characterizing the extremal distributions of a Wasserstein ball centred at a discrete reference distributions with N atoms. The earliest result in this domain is due to Wozabal (2012, Theorem 3.3), who showed that the worst-case expectation of a continuous bounded loss function is attained by a discrete distribution with at most $N + 3$ atoms. Later, Owhadi and Scovel (2017, Theorem 2.3) and Gao and Kleywegt (2023, Corollary 1) managed to sharpen this result by showing that the worst-case expectation is in fact attained by a discrete distribution with at most $N + 2$ or even only $N + 1$ atoms, respectively; see also Yue *et al.* (2022, Theorem 4). Theorem 7.26 (i) and Lemma 7.27 reveal that if $\mathcal{Z}$ is bounded and the loss function ℓ is concave, thus satisfying Assumption 7.19 (i) with $J = 1$, then the worst-case expected loss is attained by an N-point distribution. For more general loss functions, however, every N-point distributions can be strictly suboptimal even if problem (4.1) is solvable; see Kuhn *et al.* (2019, Example 5). The results in this section are based on Zhen *et al.* (2023, § 6).

7.5. Nash equilibria and adversarial attacks

The DRO problem (1.2) can be viewed as a zero-sum game in which the decision-maker first chooses a decision $x \in \mathcal{X}$, and nature subsequently responds with a distribution $\mathbb{P} \in \mathcal{P}$ that adapts to x. Throughout this section we will refer to (1.2) as the *primal* DRO problem. In addition, one can study the *dual* DRO problem

$$\sup_{\mathbb{P}\in\mathcal{P}} \inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)], \tag{7.15}$$

where nature first selects a distribution $\mathbb{P} \in \mathcal{P}$, and the decision-maker subsequently responds with a decision $x \in \mathcal{X}$ that adapts to $\mathbb{P}$. In contrast to the primal DRO problem (1.2), whose objective function is linear in $\mathbb{P}$, the objective function of the dual DRO problem (7.15) is concave in $\mathbb{P}$. This difference makes the dual DRO problem more challenging to solve. It is now natural to seek conditions that imply strong duality and thus ensure that the infimum of the primal DRO problem (1.2) coincides with the supremum of the dual DRO problem (7.15). One readily verifies that strong duality is implied, for example, by the existence of a Nash equilibrium $(x^\star, \mathbb{P}^\star) \in \mathcal{X} \times \mathcal{P}$ satisfying the saddle point condition

$$\mathbb{E}_{\mathbb{P}}[\ell(x^\star, Z)] \leq \mathbb{E}_{\mathbb{P}^\star}[\ell(x^\star, Z)] \leq \mathbb{E}_{\mathbb{P}^\star}[\ell(x, Z)] \quad \text{for all } x \in \mathcal{X},\ \mathbb{P} \in \mathcal{P}. \tag{7.16}$$

We emphasize that the reverse implication is false, that is, strong duality does not necessarily imply the existence of a Nash equilibrium. The primal DRO problem naturally arises in many applications. The practical usefulness of the dual DRO problem, on the other hand, is less evident because this problem assumes somewhat unrealistically that the decision-maker observes the distribution that governs Z. Nevertheless, the dual DRO problem has deep connections to robust statistics and machine learning, as well as several other disciplines, as we explain below.

From the perspective of robust statistics, a minimizer $x^\star$ of the primal DRO problem (1.2) can be interpreted as a *robust estimator* for the minimizer of the

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

stochastic program $\min_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)]$ corresponding to an unknown distribution $\mathbb{P}_0$. When $x^\star$ and $\mathbb{P}^\star$ satisfy the saddle point condition (7.16), then the robust estimator $x^\star$ constitutes a best response to $\mathbb{P}^\star$. Hence it solves the stochastic program corresponding to $\mathbb{P}^\star$; see also Lehmann and Casella (2006, Chapter 5). For this reason, $\mathbb{P}^\star$ is often referred to as the *least favourable distribution*. The existence of $\mathbb{P}^\star$ makes $x^\star$ a plausible estimator because it ensures that $x^\star$ is the minimizer of a stochastic program corresponding to *some* distribution in the ambiguity set.

Algorithms for computing Nash equilibria of DRO problems are also relevant for applications in machine learning. To see this, recall that adversarial training aims to immunize machine learning models against adversarial perturbations of the input data (Szegedy *et al.* 2014, Goodfellow *et al.* 2015, Mądry *et al.* 2018, Wang *et al.* 2019, Kurakin, Goodfellow and Bengio 2017). In this context, it is of interest to generate adversarial examples, that is, maliciously designed inputs that mislead prediction models encoded by parameters $x \in \mathcal{X}$. As a naïve approach to constructing adversarial examples, one could simply solve the worst-case expectation problem

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(\hat{x}, Z)], \tag{7.17}$$

which seeks a test distribution that maximizes the expected prediction loss of one particular model encoded by $\hat{x}$. Thus any solution $\mathbb{P}^\star$ of (7.17) can be viewed as an adversarial attack, and samples drawn from $\mathbb{P}^\star$ are naturally interpreted as adversarial examples. In order to develop efficient strategies for attacking as well as defending prediction models, however, it is desirable to construct adversarial attacks that fool a broad spectrum of different models. Such attacks are called *transferable* in the machine learning literature (Tramèr *et al.* 2017, Demontis *et al.* 2019, Kurakin *et al.* 2017). The dual DRO problem (7.15) can be used to construct transferable attacks in a systematic manner. Indeed, the solutions of (7.15) are *not* tailored to a particular model $\hat{x} \in \mathcal{X}$. Instead, they aim to attack *all* models $x \in \mathcal{X}$ simultaneously. If the primal DRO problem (1.2) has a unique minimizer $x^\star$, then this minimizer can be recovered by solving the stochastic program corresponding to the adversary's Nash strategy $\mathbb{P}^\star$.

To date, dual DRO problems have only been investigated in the context of specific applications. For example, it is known that the least favourable distributions in distributionally robust estimation and Kalman filtering problems with a 2-Wasserstein ambiguity set centred at a Gaussian reference distribution are themselves Gaussian and can be computed efficiently via semidefinite programming (Shafieezadeh-Abadeh *et al.* 2018, Nguyen *et al.* 2023*b*). Several recent studies describe similar results for distributionally robust optimal control problems with a 2-Wasserstein ambiguity set (Al Taha *et al.* 2023, Hajar *et al.* 2023, Kargin *et al.* 2024*a*,*b*,*c*,*d*, Taşkesen *et al.* 2024). When the Wasserstein ambiguity set is replaced with a Kullback–Leibler ambiguity set around a Gaussian reference distribution, then the least favourable distributions remain Gaussian and can be determined in

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

quasi-closed form (Levy and Nikoukhah 2004, 2012). In fact, these results even extend to generalized τ-divergence ambiguity sets (Zorzi 2016, 2017*b*). Gaussian distributions also solve several other minimax games reminiscent of DRO problems, which are relevant for applications in statistics, control and information theory (Başar and Mintz 1972, 1973, Başar and Max 1973, Başar 1977, Başar and Başar 1982, Başar 1983, Başar and Başar 1984, Başar and Wu 1985, 1986). Furthermore, it is possible to characterize the Nash equilibria of distributionally robust pricing and auction design problems with support-only and Markov ambiguity sets in closed form (Bergemann and Schlag 2008, Koçyiğit *et al.* 2020, 2022, Anunrojwong *et al.* 2024, Chen *et al.* 2024*b*). Minimax theorems establishing strong duality between primal and dual DRO problems involving more general optimal transport ambiguity sets are reported by Blanchet *et al.* (2022*b*), Shafiee *et al.* (2023), Frank and Niles-Weed (2024*b*) and Pydi and Jog (2024).

8. Regularization by robustification

Classical stochastic optimization seeks decisions that perform well under a probability distribution $\hat{\mathbb{P}}$ estimated from training data. By ignoring any information about estimation errors in $\hat{\mathbb{P}}$, however, stochastic optimization tends to output overfitted decisions that incur a low expected loss under $\hat{\mathbb{P}}$ but may perform poorly under the unknown population distribution $\mathbb{P}$. This problem becomes more acute if training data is scarce. A key advantage of DRO *vis-à-vis* stochastic optimization is that it has access to information about estimation errors. DRO uses this information to prevent overfitting. Robustifying a stochastic optimization problem against distributional uncertainty can thus be viewed as a form of implicit regularization.

We now show that there is often a deep connection between *implicit* regularization (achieved by robustifying a problem against distributional uncertainty) and *explicit* regularization (achieved by adding a penalty term to the problem's objective function). This discussion complements and extends several results from Section 6. For example, in Section 6.9 we have seen that the worst-case expected value of a linear loss function with respect to a Kullback–Leibler ambiguity set centred at a Gaussian distribution coincides with the nominal expected loss and a variance regularization term. Similarly, in Section 6.13 we have seen that the worst-case expected value of a convex loss function with respect to a 1-Wasserstein ambiguity set coincides with the nominal expected loss and a Lipschitz regularization term. See also Sections 6.14 and 6.15 for some variants and generalizations of this result.

In Section 8.1 we will show – in broad generality – that the worst-case expected loss over a ϕ-divergence ambiguity set is closely related to the nominal expected loss with a *variance* regularization term. Similarly, in Section 8.2 we will show that the worst-case expected loss over a Wasserstein ambiguity set is closely related to the nominal expected loss with *variation* and *Lipschitz* regularization terms. In Section 8.3 we will further show that many popular risk measures are Lipschitz-continuous in the distribution of the relevant risk factors with respect to

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

a Wasserstein distance. This implies that the worst-case risk over a Wasserstein ambiguity set is closely related to the nominal risk and a *Lipschitz* regularization term. We remark that the connections between robustification and regularization are less well understood for moment ambiguity sets. From Section 6.8 we know that the worst-case risk of a linear loss function over a Gelbrich ambiguity set often coincides with the nominal risk and a 2-norm regularization term. However, it is unclear whether similar results can be obtained for nonlinear loss functions or other moment ambiguity sets. Therefore we will not touch on moment ambiguity sets in this section. We emphasize that the connections between robustification and regularization often enable statistical analyses of DRO problems; see Section 10.

8.1. ϕ-divergence ambiguity sets

As a motivating example, we show that robustification with respect to a Pearson χ^2-divergence ambiguity set is closely related to variance regularization. To see this, recall first that the Pearson χ^2-divergence ambiguity set (2.17) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathbb{P}(\mathcal{Z})\colon \chi^2(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

If ℓ is a bounded Borel function, Proposition 2.13 readily implies that

$$\mathcal{P} \subseteq \left\{\mathbb{P} \in \mathbb{P}(\mathcal{Z})\colon \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}\right\},$$

and thus we may conclude that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}.$$

Hence the worst-case expected loss with respect to a Pearson χ^2-divergence ambiguity set of radius r around $\hat{\mathbb{P}}$ is bounded above by the mean–standard deviation risk measure with risk-aversion coefficient $r^{1/2}$ evaluated under $\hat{\mathbb{P}}$. By slight abuse of terminology, the scaled standard deviation $r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}$ is commonly referred to as a variance regularizer. By leveraging Theorem 4.15, the above bound can be extended to arbitrary (possibly unbounded) Borel loss functions. This extension critically relies on the following lemma.

Lemma 8.1 (Variance formula). For any reference distribution $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$, size parameter $r > 0$ and Borel function $\ell \in \mathcal{L}(\mathbb{R}^d)$ with $\mathbb{E}_{\hat{\mathbb{P}}}[|\ell(Z)|] < \infty$, we have

$$\inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \lambda r + \frac{\mathbb{E}_{\hat{\mathbb{P}}}[(\ell(Z)-\lambda_0)^2]}{4\lambda} = r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}. \tag{8.1}$$

Proof. If $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)^2] = \infty$, then both sides of (8.1) evaluate to ∞, and thus the claim follows. In the remainder of the proof, we may thus assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)^2] < \infty$. In this case, one readily verifies that the partial minimization problem over λ_0 is solved by $\lambda_0^\star = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]$. Substituting $\lambda_0^\star$ back into the objective function reveals that the infimum on the left-hand side of (8.1) equals $\inf_{\lambda\in\mathbb{R}_+} \lambda r + \mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]/4\lambda$. In order to prove (8.1), it suffices to realize that this minimization problem over λ is solved by $\lambda^\star = \sqrt{\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]/(4r)}$. This observation completes the proof. □

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Theorem 8.2 (Variance regularization). If $\mathcal{P}$ is the Pearson χ^2-divergence ambiguity set (2.17) and $\mathbb{E}_{\hat{\mathbb{P}}}[|\ell(Z)|] < \infty$, then we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}.$$

Proof. The claim trivially holds if $r = 0$. We may thus assume that $r > 0$. Recall now that the entropy function ϕ inducing the Pearson χ^2-divergence satisfies $\phi(s) = (s-1)^2$ if $s \geq 0$ and $\phi(s) = \infty$ if $s < 0$. Hence the conjugate entropy function ϕ^* satisfies $\phi^*(t) = \frac{1}{4}t^2 + t$ if $t \geq -2$ and $\phi^*(t) = -1$ if $t < -2$, and its domain is given by $\mathrm{dom}(\phi^*) = \mathbb{R}$. As $\phi^\infty(1) = \infty$, all distributions $\mathbb{P} \in \mathcal{P}$ are absolutely continuous with respect to $\hat{\mathbb{P}}$. Thus Theorem 4.15 applies, and we find

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] &= \inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \lambda_0 + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\ell(Z) - \lambda_0, \lambda)] \\ &\leq \inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \lambda r + \frac{\mathbb{E}_{\hat{\mathbb{P}}}[(\ell(Z) - \lambda_0)^2]}{4\lambda} \\ &= \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2},\end{aligned}$$

where the inequality holds because $\phi^*(t) \leq \frac{1}{4}t^2 + t$, and the second equality follows from Lemma 8.1. Thus the claim follows. □

Most ϕ-divergences are smooth and non-negative and thus resemble the Pearson χ^2-divergence locally around 1 (Polyanskiy and Wu 2024, § 7.10). Accordingly, one can use a Taylor expansion to show that robustification over a ϕ-divergence ambiguity set of sufficiently small size r is often equivalent to variance regularization. To formalize this result, we assume from now on that ϕ is differentiable.

Assumption 8.3 (Differentiability). The entropy function ϕ is twice continuously differentiable on a neighbourhood of 1 with $\phi(1) = \phi'(1) = 0$ and $\phi''(1) = 2$.

The assumption that $\phi'(1) = 0$ incurs no loss of generality. Indeed, any entropy function ϕ is equivalent to a transformed entropy function $\tilde{\phi}$ defined by $\tilde{\phi}(t) = \phi(t) - \phi'(1) \cdot t + \phi'(1)$ with $\tilde{\phi}'(1) = 0$. That is, both ϕ and $\tilde{\phi}$ induce the same divergence. Note that all entropy functions listed in Table 2.1 – except for the one associated with the total variation – satisfy $\phi'(1) = 0$. The assumption that $\phi''(1) = 2$ serves as an arbitrary normalization but will simplify calculations.

Recall now that the restricted ϕ-divergence ambiguity set (2.11) is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathbb{P} \ll \hat{\mathbb{P}},\ \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

Here $\mathcal{Z}$ is a closed support set, $r \in \mathbb{R}_+$ is a size parameter, ϕ is an entropy function in the sense of Definition 2.4, D_ϕ is the corresponding ϕ-divergence in the sense of Definition 2.5, and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. The following theorem provides a leading-order Taylor expansion of the worst-case expectation over $\mathcal{P}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Theorem 8.4 (Taylor expansion of worst-case expectation). If $\mathcal{P}$ is the restricted ϕ-divergence ambiguity set (2.11), the entropy function ϕ satisfies Assumption 8.3 and the loss $\ell(Z)$ is $\hat{\mathbb{P}}$-almost surely bounded, then we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2} + o(r^{1/2}). \tag{8.2}$$

Proof. Note that (8.2) trivially holds if $r = 0$. Similarly, if $\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)] = 0$, then $\ell(Z)$ coincides $\hat{\mathbb{P}}$-almost surely with $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]$. As $\mathcal{P}$ is a restricted ϕ-divergence ambiguity set, this readily implies that $\mathbb{E}_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]$ for all $\mathbb{P} \in \mathcal{P}$. Indeed, any $\mathbb{P} \in \mathcal{P}$ satisfies $\mathbb{P} \ll \hat{\mathbb{P}}$. Hence (8.2) is again trivially satisfied. In the remainder of the proof we my therefore assume that $r > 0$ and that $\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)] > 0$.

Assumption 8.3 implies that $\phi(s) = (s-1)^2 + o(s^2)$. By Taylor's theorem with Peano remainder, ϕ can thus be bounded from below (or above) locally around 1 by a quadratic function whose second derivative is slightly smaller (or larger) than $\phi''(1) = 2$. Thus there exists a function $\kappa\colon \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{\varepsilon\downarrow 0} \kappa(\varepsilon) = 0$ and

$$\frac{1}{1+\kappa(\varepsilon)} \cdot s^2 \le \phi(1+s) \le (1+\kappa(\varepsilon)) \cdot s^2 \quad \text{for all } s \in [-\varepsilon, +\varepsilon] \tag{8.3}$$

for all sufficiently small ε. The rest of the proof proceeds in two steps, both of which exploit (8.3). First, we show that the right-hand side of (8.2) provides a *lower* bound on the worst-case expected loss over $\mathcal{P}$ (Step 1). Next, we show that the right-hand side of (8.2) also provides an *upper* bound on the worst-case expected loss over $\mathcal{P}$ (Step 2). Taken together, Steps 1 and 2 will imply the claim.

Step 1. Every distribution $\mathbb{P}$ in the restricted ϕ-divergence ambiguity set $\mathcal{P}$ satisfies $\mathbb{P} \ll \hat{\mathbb{P}}$ and has thus a density function $f \in \mathcal{L}^1(\hat{\mathbb{P}})$ with respect to $\hat{\mathbb{P}}$. Here $\mathcal{L}^1(\hat{\mathbb{P}})$ denotes as usual the family of all Borel functions from $\mathcal{Z}$ to $\mathbb{R}$ that are integrable with respect to $\hat{\mathbb{P}}$. As $\mathbb{P} \ll \hat{\mathbb{P}}$, we have $\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}) = \mathbb{E}_{\hat{\mathbb{P}}}[\phi(f(Z))]$ (see also Section 2.2). Thus the worst-case expectation problem over $\mathcal{P}$ can be recast as

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \left\{ \begin{array}{ll} \sup\limits_{f\in\mathcal{L}^1(\hat{\mathbb{P}})} & \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)f(Z)] \\ \text{s.t.} & \hat{\mathbb{P}}(f(Z) \ge 0) = 1 \\ & \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)] = 1 \\ & \mathbb{E}_{\hat{\mathbb{P}}}[\phi(f(Z))] \le r. \end{array} \right.$$

Renaming $f(z) + 1$ as $f(z)$ further yields

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \left\{ \begin{array}{ll} \sup\limits_{f\in\mathcal{L}^1(\hat{\mathbb{P}})} & \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)f(Z)] \\ \text{s.t.} & \hat{\mathbb{P}}(f(Z) \ge -1) = 1 \\ & \mathbb{E}_{\hat{\mathbb{P}}}[f(Z)] = 0 \\ & \mathbb{E}_{\hat{\mathbb{P}}}[\phi(1+f(Z))] \le r. \end{array} \right. \tag{8.4}$$

Next, introduce an auxiliary function $\varepsilon\colon \mathbb{R}_+ \to \mathbb{R}_+$ satisfying

$$\varepsilon(r) = 2r^{1/2} \cdot \frac{\operatorname{ess\,sup}_{\hat{\mathbb{P}}}[|\ell(Z) - \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]|]}{\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}}.$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

In addition, for every $r \in \mathbb{R}_+$, define the function $f_r^\star \in \mathcal{L}^1(\hat{\mathbb{P}})$ through

$$f_r^\star(z) = \frac{r^{1/2}}{(1+\kappa(\varepsilon(r)))^{1/2}} \cdot \frac{\ell(z) - \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]}{\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}}.$$

By construction, we may thus conclude that

$$|f_r^\star(Z)| \leq r^{1/2} \cdot \frac{|\ell(Z) - \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]|}{\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}} \leq \varepsilon(r) \;\; \hat{\mathbb{P}}\text{-a.s.} \tag{8.5}$$

for every $r \in \mathbb{R}_+$, where the two inequalities follow from the definitions of $f_r^\star(z)$ and $\varepsilon(r)$, respectively. In addition, we have $\mathbb{E}_{\hat{\mathbb{P}}}[f_r^\star(Z)] = 0$ and

$$\mathbb{E}_{\hat{\mathbb{P}}}[\phi(1 + f_r^\star(Z))] \leq (1+\kappa(\varepsilon(r))) \cdot \mathbb{E}_{\hat{\mathbb{P}}}[f_r^\star(Z)^2)] = r$$

for all sufficiently small r. The inequality in the above expression follows from (8.5) and from the upper bound on ϕ in (8.3), which holds for all sufficiently small ε. The equality exploits the definition of $f_r^\star$. This shows that $f_r^\star$ constitutes a feasible solution for the maximization problem in (8.4) if r is sufficiently small. Substituting $f_r^\star$ into (8.4) then yields the desired lower bound. Indeed, we have

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] &\geq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z) f_r^\star(Z)] \\
&= \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \frac{r^{1/2}}{(1+\kappa(\varepsilon(r)))^{1/2}} \cdot \frac{\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)(\ell(Z) - \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)])]}{\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}} \\
&= \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2} + o(r^{1/2}),
\end{aligned}$$

for all sufficiently small r, where the first equality follows from the definition of $f_r^\star$. The second equality exploits the Taylor expansion of the inverse square root function around 1 and the elementary observation that $\lim_{r\downarrow 0} \kappa(\varepsilon(r)) = 0$.

Step 2. The Huber loss $h_\varepsilon : \mathbb{R} \to \mathbb{R}$ with tuning parameter $\varepsilon > 0$ is defined by

$$h_\varepsilon(s) = \begin{cases} \frac{1}{2}s^2 & \text{if } |s| \leq \varepsilon, \\ \varepsilon|s| - \frac{1}{2}\varepsilon^2 & \text{otherwise.} \end{cases}$$

By construction, h_ε is continuously differentiable, depends quadratically on s if $|s| \leq \varepsilon$ and depends linearly on s if $|s| > \varepsilon$. Its conjugate $h_\varepsilon^* : \mathbb{R} \to \overline{\mathbb{R}}$ satisfies

$$h_\varepsilon^*(t) = \begin{cases} \frac{1}{2}t^2 & \text{if } |t| \leq \varepsilon, \\ \infty & \text{otherwise.} \end{cases}$$

The lower bound on ϕ in (8.3) and the convexity of ϕ imply that

$$\phi(s) \geq \frac{2}{1+\kappa(\varepsilon)} h_\varepsilon(s-1) \quad \text{for all } s \in \mathbb{R}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

whenever ε is sufficiently small. This uniform lower bound on ϕ in terms of h_ε gives rise to a uniform upper bound on ϕ^* in terms of h_ε^*. Indeed, we have

$$\begin{aligned}\phi^*(t) &\leq \sup_{s\in\mathbb{R}}\; st - \frac{2}{1+\kappa(\varepsilon)} h_\varepsilon(s-1) \\ &= t + \frac{2}{1+\kappa(\varepsilon)} h_\varepsilon^*\big(\tfrac{1}{2}t(1+\kappa(\varepsilon))\big) \\ &= t + \begin{cases} \dfrac{(1+\kappa(\varepsilon))t^2}{4} & \text{if } |t| \leq \dfrac{2\varepsilon}{1+\kappa(\varepsilon)}, \\ \infty & \text{otherwise,} \end{cases} \end{aligned} \tag{8.6}$$

for all sufficiently small ε. The first equality in (8.6) is obtained by applying the variable transformation $s \leftarrow s-1$ and by extracting the constant $2/(1+\kappa(\varepsilon))$ from the supremum. The second equality follows from the definition of h_ε^*. By weak duality as established in Theorem 4.15, we then find

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] &\leq \inf_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} \lambda_0 + \lambda r + \mathbb{E}_{\hat{\mathbb{P}}}[(\phi^*)^\pi(\ell(Z)-\lambda_0,\lambda)] \\ &\leq \begin{cases} \inf\limits_{\lambda_0\in\mathbb{R},\lambda\in\mathbb{R}_+} & \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \lambda r + \dfrac{1+\kappa(\varepsilon(r))}{4\lambda}\mathbb{E}_{\hat{\mathbb{P}}}[(\ell(Z)-\lambda_0)^2] \\ \text{s.t.} & \hat{\mathbb{P}}\left(|\ell(Z)-\lambda_0| \leq \dfrac{2\varepsilon(r)\lambda}{1+\kappa(\varepsilon(r))}\right) = 1, \end{cases} \end{aligned} \tag{8.7}$$

where the second inequality follows from the definition of the perspective function and from (8.6), which holds for all sufficiently small ε. Here we have re-used the function $\varepsilon(r)$ introduced in Step 1. Next, we set $\lambda_0^\star = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]$ and define

$$\lambda_r^\star = \frac{(1+\kappa(\varepsilon(r)))^{1/2}}{2r^{1/2}} \cdot \mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}$$

for any $r > 0$. Note that $(\lambda_0^\star, \lambda_r^\star)$ is feasible in (8.7) provided that r is sufficiently small; in particular, r must be small enough to satisfy $\kappa(\varepsilon(r)) \leq 3$. Indeed, we have

$$\begin{aligned}&\hat{\mathbb{P}}\left(|\ell(Z)-\lambda_0^\star| \leq \frac{2\varepsilon(r)\lambda_r^\star}{1+\kappa(\varepsilon(r))}\right) \\ &\quad= \hat{\mathbb{P}}\left(|\ell(Z)-\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]| \leq \frac{\varepsilon(r)\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}}{r^{1/2}(1+\kappa(\varepsilon(r)))^{1/2}}\right) \\ &\quad= \hat{\mathbb{P}}\left(|\ell(Z)-\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]| \leq \frac{2}{(1+\kappa(\varepsilon(r)))^{1/2}} \operatorname{ess\,sup}_{\hat{\mathbb{P}}}[|\ell(Z)-\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)]|]\right) \\ &\quad= 1, \end{aligned}$$

where the first equality follows from the definitions of $\lambda_0^\star$ and $\lambda_r^\star$, the second equality follows from the definition of $\varepsilon(r)$, and the last equality holds because

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$\kappa(\varepsilon(r)) \leq 3$. Substituting $(\lambda_0^\star, \lambda_r^\star)$ into (8.7) then yields the desired upper bound:

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] &\leq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \lambda_r^\star r + \frac{1+\kappa(\varepsilon(r))}{4\lambda_r^\star}\mathbb{E}_{\hat{\mathbb{P}}}\big[(\ell(Z)-\lambda_0^\star)^2\big] \\ &= \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \frac{(1+\kappa(\varepsilon(r)))^{1/2}}{2}\cdot r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2} \\ &\quad + \frac{(1+\kappa(\varepsilon(r)))^{1/2}}{2\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}} r^{1/2}\mathbb{E}_{\hat{\mathbb{P}}}\big[(\ell(Z)-\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)])^2\big] \\ &= \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2} + o(r^{1/2})\end{aligned}$$

Here the first equality follows from the definitions of $\lambda_0^\star$ and $\lambda_r^\star$, and the second equality holds because $\lim_{r\downarrow 0}\kappa(\varepsilon(r)) = 0$. Hence the claim follows. □

Theorem 8.4 reveals that, up to leading order in r, robustification with respect to a restricted divergence ambiguity set is equivalent to variance regularization. The requirement that the loss must be almost surely bounded is restrictive but necessary. However, it can be relaxed if the entropy function ϕ grows superlinearly. As an example, recall from Proposition 6.12 that the worst-case expectation of a linear loss function with respect to a Kullback–Leibler ambiguity set centred at a Gaussian distribution equals precisely $\mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + (2r)^{1/2}\mathbb{V}_{\hat{\mathbb{P}}}[\ell(Z)]^{1/2}$ without any higher-order error terms. This formula is consistent with Theorem 8.4 because the entropy function of the Kullback–Leibler divergence satisfies $\phi''(1) = 1$. Thus it must be scaled by 2 to satisfy Assumption 8.3. Note that any (non-constant) linear loss functions fails to be $\hat{\mathbb{P}}$-almost surely bounded with respect to any (non-degenerate) Gaussian distribution $\hat{\mathbb{P}}$. However, the conclusions of Theorem 8.4 hold nevertheless because the underlying entropy function grows faster than linearly.

A Taylor expansion akin to (8.2) for *empirical* reference distributions and for the Kullback–Leibler divergence ambiguity set (2.13) is due to Lam (2019). Duchi *et al.* (2021) generalize this result to other ϕ-divergences. Similar results for empirical reference distributions are also reported by Lam (2016, 2018), Duchi and Namkoong (2019) and Blanchet and Shapiro (2024) in different contexts. In a parallel line of research, Gotoh, Kim and Lim (2018, 2021) derive a Taylor expansion of the penalty-based worst-case expected loss $\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})}\mathbb{E}_{\mathbb{P}}[\ell(Z)] - \frac{1}{r}\mathrm{D}_\phi(\mathbb{P},\hat{\mathbb{P}})$. They focus again on discrete empirical reference distributions and provide both first- and higher-order terms of the corresponding Taylor expansion.

Maurer and Pontil (2009) show that variance-regularized empirical risk minimization may provide faster rates of convergence to the expected loss under the population distribution compared to standard empirical risk minimization. This improved convergence highlights the potential benefits of incorporating variance regularization in the learning process. Unfortunately, simple stochastic optimization problems with a mean–variance objective are NP-hard even if the underlying loss function is convex in the decision variables (Ahmed 2006). In contrast, the worst-case expectation with respect to any ambiguity set preserves the convexity of

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

the underlying loss function. Theorem 8.4 thus suggests that the worst-case expected loss over a restricted ϕ-divergence ambiguity set provides a convex surrogate for the non-convex – but statistically attractive – variance-regularized empirical loss.

8.2. *Wasserstein ambiguity sets*

As a motivating example, we show that robustification with respect to a 1-Wasserstein ambiguity set is closely related to Lipschitz regularization. To see this, recall first that the p-Wasserstein ambiguity set (2.28) for $p \in [1, \infty)$ is defined as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) \colon \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}.$$

Here $\mathcal{Z}$ is a closed support set, $r \in \mathbb{R}_+$ is a size parameter, W_p is the p-Wasserstein distance induced by a norm $\|\cdot\|$ on $\mathbb{R}^d$ (see Definition 2.18) and $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$ is a reference distribution. If the loss function ℓ is piecewise concave, then the worst-case expectation problem (4.1) over $\mathcal{P}$ can be reformulated as a finite convex program (see Theorem 7.20). For more general loss functions, however, exact reformulations of (4.1) are unavailable. We now show that if $p = 1$ and ℓ is Lipschitz-continuous as well as $\hat{\mathbb{P}}$-integrable, then the worst-case expectation problem (4.1) admits a simple upper bound that involves the Lipschitz modulus of ℓ.

Proposition 8.5 (Lipschitz regularization). Suppose that $\mathcal{P}$ is the 1-Wasserstein ambiguity set of radius $r \in \mathbb{R}_+$ around $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$, and W_1 is induced by a norm $\|\cdot\|$ on $\mathbb{R}^d$. In addition, suppose that ℓ is Lipschitz-continuous on $\mathcal{Z}$ with respect to the same norm $\|\cdot\|$ and that $\mathbb{E}_{\hat{\mathbb{P}}}[|\ell(Z)|] < \infty$. Then we have

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r \cdot \mathrm{lip}(\ell). \tag{8.8}$$

We emphasize that evaluating the Lipschitz modulus of a generic loss function is computationally challenging. For example, one can show that computing $\mathrm{lip}(\ell)$ is NP-hard even if $\|\cdot\|$ is the ∞-norm and even if ℓ is a (convex) conic quadratic loss function; see e.g. Kuhn *et al.* (2019, Remark 3) for a simple proof.

Proof of Proposition 8.5. The Kantorovich–Rubinstein duality implies that

$$\begin{aligned}\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] &= \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + \mathrm{lip}(\ell) \cdot \left(\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\frac{\ell(Z)}{\mathrm{lip}(\ell)}\right] - \mathbb{E}_{\hat{\mathbb{P}}}\left[\frac{\ell(Z)}{\mathrm{lip}(\ell)}\right] \right) \\ &\leq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r \cdot \mathrm{lip}(\ell).\end{aligned}$$

Indeed, the normalized function $\ell/\mathrm{lip}(\ell)$ is Lipschitz-continuous and has Lipschitz modulus at most 1. By Corollary 2.19, we thus have for every $\mathbb{P} \in \mathcal{P}$ that

$$\mathbb{E}_{\mathbb{P}}\left[\frac{\ell(Z)}{\mathrm{lip}(\ell)}\right] - \mathbb{E}_{\hat{\mathbb{P}}}\left[\frac{\ell(Z)}{\mathrm{lip}(\ell)}\right] \leq \mathrm{W}_1(\mathbb{P}, \hat{\mathbb{P}}) \leq r.$$

Therefore the claim follows. □

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Close connections between Wasserstein distributionally robust optimization and Lipschitz regularization have been discovered in different contexts (Mohajerin Esfahani and Kuhn 2018, Shafieezadeh-Abadeh *et al.* 2015, 2019, Gao *et al.* 2024*b*). Recall that the upper bound in (8.8) is tight. Indeed, Proposition 6.17 implies that (8.8) collapses to an equality if ℓ is convex and $\mathcal{Z} = \mathbb{R}^d$. The Lipschitz modulus of the loss function encodes its variability. Thus the Lipschitz regularization term in (8.8) penalizes loss functions that display a high degree of variability. In the following we will derive generalized variation regularization bounds akin to (8.8) for worst-case expectation problems over p-Wasserstein ambiguity sets for $p \in \mathbb{N}$.

Toward this goal, for any $k \in \mathbb{Z}_+$ we use $D^k\ell(\hat{z})$, to denote the totally symmetric tensor of all kth-order partial derivatives of $\ell(z)$ at $z = \hat{z}$. Accordingly, $D^k\ell(\hat{z})[z_1, \ldots, z_k]$ stands for the directional derivative of $\ell(z)$ along the directions $z_i \in \mathbb{R}^d$ for $i \in [k]$. If $z_i = z$ for all $i \in [k]$, then we use $D^k\ell(\hat{z})[z]^k$ as shorthand for $D^k\ell(\hat{z})[z, \ldots, z]$. Any norm $\|\cdot\|$ on $\mathbb{R}^d$ induces a norm on the space of totally symmetric kth-order tensors through

$$\begin{aligned}\|D^k\ell(\hat{z})\| &= \sup_{z_1,\ldots,z_k\in\mathbb{R}^d} \{|D^k\ell(\hat{z})[z_1, \ldots, z_k]| : \|z_i\| \le 1 \ \ \forall i \in [k]\} \\ &= \sup_{z\in\mathbb{R}^d} \{|D^k\ell(\hat{z})[z]^k| : \|z\| \le 1\},\end{aligned}$$

where the second equality exploits the symmetry of $D^k\ell(\hat{z})$ (Banach 1938, Satz 1). By slight abuse of notation, we use the same symbol $\|\cdot\|$ for the tensor norm as for the underlying vector norm $\|\cdot\|$. The following theorem generalizes Proposition 8.5 to any $p \in \mathbb{N}$. This result is due to Shafiee *et al.* (2023, Theorem 3.2).

Theorem 8.6 (Variation and Lipschitz regularization). If $\mathcal{P}$ is the p-Wasserstein ambiguity set (2.28) for some $p \in \mathbb{N}$, where W_p is induced by a norm $\|\cdot\|$ on $\mathbb{R}^d$, $\mathcal{Z}$ is convex and ℓ is $p-1$ times continuously differentiable, then we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \le \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + \sum_{k=1}^{p-1} \frac{r^k}{k!} \mathbb{E}_{\hat{\mathbb{P}}}[\|D^k\ell(\hat{Z})\|^{q_k}]^{1/q_k} + \frac{r^p}{p!} \operatorname{lip}(D^{p-1}\ell),$$

where $p_k = p/k$ and $q_k = p/(p-k)$ for all $k \in [p-1]$.

Proof. Select any $\mathbb{P} \in \mathcal{P}$ and any optimal coupling $\gamma^\star \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$ with $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) = \mathbb{E}_{\gamma^\star}[\|Z - \hat{Z}\|^p]^{1/p}$, which exists by Lemma 3.17. As $\gamma^\star \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$, we have

$$\mathbb{E}_{\mathbb{P}}[\ell(Z)] - \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] = \mathbb{E}_{\gamma^\star}[\ell(Z) - \ell(\hat{Z})].$$

By Krantz and Parks (2002, Theorem 2.2.5), we can expand $\ell(z) - \ell(\hat{z})$ as a Taylor series with Lagrange remainder. Thus there exists a Borel function $f : \mathcal{Z}\times\mathcal{Z} \to \mathcal{Z}$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

that maps any pair $(z, \hat{z})$ to a point on the line segment between z and $\hat{z}$ such that

$$\ell(z) - \ell(\hat{z}) = \sum_{k=1}^{p-1} \frac{1}{k!} D^k \ell(\hat{z})[z - \hat{z}]^k + \frac{1}{p!} D^p \ell(f(z, \hat{z}))[z - \hat{z}]^p$$
$$\leq \sum_{k=1}^{p-1} \frac{1}{k!} \|D^k \ell(\hat{z})\| \|z - \hat{z}\|^k + \frac{1}{p!} \|D^p \ell(f(z, \hat{z}))\| \|z - \hat{z}\|^p. \qquad (8.9)$$

The inequality in (8.9) follows from the definition of the tensor norm. By Hölder's inequality, the expected value of the kth term in (8.9) with respect to $\gamma^\star$ satisfies

$$\mathbb{E}_{\gamma^\star}[\|D^k \ell(\hat{Z})\| \|Z - \hat{Z}\|^k] \leq \mathbb{E}_{\gamma^\star}[\|Z - \hat{Z}\|^{k p_k}]^{1/p_k} \mathbb{E}_{\gamma^\star}[\|D^k \ell(\hat{Z})\|^{q_k}]^{1/q_k}$$
$$\leq r^k \mathbb{E}_{\hat{\mathbb{P}}}[\|D^k \ell(\hat{Z})\|^{q_k}]^{1/q_k},$$

where $p_k = p/k$ and $q_k = p/(p-k)$ represent conjugate exponents. The second inequality in the above expression holds because $\gamma^\star \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$, which implies that

$$\mathbb{E}_{\gamma^\star}[\|Z - \hat{Z}\|^{k p_k}]^{1/p_k} = \mathbb{E}_{\gamma^\star}[\|Z - \hat{Z}\|^p]^{k/p} = \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}})^k \leq r^k.$$

As $\mathcal{Z}$ is convex, we may conclude that $f(z, \hat{z}) \in \mathcal{Z}$ for all $z, \hat{z} \in \mathcal{Z}$. Thus the expected value of the Lagrange remainder in (8.9) with respect to $\gamma^\star$ satisfies

$$\mathbb{E}_{\gamma^\star}[\|D^p \ell(f(Z, \hat{Z}))\| \|Z - \hat{Z}\|^p] \leq \sup_{\hat{z} \in \mathcal{Z}} \|D^p \ell(\hat{z})\| \; \mathbb{E}_{\gamma^\star}[\|Z - \hat{Z}\|^p]$$
$$\leq r^p \sup_{\hat{z} \in \mathcal{Z}} \|D^p \ell(\hat{z})\|$$
$$\leq r^p \operatorname{lip}(D^{p-1} \ell),$$

where the second inequality again exploits Hölder's inequality and the properties of the optimal coupling $\gamma^\star$. The third inequality follows from the mean value theorem. The desired inequality is finally obtained by combining the upper bounds on the expected values of all terms in (8.9) with respect to $\gamma^\star$. □

Theorem 8.6 shows that the worst-case expected loss over a p-Wasserstein ball is bounded above by the sum of the expected loss under the reference distribution, $p-1$ variation regularization terms, and a Lipschitz regularization term. Note that $p_1 = p$ and $q = q_1 = p/(p-1)$ are Hölder conjugates and that $D^1 \ell = \nabla \ell$. Thus the term corresponding to $k = 1$ in the upper bound of Theorem 8.6 can be expressed more explicitly as $\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla \ell(\hat{Z})\|^q]^{1/q}$. The next theorem, which is adapted from Bartl, Drapeau, Oblój and Wiesel (2021) and Gao *et al.* (2024*b*), reveals that this variation regularizer matches the leading term of a Taylor expansion of the worst-case expected loss in the radius r of the p-Wasserstein ball for any $p > 1$.

Theorem 8.7 (Taylor expansion of worst-case expectation). Suppose that $\mathcal{P}$ is the p-Wasserstein ambiguity set (2.28) for some $p \geq 1$, where W_p is induced by a norm $\|\cdot\|$ on $\mathbb{R}^d$, and $\mathcal{Z}$ is convex. Suppose also that the following hold.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

(i) *Growth condition.* There exist $g, \delta_0 > 0$ such that $\ell(z) - \ell(\hat{z}) \leq g\|z - \hat{z}\|^p$ for all $z, \hat{z} \in \mathcal{Z}$ with $\|z - \hat{z}\| > \delta_0$.

(ii) *Smoothness condition.* There exists $L > 0$ such that $\|\nabla\ell(z) - \nabla\ell(\hat{z})\|_* \leq L\|z - \hat{z}\|$ for all $z, \hat{z} \in \mathcal{Z}$, where $\|\cdot\|_*$ is the norm dual to $\|\cdot\|$.

(iii) *Integrability condition.* Both $\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]$ and $\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^{2q-2}]$ are finite, where $q = p/(p-1)$ is the Hölder conjugate of p.

Then we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(Z)] + r \cdot \mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(Z)\|_*^q]^{1/q} + o(r). \tag{8.10}$$

Recall that all norms on $\mathbb{R}^d$ are topologically equivalent. Thus, in the smoothness condition we could equivalently use the primal norm instead of the dual norm to measure differences between gradients. However, working with the dual norm is more convenient and will simplify the proof of Theorem 8.7.

Proof of Theorem 8.7. For any fixed $\delta \in \mathbb{R}_+$ and $\hat{z} \in \mathcal{Z}$, we define the variation of the loss function ℓ over a norm ball of radius δ around $\hat{z}$ as

$$V_\delta(\hat{z}) = \sup_{z\in\mathcal{Z}}\{\ell(z) - \ell(\hat{z})\colon \|z - \hat{z}\| \leq \delta\}.$$

Note that $V_\delta(\hat{z})$ is finite because ℓ is continuous thanks to the smoothness condition. As a preparation to prove the theorem, we first establish simple upper and lower bounds on $V_\delta(\hat{z})$. As $\mathcal{Z}$ is convex, the line segment from $\hat{z}$ to any $z \in \mathcal{Z}$ is contained in $\mathcal{Z}$. The mean value theorem then implies that there exists a point $\bar{z} \in \mathcal{Z}$ on this line segment that satisfies $\ell(z) - \ell(\hat{z}) = \nabla\ell(\bar{z})^\top(z - \hat{z})$. Thus we have

$$\begin{aligned}|\ell(z) - \ell(\hat{z}) - \nabla\ell(\hat{z})^\top(z - \hat{z})| &= |\nabla\ell(\bar{z})^\top(z - \hat{z}) - \nabla\ell(\hat{z})^\top(z - \hat{z})| \\ &\leq \|\nabla\ell(\bar{z}) - \nabla\ell(\hat{z})\|_*\|z - \hat{z}\| \\ &\leq L\|z - \hat{z}\|^2,\end{aligned}$$

where the two inequalities follow from the definition of the dual norm and from the smoothness condition, respectively. This implies that

$$\nabla\ell(\hat{z})^\top(z - \hat{z}) - L\|z - \hat{z}\|^2 \leq \ell(z) - \ell(\hat{z}) \leq \nabla\ell(\hat{z})^\top(z - \hat{z}) + L\|z - \hat{z}\|^2. \tag{8.11}$$

The first inequality in (8.11) gives rise to a lower bound on $V_\delta(\hat{z})$. Indeed, we find

$$\begin{aligned}V_\delta(\hat{z}) &\geq \sup_{z\in\mathcal{Z}}\{\nabla\ell(\hat{z})^\top(z - \hat{z}) - L\|z - \hat{z}\|^2\colon \|z - \hat{z}\| \leq \delta\} \\ &\geq \sup_{z\in\mathcal{Z}}\{\nabla\ell(\hat{z})^\top(z - \hat{z})\colon \|z - \hat{z}\| \leq \delta\} - L\delta^2 \\ &= \|\nabla\ell(\hat{z})\|_*\delta - L\delta^2,\end{aligned} \tag{8.12}$$

where the equality follows from the definition of the dual norm. Similarly, the second inequality in (8.11) gives rise to the following upper bound on $V_\delta(\hat{z})$:

$$V_\delta(\hat{z}) \leq \|\nabla\ell(\hat{z})\|_*\delta + L\delta^2 \quad \text{for all } \delta \in \mathbb{R}_+ \tag{8.13}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

This upper bound grows quadratically with δ and is therefore too loose for our purposes if $p < 2$. In this case, we must establish an alternative upper bound that grows only as δ^p. This is possible thanks to the growth condition on ℓ. To see this, define the worst-case variation of ℓ over any ball of radius δ_0 as

$$\overline{V} = \sup\{\ell(z) - \ell(\hat{z}) \colon z, \hat{z} \in \mathcal{Z},\ \|z - \hat{z}\| \leq \delta_0\}.$$

One can show that $\overline{V}$ is finite. If $\mathcal{Z}$ is compact, then this is a consequence of Weierstrass's maximum theorem, which applies because ℓ is continuous. If $\mathcal{Z}$ is unbounded, on the other hand, then this is a consequence of the convexity of $\mathcal{Z}$ and the growth condition on ℓ. In this case, there exists a recession direction d of $\mathcal{Z}$ with $\|d\| = 2\delta_0$. Thus, for all $z, \hat{z} \in \mathcal{Z}$ with $\|z - \hat{z}\| \leq \delta$, we have

$$\begin{aligned}\ell(z) - \ell(\hat{z}) &\leq |\ell(z) - \ell(z + d)| + |\ell(z + d) - \ell(\hat{z})| \\ &\leq g\|d\|^p + g\|z + d - \hat{z}\|^p \\ &\leq g((2\delta_0)^p + (3\delta_0)^p).\end{aligned}$$

The second inequality follows from the growth condition on ℓ and the estimates $\|z - (z + d)\| = \delta_0$ and $\|(z + d) - \hat{z}\| \geq \|d\| - \|z - \hat{z}\| \geq \delta_0$. Thus $\ell(z) - \ell(\hat{z})$ admits a finite upper bound independent of z and $\hat{z}$, which confirms that $\overline{V}$ is finite.

The growth condition on ℓ ensures that $V_\delta(\hat{z}) \leq \max\{\overline{V}, g\delta^p\}$. Combining this estimate with (8.13) and defining $u(\delta) = \min\{\max\{\overline{V}, g\delta^p\}, L\delta^2\}$ yields

$$V_\delta(\hat{z}) \leq \min\{\max\{\overline{V}, g\delta^p\}, \|\nabla\ell(\hat{z})\|_*\delta + L\delta^2\} \leq \|\nabla\ell(\hat{z})\|_*\delta + u(\delta).$$

Note that $u(\delta) = g\delta^p$ for all sufficiently large δ and $u(\delta) = L\delta^2$ for all sufficiently small δ. In between there is a (possibly empty) interval on which $u(\delta) = \overline{V}$ is constant. Since $p \leq 2$, in all three regimes, $u(\delta)$ can be bounded above by $g'\delta^p$ for some growth parameter $g' \in \mathbb{R}_+$. Setting G to the largest of these three growth parameters, we may thus conclude that

$$V_\delta(\hat{z}) \leq \|\nabla\ell(\hat{z})\|_*\delta + G\delta^p \quad \text{for all } \delta \in \mathbb{R}_+. \tag{8.14}$$

Thus, if $p \leq 2$, then $V_\delta(\hat{z})$ admits an upper bound that grows only as δ^p.

The remainder of the proof proceeds in two steps. First, we show that the right-hand side of (8.10) provides a *lower* bound on the worst-case expected loss over $\mathcal{P}$ (Step 1). Next, we show that the right-hand side of (8.10) also provides an *upper* bound on the worst-case expected loss over $\mathcal{P}$ (Step 2). This will prove the claim.

Step 1. Define $\mathcal{F}$ as the family of all Borel functions $f \colon \mathcal{Z} \to \mathcal{Z}$. Any $f \in \mathcal{F}$ induces a pushforward distribution $\mathcal{P} = \hat{\mathbb{P}} \circ f^{-1}$ supported on $\mathcal{Z}$. By restricting the Wasserstein ball around $\hat{\mathbb{P}}$ to contain only such pushforward distributions, we find

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \geq \sup_{f\in\mathcal{F}} \{\mathbb{E}_{\hat{\mathbb{P}}}[\ell(f(\hat{Z}))] \colon \mathbb{E}_{\hat{\mathbb{P}}}[\|f(\hat{Z}) - \hat{Z}\|^p] \leq r^p\} \tag{8.15a}$$

$$\geq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + \sup_{\delta\in\Delta} \big\{\mathbb{E}_{\hat{\mathbb{P}}}\big[V_{\delta(\hat{Z})}(\hat{Z})\big] \colon \mathbb{E}_{\hat{\mathbb{P}}}[\delta(\hat{Z})^p] \leq r^p\big\}, \tag{8.15b}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where the set Δ in (8.15b) represents the family of all Borel functions $\delta\colon \mathcal{Z} \to \mathbb{R}_+$. The second inequality in the above expression can be justified as follows. Select any $\delta \in \Delta$ feasible in (8.15b), and define $f \in \mathcal{F}$ as any Borel function satisfying

$$f(\hat{z}) \in \arg\max_{z\in\mathcal{Z}} \{\ell(z)\colon \|z - \hat{z}\| \le \delta(\hat{z})\} \quad \text{for all } \hat{z} \in \mathcal{Z}.$$

Such a Borel function exists thanks to Rockafellar and Wets (2009, Corollary 14.6 and Theorem 14.37). As δ is feasible in (8.15b), this function f satisfies

$$\mathbb{E}_{\hat{\mathbb{P}}}[\|f(\hat{Z}) - \hat{Z}\|^p] \le \mathbb{E}_{\hat{\mathbb{P}}}[\delta(\hat{Z})^p] \le r^p$$

and is thus feasible in (8.15a). Its objective function value in (8.15a) satisfies

$$\mathbb{E}_{\hat{\mathbb{P}}}[\ell(f(\hat{Z}))] = \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + \mathbb{E}_{\hat{\mathbb{P}}}[V_{\delta(\hat{Z})}(\hat{Z})].$$

Hence any feasible solution in (8.15b) gives rise to a feasible solution in (8.15a) with the same objective function value. This proves the inequality in (8.15a). Substituting the lower bound (8.12) on $V_\delta(\hat{z})$ into (8.15b) then yields the estimate

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \ge \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + \begin{cases} \sup\limits_{\delta\in\Delta} & \mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*\delta(\hat{Z}) - L\delta(\hat{Z})^2] \\ \text{s.t.} & \mathbb{E}_{\hat{\mathbb{P}}}[\delta(\hat{Z})^p] \le r^p. \end{cases} \tag{8.16}$$

If $\|\nabla\ell(\hat{Z})\|_* = 0$ $\hat{\mathbb{P}}$-almost surely, then we have established the desired lower bound. From now on we may thus assume that $\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*] > 0$. Next, we construct a function $\delta^\star \in \Delta$ feasible in the maximization problem in (8.16) and use its objective function value as a lower bound on the problem's supremum. Specifically, we set

$$\delta^\star(\hat{z}) = \frac{\|\nabla\ell(\hat{z})\|_*^{q-1} r}{\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/p}} \quad \text{for all } \hat{z} \in \mathcal{Z},$$

which is well-defined by the integrability condition. As $q - 1 = q/p$, we find

$$\mathbb{E}_{\hat{\mathbb{P}}}[\delta^\star(\hat{Z})^p] = r^p \quad \text{and} \quad \mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*\delta^\star(\hat{Z})^p] = r \cdot \mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q}.$$

Hence $\delta^\star$ is feasible in (8.16), and its objective function value amounts to

$$\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*\delta^\star(\hat{Z}) - L\delta^\star(\hat{Z})^2] = r \cdot \mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q} - Lr^2 \cdot \frac{\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^{2q-2}]}{\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{2/p}}.$$

Note that the last term is again finite thanks to the integrability condition. Substituting this expression back into (8.16) yields the desired lower bound

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \ge \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + r \cdot \mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q} + O(r^2).$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Step 2. By strong duality as established in Theorem 4.18, we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] = \inf_{\lambda\in\mathbb{R}_+} \lambda r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}} \ell(z) - \lambda\|z-\hat{Z}\|^p\right]$$
$$= \inf_{\lambda\in\mathbb{R}_+} \lambda r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\ell(\hat{Z}) + \sup_{\delta\in\mathbb{R}_+} V_\delta(\hat{Z}) - \lambda\delta^p\right], \tag{8.17}$$

where the second equality follows from the observation that

$$\sup_{z\in\mathcal{Z}} \ell(z) - \lambda\|z-\hat{z}\|^p = \sup_{z\in\mathcal{Z}} \sup_{\delta\in\mathbb{R}_+} \{\ell(z) - \lambda\delta^p : \|z-\hat{z}\| \le \delta\}$$
$$= \ell(\hat{z}) + \sup_{\delta\in\mathbb{R}_+} V_\delta(\hat{z}) - \lambda\delta^p.$$

Next, we construct an upper bound on (8.17). In fact, we need separate constructions for $p > 2$ and $p \le 2$. Assume first that $p > 2$. In this case, we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] - \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})]$$
$$\le \inf_{\lambda_1,\lambda_2\in\mathbb{R}_+} (\lambda_1+\lambda_2) r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{\delta\in\mathbb{R}_+} \|\nabla\ell(\hat{Z})\|_*\delta + L\delta^2 - (\lambda_1+\lambda_2)\delta^p\right]$$
$$\le \inf_{\lambda_1\in\mathbb{R}_+} \lambda_1 r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{\delta\in\mathbb{R}_+} \|\nabla\ell(\hat{Z})\|_*\delta - \lambda_1\delta^p\right] \tag{8.18a}$$
$$+ \inf_{\lambda_2\in\mathbb{R}_+} \lambda_2 r^p + \sup_{\delta\in\mathbb{R}_+} L\delta^2 - \lambda_2\delta^p, \tag{8.18b}$$

where the first inequality follows from the estimate (8.13), and the second inequality holds because the supremum over δ is duplicated. The resulting upper bound on the worst-case expected loss thus coincides with the sum of two infima. One readily verifies that the maximization problem over δ in (8.18a) is solved by $\delta^\star = (p\lambda_1)^{-q/p}\|\nabla\ell(\hat{Z})\|_*^{q/p}$. Thus the infimum in (8.18a) equals

$$\inf_{\lambda_1\in\mathbb{R}_+} \lambda_1 r^p + \frac{1}{q}(\lambda_1 p)^{-q/p}\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q] = r\cdot\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q}, \tag{8.19a}$$

where the equality holds because the resulting minimization problem over λ_1 is solved by $\lambda_1^\star = pr^{-p/q}\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q}$. Similarly, the maximization problem over δ in (8.18b) is solved by $\delta^\star = C_1\lambda_2^{-1/(p-2)}$, where C_1 represents a positive constant that only depends on p and L. Thus the infimum in (8.18b) equals

$$\inf_{\lambda_2\in\mathbb{R}_+} \lambda_2 r^p + C_2\lambda_2^{-2/(p-2)} = C_3 r^2, \tag{8.19b}$$

where C_2 and C_3 are other positive constants depending on p and L. The equality in (8.19b) is obtained by solving the minimization problem over λ_2 in closed form. Replacing (8.18a) with (8.19a) and (8.18b) with (8.19b) finally yields

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \le \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + r\cdot\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q} + O(r^2).$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Assume next that $p \leq 2$. In this case, we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \inf_{\lambda_1,\lambda_2\in\mathbb{R}_+} (\lambda_1+\lambda_2)r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{\delta\in\mathbb{R}_+} \|\nabla\ell(\hat{Z})\|_*\delta + G\delta^p - (\lambda_1+\lambda_2)\delta^p\right]$$

$$\leq \inf_{\lambda_1\in\mathbb{R}_+} \lambda_1 r^p + \mathbb{E}_{\hat{\mathbb{P}}}\left[\sup_{\delta\in\mathbb{R}_+} \|\nabla\ell(\hat{Z})\|_*\delta - \lambda_1\delta^p\right] \tag{8.20a}$$

$$+ \inf_{\lambda_2\in\mathbb{R}_+} \lambda_2 r^p + \sup_{\delta\in\mathbb{R}_+} G\delta^p - \lambda_2\delta^p, \tag{8.20b}$$

where the first inequality follows from the estimate (8.14). Note that the infimum in (8.20a) is identical to that in (8.18a) and thus simplifies to (8.19a). Next, note that the maximization problem over δ in (8.20b) is unbounded unless $\lambda_2 \geq G$. This condition thus constitutes an implicit constraint for the minimization problem over λ_2. Whenever λ_2 satisfies this constraint, however, the supremum over δ evaluates to 0, and therefore the infimum over λ_2 evaluates to Gr^p. Replacing (8.20a) with (8.19a) and (8.20b) with Gr^p finally yields

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Z)] \leq \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\hat{Z})] + r\cdot\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q} + O(r^p).$$

As both $O(r^2)$ and $O(r^p)$ for $0 < p \leq 2$ are of the order $o(r)$, the claim follows. □

The proof of Theorem 8.7 reveals that the variation $V_\delta(\hat{z})$ equals $\|\nabla\ell(\hat{z})\|_*\delta$ to first order in δ. Hence it is natural to refer to the regularization term $\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla\ell(\hat{Z})\|_*^q]^{1/q}$ appearing in (8.10) as the *total variation*.

Regularizers penalizing the Lipschitz moduli, gradients, Hessians or tensors of higher-order partial derivatives are successfully used in the adversarial training of neural networks (Lyu, Huang and Liang 2015, Jakubovitz and Giryes 2018, Finlay and Oberman 2021, Bai, He, Jiang and Obloj 2023*a*) and in the stabilizing training of generative adversarial networks (Roth, Lucchi, Nowozin and Hofmann 2017, Nagarajan and Kolter 2017, Gulrajani *et al.* 2017). However, these regularizers introduce non-convexity into an otherwise convex optimization problem. Theorems 8.6 and 8.7 thus suggest that the worst-case expected loss with respect to a Wasserstein ambiguity set provides a convex surrogate for the empirical loss with Lipschitz and/or variation regularizers.

8.3. Lipschitz continuity of law-invariant convex risk measures

Let ϱ be a law-invariant convex risk measure as introduced in Section 5. Recall that all convex risk measures are translation-invariant, monotone and convex. Assume also that ϱ is an $\mathcal{L}_p$-risk measure for some $p \geq 1$. By this we mean that $\varrho_{\mathbb{P}}[\ell(Z)]$ is finite whenever $\ell \in \mathcal{L}_p(\mathbb{P})$ and $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$, that is, whenever $\mathbb{E}_{\mathbb{P}}[|\ell(Z)|^p] < +\infty$. The aim of this section is to derive interpretable and easily computable upper bounds on the worst case of $\varrho_{\mathbb{P}}[\ell(Z)]$ with respect to all distributions $\mathbb{P}$ of Z in a p-Wasserstein ball. To this end, we first recall the definition of a subgradient.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Definition 8.8 (Subgradient). If ϱ is a law-invariant convex $\mathcal{L}_p$-risk measure for some $p \geq 1$, then $h \in \mathcal{L}_q(\mathbb{P})$ is a subgradient of $\varrho_{\mathbb{P}}$ at $\ell_0 \in \mathcal{L}_p(\mathbb{P})$ if $\frac{1}{p} + \frac{1}{q} = 1$ and

$$\varrho_{\mathbb{P}}[\ell(Z)] \geq \varrho_{\mathbb{P}}[\ell_0(Z)] + \mathbb{E}_{\mathbb{P}}[h(Z) \cdot (\ell(Z) - \ell_0(Z))] \quad \text{for all } \ell \in \mathcal{L}_p(\mathbb{P}).$$

We say that $\varrho_{\mathbb{P}}$ is subdifferentiable at ℓ_0 if it has at least one subgradient at ℓ_0.

Definition 8.9 (Lipschitz continuity). Let ϱ be a law-invariant convex $\mathcal{L}_p$-risk measure for some $p \geq 1$. Then ϱ is Lipschitz-continuous if there exists $L \geq 0$ with

$$|\varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\mathbb{P}}[\ell_0(Z)]| \leq L \cdot \mathbb{E}_{\mathbb{P}}[|\ell(Z) - \ell_0(Z)|^p]^{1/p}$$

for all $\ell, \ell_0 \in \mathcal{L}_p(\mathbb{P})$, $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$. We use $\mathrm{lip}(\varrho)$ to denote the Lipschitz modulus, i.e. the smallest L with this property.

Lemma 8.10 (Subgradient bounds). Let ϱ be a law-invariant convex $\mathcal{L}_p$-risk measure and $h \in \mathcal{L}_q(\mathbb{P})$ a subgradient of $\varrho_{\mathbb{P}}$ at $\ell_0 \in \mathcal{L}_p(\mathbb{P})$ for some $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$, where $\frac{1}{p} + \frac{1}{q} = 1$. If ϱ is Lipschitz-continuous, then $\mathbb{E}_{\mathbb{P}}[|h(Z)|^q]^{1/q} \leq \mathrm{lip}(\varrho)$.

Proof. By the Lipschitz continuity of ϱ and the definition of subgradients, we have

$$\begin{aligned}
&\varrho_{\mathbb{P}}[\ell_0(Z)] + \mathrm{lip}(\varrho) \cdot \mathbb{E}_{\mathbb{P}}[|\ell(Z) - \ell_0(Z)|^p]^{1/p} \\
&\quad \geq \varrho_{\mathbb{P}}[\ell(Z)] \\
&\quad \geq \varrho_{\mathbb{P}}[\ell_0(Z)] + \mathbb{E}_{\mathbb{P}}[h(Z) \cdot (\ell(Z) - \ell_0(Z))]
\end{aligned}$$

for every $\ell \in \mathcal{L}_p(\mathbb{P})$. This inequality is equivalent to

$$\mathrm{lip}(\varrho) \geq \sup_{\substack{\ell \in \mathcal{L}_p(\mathbb{P}) \\ \ell \neq \ell_0}} \mathbb{E}_{\mathbb{P}}\left[h(Z) \cdot \frac{\ell(Z) - \ell_0(Z)}{\mathbb{E}_{\mathbb{P}}[|\ell(Z) - \ell_0(Z)|^p]^{1/p}} \right] = \mathbb{E}_{\mathbb{P}}[|h(Z)|^q]^{1/q},$$

where the equality holds because the $\mathcal{L}_q$-norm is dual to the $\mathcal{L}_p$-norm. □

The results of this section also rely on the fundamentals of comonotonicity theory, which we review next. For any Borel-measurable function $f\colon \mathbb{R}^d \to \mathbb{R}$, the distribution function $F\colon \mathbb{R} \to [0,1]$ of the random variable $f(Z)$ under $\mathbb{P}$ is defined by $F(\tau) = \mathbb{P}(f(Z) \leq \tau)$ for every $\tau \in \mathbb{R}$, and the corresponding (left) quantile function $F^{\leftarrow}\colon [0,1] \to \overline{\mathbb{R}}$ is defined by $F^{\leftarrow}(q) = \inf\{\tau \in \mathbb{R}\colon F_1(\tau) \geq q\}$ for every $q \in [0,1]$. Note that if F is invertible, then $F^{\leftarrow} = F^{-1}$. Note also that F is generally right-continuous, whereas $F^{\leftarrow}$ is generally left-continuous. The definition of the quantile function $F^{\leftarrow}$ also readily implies the equivalence

$$F(\tau) \geq q \quad \Longleftrightarrow \quad \tau \geq F^{\leftarrow}(q) \quad \text{for all } \tau \in \mathbb{R},\ q \in [0,1]. \tag{8.21}$$

Definition 8.11 (Comonotonicity). Two random variables $f(Z)$ and $g(Z)$ induced by Borel-measurable functions $f, g\colon \mathbb{R}^d \to \mathbb{R}$ are comonotonic under $\mathbb{P}$ if

$$\mathbb{P}(f(Z) \leq \tau_1 \wedge g(Z) \leq \tau_2) = \min\{F(\tau_1), G(\tau_2)\} \quad \text{for all } \tau_1, \tau_2 \in \mathbb{R},$$

where F and G denote the distribution functions of $f(Z)$ and $g(Z)$ under $\mathbb{P}$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The following proposition sheds more light on Definition 8.11. It shows that comonotonic random variables can essentially always be expressed as functions of each other (McNeil, Frey and Embrechts 2015, Corollary 5.17).

Proposition 8.12 (Comonotonicity). Let $f(Z)$ and $g(Z)$ be two random variables with respective distribution functions F and G under $\mathbb{P}$ as in Definition 8.11. If F is continuous, then $f(Z)$ and $g(Z)$ are comonotonic under $\mathbb{P}$ if and only if

$$g(Z) = G^{\leftarrow}(F(f(Z))) \ \ \mathbb{P}\text{-a.s.}$$

Proof. Note first that $F(f(Z))$ follows the standard uniform distribution on $[0, 1]$ under $\mathbb{P}$. To see this, note that for any $q \in [0, 1]$ we have

$$\mathbb{P}(F(f(Z)) \leq q) = \mathbb{P}(f(Z) \leq F^{\leftarrow}(q)) = F(F^{\leftarrow}(q)) = q,$$

where the first two equalities follow from the definitions of $F^{\leftarrow}$ and F, respectively, while the last equality holds because F is continuous.

Assume now that $f(Z)$ and $g(Z)$ are comonotonic under $\mathbb{P}$. Hence we have

$$\begin{aligned}\mathbb{P}(f(Z) \leq \tau_1 \wedge g(Z) \leq \tau_2) &= \min\{F(\tau_1), G(\tau_2)\} \\ &= \mathbb{P}(F(f(Z)) \leq \min\{F(\tau_1), G(\tau_2)\}) \\ &= \mathbb{P}(F(f(Z)) \leq F(\tau_1) \wedge F(f(Z)) \leq G(\tau_2)) \\ &= \mathbb{P}(F^{\leftarrow}(F(f(Z))) \leq \tau_1 \wedge G^{\leftarrow}(F(f(Z))) \leq \tau_2)\end{aligned}$$

for all $\tau_1, \tau_2 \in \mathbb{R}$. Here the second equality holds because $F(f(Z))$ follows the standard uniform distribution under $\mathbb{P}$. The last equality holds thanks to (8.21). As $F^{\leftarrow}(F(f(Z)))$ is $\mathbb{P}$-almost surely equal to $f(Z)$, we thus have

$$\mathbb{P}(f(Z) \leq \tau_1 \wedge g(Z) \leq \tau_2) = \mathbb{P}(f(Z) \leq \tau_1 \wedge G^{\leftarrow}(F(f(Z))) \leq \tau_2)$$

for all $\tau_1, \tau_2 \in \mathbb{R}$. Hence $(f(Z), g(Z))$ and $(f(Z), G^{\leftarrow}(F(f(Z))))$ are equal in law under $\mathbb{P}$. This implies in particular that the distribution of $g(Z)$ conditional on $f(Z)$ coincides with the distribution of $G^{\leftarrow}(F(f(Z)))$ conditional on $f(Z)$ under $\mathbb{P}$. As the latter distribution is given by the Dirac point mass at $G^{\leftarrow}(F(f(Z)))$, we may conclude that $g(Z)$ is $\mathbb{P}$-almost surely equal to $G^{\leftarrow}(F(f(Z)))$.

Assume now that $g(Z) = G^{\leftarrow}(F(f(Z)))$ $\mathbb{P}$-almost surely. Thus we have

$$\begin{aligned}\mathbb{P}(f(Z) \leq \tau_1 \wedge g(Z) \leq \tau_2) &= \mathbb{P}(f(Z) \leq \tau_1 \wedge G^{\leftarrow}(F(f(Z))) \leq \tau_2) \\ &= \min\{F(\tau_1), G(\tau_2)\},\end{aligned}$$

where the second equality follows from the first part of the proof. □

Next, we show that the correlation of two random variables with fixed marginals is maximal if they are comonotonic (McNeil *et al.* 2015, Theorem 5.25).

Theorem 8.13 (Attainable correlations). Let f, $f^\star$, g and $g^\star$ be real-valued Borel-measurable functions on $\mathbb{R}^d$. Assume that if Z is governed by $\mathbb{P}$, then $f(Z)$ and $f^\star(Z)$ have the same distribution function F, whereas $g(Z)$ and $g^\star(Z)$ have the

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

same distribution function G. If $f^\star(Z)$ and $g^\star(Z)$ are comonotonic, then

$$\mathbb{E}_{\mathbb{P}}[f(Z)\cdot g(Z)] \leq \mathbb{E}_{\mathbb{P}}[f^\star(Z)\cdot g^\star(Z)].$$

Proof. Define the joint distribution function $H\colon \mathbb{R}^2 \to [0,1]$ of $f(Z)$ and $g(Z)$ under $\mathbb{P}$ via $H(\tau_1,\tau_2) = \mathbb{P}(f(Z)\leq \tau_1 \wedge g(Z)\leq \tau_2)$ for all $\tau_1,\tau_2\in\mathbb{R}$. By McNeil *et al.* (2015, Lemma 5.24), the covariance of $f(Z)$ and $g(Z)$ under $\mathbb{P}$ satisfies

$$\operatorname{cov}_{\mathbb{P}}(f(Z),g(Z)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (H(\tau_1,\tau_2) - F(\tau_1)\,G(\tau_2))\,\mathrm{d}\tau_1\,\mathrm{d}\tau_2. \tag{8.22}$$

In addition, by the classical Fréchet bounds for copulas (McNeil *et al.* 2015, Remark 5.8), we know that $H(\tau_1,\tau_2)\leq \min\{F(\tau_1),G(\tau_2)\}$ for all $\tau_1,\tau_2\in\mathbb{R}$. As the marginal distribution functions F and G are fixed, it is evident from (8.22) that the covariance of the random variables $f(Z)$ and $g(Z)$ is maximized if their joint distribution function $H(\tau_1,\tau_2)$ coincides with its Fréchet upper bound. This, however, happens if and only if $f(Z)$ and $g(Z)$ are comonotonic under $\mathbb{P}$. We have thus shown that $\operatorname{cov}_{\mathbb{P}}(f(Z),g(Z)) \leq \operatorname{cov}_{\mathbb{P}}(f^\star(Z),g^\star(Z))$, which in turn implies that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[f(Z)\cdot g(Z)] &= \operatorname{cov}_{\mathbb{P}}(f(Z),g(Z)) + \mathbb{E}_{\mathbb{P}}[f(Z)]\cdot\mathbb{E}_{\mathbb{P}}[g(Z)]\\ &\leq \operatorname{cov}_{\mathbb{P}}(f^\star(Z),g^\star(Z)) + \mathbb{E}_{\mathbb{P}}[f^\star(Z)]\cdot\mathbb{E}_{\mathbb{P}}[g^\star(Z)]\\ &= \mathbb{E}_{\mathbb{P}}[f^\star(Z)\cdot g^\star(Z)].\end{aligned}$$

Here the inequality exploits the assumption that $f(Z)$ equals $f^\star(Z)$ in law and that $g(Z)$ equals $g^\star(Z)$ in law under $\mathbb{P}$. Hence the claim follows. □

We are now ready to show that if ϱ is a Lipschitz-continuous $\mathcal{L}_p$-risk measure and ℓ is a Lipschitz-continuous loss function, then the risk $\varrho_{\mathbb{P}}[\ell(Z)]$ is Lipschitz-continuous in the distribution $\mathbb{P}$ with respect to the p-Wasserstein distance.

Theorem 8.14 (Lipschitz continuity of risk measures). If $\ell\colon \mathbb{R}^d\to\mathbb{R}$ is a Lipschitz-continuous loss function with respect to some norm $\|\cdot\|$ on $\mathbb{R}^d$, $p\geq 1$ and ϱ a Lipschitz-continuous and law-invariant convex $\mathcal{L}_p$-risk measure, then

$$|\varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}}[\ell(\hat{Z})]| \leq \operatorname{lip}(\varrho)\cdot\operatorname{lip}(\ell)\cdot \mathrm{W}_p(\mathbb{P},\hat{\mathbb{P}})$$

for all $\mathbb{P},\hat{\mathbb{P}}\in\mathcal{P}(\mathbb{R}^d)$. Here W_p is defined with respect to $\|\cdot\|$, and $\frac{1}{p}+\frac{1}{q}=1$.

Proof. Consider an arbitrary $\mathbb{P}\in\mathcal{P}(\mathbb{R}^d)$. By Ruszczyński and Shapiro (2006, Corollary 3.1), $\varrho_{\mathbb{P}}$ is continuous and subdifferentiable on the whole Banach space $\mathcal{L}_p(\mathbb{P})$ equipped with its norm topology. The Fenchel–Moreau theorem thus implies that

$$\varrho_{\mathbb{P}}[\ell'(Z)] = \sup_{h'\in\mathcal{L}_q(\mathbb{P})} \mathbb{E}_{\mathbb{P}}[h'(Z)\cdot\ell'(Z)] - \varrho^*_{\mathbb{P}}[h'(Z)] \tag{8.23a}$$

for all $\ell'\in\mathcal{L}_p(\mathbb{P})$, where

$$\varrho^*_{\mathbb{P}}[h'(Z)] = \sup_{\ell'\in\mathcal{L}_p(\mathbb{P})} \mathbb{E}_{\mathbb{P}}[h'(Z)\cdot\ell'(Z)] - \varrho_{\mathbb{P}}[\ell'(Z)] \tag{8.23b}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

for all $h' \in \mathcal{L}_q(\mathbb{P})$ (Rockafellar 1974, Theorem 5). The relation (8.23b) defines a law-invariant convex risk measure ϱ^*. Indeed, ϱ^* is convex because pointwise suprema of affine functions are convex. In addition, ϱ^* inherits law-invariance from ϱ. Note that $h \in \mathcal{L}_q(\mathbb{P})$ attains the supremum in (8.23a) at $\ell' = \ell$ if and only if

$$\begin{aligned} & \varrho_{\mathbb{P}}[\ell(Z)] = \mathbb{E}_{\mathbb{P}}[h(Z) \cdot \ell(Z)] - \varrho^*_{\mathbb{P}}[h(Z)] \\ \iff\ & \varrho^*_{\mathbb{P}}[h(Z)] = \mathbb{E}_{\mathbb{P}}[h(Z) \cdot \ell(Z)] - \varrho_{\mathbb{P}}[\ell(Z)] \\ \iff\ & \mathbb{E}_{\mathbb{P}}[h(Z) \cdot \ell'(Z)] - \varrho_{\mathbb{P}}[\ell'(Z)] \\ & \qquad \leq \mathbb{E}_{\mathbb{P}}[h(Z) \cdot \ell(Z)] - \varrho_{\mathbb{P}}[\ell(Z)] \ \ \forall \ell' \in \mathcal{L}_p(\mathbb{P}), \end{aligned}$$

where the last equivalence follows from the definition of $\varrho^*_{\mathbb{P}}[h(Z)]$ in (8.23b). By rearranging terms, we then find that the last inequality is equivalent to

$$\varrho_{\mathbb{P}}[\ell(Z)] + \mathbb{E}_{\mathbb{P}}[h(Z) \cdot (\ell'(Z) - \ell(Z))] \leq \varrho_{\mathbb{P}}[\ell'(Z)] \ \ \forall \ell' \in \mathcal{L}_p(\mathbb{P}).$$

Thus h attains the supremum in (8.23a) at ℓ if and only if it represents a subgradient of $\varrho_{\mathbb{P}}$ at ℓ. As $\varrho_{\mathbb{P}}$ is subdifferentiable throughout $\mathcal{L}_p(\mathbb{P})$, the above reasoning implies that the supremum in (8.23a) is always attained.

Now select any $\mathbb{P}, \hat{\mathbb{P}} \in \mathcal{P}(\mathbb{R}^d)$ with $\mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) < +\infty$. We assume temporarily that $\mathbb{P}$ and $\hat{\mathbb{P}}$ are non-atomic, that is, $\mathbb{P}(Z = z) = \hat{\mathbb{P}}(Z = z) = 0$ for all $z \in \mathbb{R}^d$. Thus, for any admissible distribution function F, there exists a Borel-measurable function $f\colon \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{P}(f(Z) \leq \tau) = F(\tau)$ for all $\tau \in \mathbb{R}$; see e.g. Delage, Kuhn and Wiesemann (2019, Lemma 1). Note that non-atomicity will later be relaxed. Now also select any $h \in \mathcal{L}_q(\mathbb{P})$ that attains the supremum in (8.23a) at $\ell' = \ell$, which is guaranteed to exist. The representation (8.23a) then implies that

$$\begin{aligned} & \varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}}[\ell(\hat{Z})] \\ & \quad = \mathbb{E}_{\mathbb{P}}[h(Z) \cdot \ell(Z)] - \varrho^*_{\mathbb{P}}[h(Z)] - \sup_{\hat{h} \in \mathcal{L}_q(\hat{\mathbb{P}})} \left\{ \mathbb{E}_{\hat{\mathbb{P}}}[\hat{h}(\hat{Z}) \cdot \ell(\hat{Z})] - \varrho^*_{\hat{\mathbb{P}}}[\hat{h}(\hat{Z})] \right\}. \end{aligned}$$

In the following, we use F to denote the distribution function of $h(Z)$ under $\mathbb{P}$ and $\hat{F}$ to denote the distribution function of $\ell(\hat{Z})$ under $\hat{\mathbb{P}}$. In addition, we restrict the above maximization problem to functions $\hat{h}$ for which the distribution function of the random variable $\hat{h}(\hat{Z})$ coincides with F. As restricting the feasible set of a maximization problem leads to a lower bound on its optimal value, we find

$$\begin{aligned} & \varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}}[\ell(\hat{Z})] \\ & \quad \leq \mathbb{E}_{\mathbb{P}}[h(Z) \cdot \ell(Z)] - \begin{cases} \sup & \mathbb{E}_{\hat{\mathbb{P}}}[\hat{h}(\hat{Z}) \cdot \ell(\hat{Z})] \\ \hat{h} \in \mathcal{L}_q(\hat{\mathbb{P}}) & \\ \text{s.t.} & \hat{\mathbb{P}}(\hat{h}(\hat{Z}) \leq \tau) = F(\tau) \ \ \forall \tau \in \mathbb{R}. \end{cases} \end{aligned} \tag{8.24}$$

Here we have exploited the law-invariance of the risk measure ϱ^*, which implies that $\varrho^*_{\mathbb{P}}[h(Z)]$ and $\varrho^*_{\hat{\mathbb{P}}}[\hat{h}(\hat{Z})]$ match. Next, define the function $\hat{h}^\star\colon \mathbb{R}^d \to \mathbb{R}$ through

$$\hat{h}^\star(\hat{z}) = F^{\leftarrow}(\hat{F}(\ell(\hat{z}))) \quad \text{for all } \hat{z} \in \mathbb{R}^d.$$

Note that $\hat{F}$ is continuous because $\hat{\mathbb{P}}$ is non-atomic and ℓ is (Lipschitz-) continuous.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

By Proposition 8.12, the random variables $\hat{h}^\star(\hat{Z})$ and $\ell(\hat{Z})$ are thus comonotonic and have distribution functions F and $\hat{F}$ under $\hat{\mathbb{P}}$, respectively. Hence $\hat{h}^\star$ is feasible in the maximization problem in (8.24). In addition, by Theorem 8.13, $\hat{h}^\star$ is optimal.

Next, select any transportation plan $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$. As the marginal distributions of Z and $\hat{Z}$ under γ are given by $\mathbb{P}$ and $\hat{\mathbb{P}}$, respectively, the above implies that

$$\begin{aligned}&\varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}}[\ell(\hat{Z})] \\ &\quad\le \mathbb{E}_\gamma[h(Z)\cdot\ell(Z)] - \begin{cases} \sup & \mathbb{E}_\gamma[\hat{h}(Z,\hat{Z})\cdot\ell(\hat{Z})] \\ \hat{h}\in\mathcal{L}_q(\gamma) & \\ \text{s.t.} & \gamma(\hat{h}(Z,\hat{Z})\le\tau) = F(\tau)\ \ \forall \tau\in\mathbb{R}. \end{cases}\end{aligned} \tag{8.25}$$

Note that we have relaxed the maximization problem in (8.25) by allowing the function $\hat{h}$ to depend on both Z and $\hat{Z}$. However, this extra flexibility does not result in a higher optimal value. Indeed, Theorem 8.13 ensures that the supremum is attained by any function $\hat{h}$ for which the random variables $\hat{h}(Z,\hat{Z})$ and $\ell(\hat{Z})$ are comonotonic and for which $\hat{h}(Z,\hat{Z})$ has distribution function F. As we have seen before, there exists a function with these properties that does not depend on Z. Hence the right to select a function $\hat{h}$ that depends on Z is worthless.

Observe now that the function $\hat{h}(Z,\hat{Z}) = h(Z)$ is feasible in (8.25). Thus we find

$$\begin{aligned}\varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}}[\ell(\hat{Z})] &\le \mathbb{E}_\gamma[h(Z)\cdot\ell(Z)] - \mathbb{E}_\gamma[h(Z)\cdot\ell(\hat{Z})] \\ &\le \mathbb{E}_\gamma[h(Z)\cdot|\ell(Z)-\ell(\hat{Z})|] \\ &\le \mathbb{E}_\gamma[h(Z)\cdot \mathrm{lip}(\ell)\cdot\|Z-\hat{Z}\|] \\ &\le \mathrm{lip}(\ell)\cdot\mathbb{E}_\gamma[\|Z-\hat{Z}\|^p]^{1/p}\cdot\mathbb{E}_{\mathbb{P}}[h(Z)^q]^{1/q},\end{aligned}$$

where the second inequality holds because all convex risk measures are monotonic, which implies that the subgradient $h(Z)$ is $\mathbb{P}$-almost surely non-negative. The third inequality exploits the Lipschitz continuity of the loss function, and the fourth inequality follows from Hölder's inequality. As the resulting inequality holds for all couplings $\gamma \in \Gamma(\mathbb{P}, \hat{\mathbb{P}})$, the definition of the p-Wasserstein distance implies that

$$\begin{aligned}\varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}}[\ell(\hat{Z})] &\le \mathrm{lip}(\ell)\cdot \mathrm{W}_p(\mathbb{P},\hat{\mathbb{P}})\cdot\mathbb{E}_{\mathbb{P}}[h(Z)^q]^{1/q} \\ &\le \mathrm{lip}(\varrho)\cdot\mathrm{lip}(\ell)\cdot\mathrm{W}_p(\mathbb{P},\hat{\mathbb{P}}),\end{aligned}$$

where the second inequality follows from Lemma 8.10. The claim then follows by interchanging the roles of $\mathbb{P}$ and $\hat{\mathbb{P}}$.

Recall now that we assumed $\mathbb{P}$ and $\hat{\mathbb{P}}$ are non-atomic. This assumption was needed to show that the supremum in (8.24) is attained. In general, one can extend $\mathbb{P}$ to a distribution $\mathbb{P}'$ on $\mathbb{R}^{d+1}$ under which $(Z_1,\ldots,Z_d)$ and Z_{d+1} are independent and have marginal distributions equal to $\mathbb{P}$ and to the uniform distribution on $[0,1]$, respectively. In the same way, $\hat{\mathbb{P}}$ can be extended to a distribution $\hat{\mathbb{P}}'$ on $\mathbb{R}^{d+1}$. By construction, $\mathbb{P}'$ and $\hat{\mathbb{P}}'$ are non-atomic. As ϱ is law-invariant, we further have

$$|\varrho_{\mathbb{P}}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}}[\ell(\hat{Z})]| = |\varrho_{\mathbb{P}'}[\ell(Z)] - \varrho_{\hat{\mathbb{P}}'}[\ell(\hat{Z})]|.$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The right-hand side of this equation can now be bounded as above. □

Theorem 8.14 immediately implies the following worst-case risk bound.

Corollary 8.15. If all assumptions of Theorem 8.14 hold and

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)\colon \mathrm{W}_p(\mathbb{P}, \hat{\mathbb{P}}) \leq r\}$$

is a p-Wasserstein ball of radius $r \geq 0$ for any $p \geq 1$, then

$$\sup_{\mathbb{P}\in\mathcal{P}} \varrho_{\mathbb{P}}[\ell(Z)] \leq \varrho_{\hat{\mathbb{P}}}[\ell(Z)] + r \cdot \mathrm{lip}(\varrho) \cdot \mathrm{lip}(\ell).$$

Theorem 8.14 and Corollary 8.15 are due to Pichler (2013). Corollary 8.15 shows that the worst-case risk over all distributions in a p-Wasserstein ball is upper-bounded by the sum of the nominal risk and a Lipschitz regularization term for a broad spectrum of law-invariant convex risk measures. If the loss function ℓ is linear, that is, if $\ell(z) = \theta^\top z$ for some $\theta \in \mathbb{R}^d$, then this upper bound is often tight (Pflug *et al.* 2012, Wozabal 2014). In this case the Lipschitz modulus of ℓ simplifies to $\|\theta\|_*$. For example, the CVaR at level $\beta \in (0, 1]$ is a law-invariant convex $\mathcal{L}_p$-risk measure, and it is Lipschitz-continuous with Lipschitz modulus $\beta^{-1/p}$. Thus Corollary 8.15 applies. From Proposition 6.20 we know, however, that the upper bound is exact in this case. If additionally $p = 1$, then Proposition 6.18 implies that the upper bound remains exact whenever ℓ is convex and Lipschitz-continuous.

9. Numerical solution methods for DRO problems

The finite convex reformulations of the worst-case expectation problem (4.1) presented in Section 7 are often susceptible to standard optimization software, that is, they obviate the need for tailored algorithms. However, these reformulations can have two significant drawbacks. First, the corresponding monolithic optimization problems can become large and hence challenging to solve. Second, depending on the chosen ambiguity set, the emerging reformulations may belong to a class of optimization problems that are more difficult to solve than a deterministic version of the original problem. For instance, even if the loss function ℓ in the worst-case expectation (4.1) is piecewise affine and the support set $\mathcal{Z}$ is an ellipsoid, the finite dual reformulation over Chebyshev ambiguity sets, as provided by Theorem 7.9, results in a semidefinite program, as opposed to a numerically favourable quadratically constrained quadratic program. Both disadvantages can be alleviated by resorting to tailored algorithms, which we discuss in this section.

Most numerical methods for solving the DRO problem (1.2) address an equivalent reformulation of (1.2) obtained by dualizing the inner worst-case expectation problem. This reformulation is usually constructed by leveraging one of the strong duality theorems from Section 4. The resulting reformulation of (1.2) is thus representable as a semi-infinite program of the form

$$\inf\{f(y)\colon y \in \mathcal{Y},\ g_j(y, z_j) \leq 0\ \ \forall z_j \in \mathcal{Z},\ j \in [m]\}. \tag{9.1}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Note that (9.1) is naturally interpreted as a classical robust optimization problem.

As an example, assume that $\mathcal{P}$ is the generic moment ambiguity set (2.1) and that some mild regularity conditions hold. In this case, Theorem 4.5 implies that

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(x,Z)] = \begin{cases} \inf & \lambda_0 + \delta^*_{\mathcal{F}}(\lambda) \\ \text{s.t.} & x\in\mathcal{X}, \lambda_0\in\mathbb{R}, \lambda\in\mathbb{R}^m \\ & \lambda_0 + f(z)^\top\lambda \geq \ell(x,z) \;\; \forall z\in\mathcal{Z}. \end{cases}$$

If the support function $\delta^*_{\mathcal{F}}(\lambda)$ is known in closed form, then the resulting minimization problem becomes an instance of (9.1) with $y=(x,\lambda_0,\lambda)$, $\mathcal{Y}=\mathcal{X}\times\mathbb{R}\times\mathbb{R}^m$, $f(y)=\lambda_0+\delta^*_{\mathcal{F}}(\lambda)$, $m=1$ and $g_1(y,z_1)=\lambda_0+f(z_1)^\top\lambda-\ell(x,z_1)$. Alternatively, $\delta^*_{\mathcal{F}}(\lambda)$ can be recast as the optimal value of a dual minimization problem, and the underlying decision variables can be appended to y. As another example, if $\mathcal{P}$ is the ϕ-divergence ambiguity set (2.10) centred at a discrete distribution $\hat{\mathbb{P}}=\sum_{i\in[N]}\hat{p}_i\delta_{\hat{z}_i}$ and if mild regularity conditions hold, then Theorem 4.14 implies that

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(x,Z)] = \begin{cases} \inf & \lambda_0 + \lambda r + \displaystyle\sum_{i\in[N]} \hat{p}_i\cdot(\phi^*)^\pi(\ell(\hat{z}_i)-\lambda_0,\lambda) \\ \text{s.t.} & x\in\mathcal{X},\ \lambda_0\in\mathbb{R},\ \lambda\in\mathbb{R}_+ \\ & \lambda_0 + \lambda\,\phi^\infty(1) \geq \ell(x,z) \;\; \forall z\in\mathcal{Z}. \end{cases}$$

This minimization problem is readily recognized as an instance of (9.1). Note also that if $\mathcal{P}$ is the *restricted* ϕ-divergence ambiguity set (2.10) and $\hat{\mathbb{P}}$ is discrete, then, under mild regularity conditions, Theorem 4.15 implies that the above reformulation remains valid provided that $\mathcal{Z}$ is replaced with $\{\hat{z}_i : i\in[N]\}$. Finally, when $\mathcal{P}$ is the optimal transport ambiguity set (2.27) centred at a discrete reference distribution and if mild regularity conditions hold, then Theorem 4.18 implies that

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(x,Z)] = \begin{cases} \inf & \lambda r + \displaystyle\sum_{i\in[N]} \hat{p}_i s_i \\ \text{s.t.} & x\in\mathcal{X},\ \lambda\in\mathbb{R}_+,\ s\in\mathbb{R}^N \\ & \ell(x,z) - \lambda c(z,\hat{z}_i) \leq s_i \;\; \forall z_i\in\mathcal{Z},\ i\in[N]. \end{cases}$$

This minimization problem is again an instance of (9.1).

In the remainder of this section we discuss various numerical methods for solving the semi-infinite program (9.1). Some of these methods solve one or several relaxations of (9.1) that enforce the uncertainty-affected constraint only for a finite subset $\tilde{\mathcal{Z}}$ of $\mathcal{Z}$. Hence these methods assume access to a scenario oracle.

Definition 9.1 (Scenario oracle). Given any finite scenario set $\tilde{\mathcal{Z}}\subseteq\mathcal{Z}$, a scenario oracle outputs a solution to the scenario problem

$$\inf\{f(y) : y\in\mathcal{Y},\ g_j(y,z_j)\leq 0 \;\; \forall z_j\in\tilde{\mathcal{Z}}, \forall j\in[m]\}. \tag{9.2}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

As we will see below, cutting-plane algorithms refine scenario relaxations of the semi-infinite program (9.1) by iteratively adding those parameter realizations $z \in \mathcal{Z} \setminus \tilde{\mathcal{Z}}$ for which the constraint violation is maximal. Identifying such realizations requires a noise oracle as per the following definition.

Definition 9.2 (Noise oracle). Given any fixed decision $\tilde{y} \in \mathcal{Y}$, a noise oracle outputs a solution to the noise problem

$$\sup_{z \in \mathcal{Z}} \max_{j \in [m]} g_j(\tilde{y}, z). \tag{9.3}$$

In the following, we first survey the scenario approach, which replaces the semi-infinite program (9.1) with a finite scenario problem that offers stochastic approximation guarantees. This approach calls the scenario oracle only *once*. We then review cutting-plane techniques that iteratively call scenario and noise oracles to generate a solution sequence that attains the optimal value of problem (9.1), either within finitely many iterations or asymptotically. Next, we study online convex optimization algorithms, which do not require expensive scenario and/or noise oracles and instead solve only *deterministic* versions of problem (9.1) and use cheap first-order updates of the candidate decisions and/or incumbent worst-case parameter realizations. We close with an overview of specialized numerical solution methods that are tailored to specific ambiguity sets.

9.1. The scenario approach

The scenario approach was pioneered by De Farias and Van Roy (2004) in the context of robust Markov decision processes, and by Calafiore and Campi (2005, 2006) and Campi and Garatti (2008, 2011) in the context of generic robust optimization problems of the form (9.1). The scenario approach replaces the semi-infinite constraint in (9.1) with a collection of finitely many constraints corresponding to uncertainty realizations sampled from some fixed distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$.

Algorithm 1 (Scenario approach).

(1) *Initialization.* Fix a distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$.

(2) *Sampling.* Draw N independent samples $Z_1, \ldots, Z_N$ from $\mathbb{Q}$ and construct the scenario set $\tilde{\mathcal{Z}} = \{Z_1, \ldots, Z_N\}$.

(3) *Termination.* Return the output $\tilde{Y}$ of the scenario oracle (9.2) with input $\tilde{\mathcal{Z}}$.

Note that, as the input to the scenario oracle (9.2) is a random scenario set governed by the N-fold product distribution $\mathbb{Q}^N$, its output $\tilde{Y}$ is also random. Now fix a constraint violation probability $\delta \in (0, 1)$, a significance level $\eta \in (0, 1)$, and ensure that the sample size N in step (2) of Algorithm 1 satisfies $N \geq N(d_y, \delta, \eta)$, where d_y is the dimension of the decision vector y and

$$N(d_y, \delta, \eta) = \min\left\{ N \in \mathbb{N} \colon \sum_{i=0}^{d_y - 1} \binom{N}{i} \delta^i (1 - \delta)^{N-i} \leq \eta \right\}.$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Assuming that the objective and constraint functions of problem 9.1 are convex in y for any fixed z_j, $j \in [m]$, and that the optimal solution to (9.2) exists and is unique for any fixed scenario set $\tilde{\mathcal{Z}}$, Algorithm 1 then guarantees that

$$\mathbb{Q}^N(\mathbb{Q}(g_j(\tilde{Y}, Z) \leq 0 \ \forall j \in [m]) \geq 1 - \delta) \geq 1 - \eta,$$

where Z follows $\mathbb{Q}$ and $\tilde{Y}$ is governed by $\mathbb{Q}^N$; see Campi and Garatti (2008, Theorem 1). In other words, the output $\tilde{Y}$ of the scenario oracle (9.2) constitutes a feasible solution of the chance-constrained program

$$\inf\{f(y)\colon y \in \mathcal{Y},\ \mathbb{Q}(g_j(y, Z) \leq 0 \ \forall j \in [m]) \geq 1 - \delta\}.$$

with probability at least $1 - \eta$, where η can be interpreted as the (small) chance of poorly approximating $\mathbb{Q}$ in step (2) of Algorithm 1. We emphasize that the convexity of (9.1) plays a crucial role in the derivation of this probabilistic guarantee.

Two remarks on the scenario approach are in order. First, its performance guarantee is stochastic as it relates to a chance-constrained program that relaxes the semi-infinite program (9.1). Second, the sample size $N(d_y, \delta, \eta)$ needed for a probabilistic guarantee is of the order $\tilde{O}((d_y + \log(1/\eta))/\delta)$, that is, it grows linearly with the dimension d_y of the decision vector y. This dependence limits the problem dimensions that can be handled in practice. Robust performance guarantees for the scenario approach have been studied by Mohajerin Esfahani, Sutter and Lygeros (2015). The dependence of the probabilistic performance guarantees on the dimension of the decision vector y can be improved by using regularization (Campi and Caré 2013), one-off calibration schemes (Caré, Garatti and Campi 2014) and sequential validation (Calafiore, Dabbene and Tempo 2011), or by exploiting limited support ranks (Schildbach, Fagiano and Morari 2013) and solution-dependent numbers of support constraints (Campi and Garatti 2018). In general, however, the number of sampled constraints may remain large, which can be an impediment to the adoption of the scenario approach in large-scale problems.

9.2. Cutting-plane algorithms

Mutapcic and Boyd (2009) propose an iterative method for solving the semi-infinite program (9.1), which is based on Kelley's cutting-plane algorithm (Kelley Jr 1960). Their method can be described as follows.

Algorithm 2 (Cutting-plane algorithm).

(1) *Initialization.* Select a non-empty finite scenario set $\tilde{\mathcal{Z}} \subseteq \mathcal{Z}$, and set the feasibility threshold parameter ε to a small value.

(2) *Master problem.* Solve the scenario oracle problem (9.2) to find $\tilde{y}$.

(3) *Sub-problem.* Solve the noise oracle problem (9.3) to find $\tilde{z}$.

(4) *Termination.* If $g_j(\tilde{y}, \tilde{z}) \leq \varepsilon$ for all $j \in [m]$, terminate with $\tilde{y}$ as a ε-feasible solution to (9.1). Otherwise, update $\tilde{\mathcal{Z}} \leftarrow \tilde{\mathcal{Z}} \cup \{\tilde{z}\}$ and return to step (2).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Algorithm 2 alternates between two steps. Step (2) solves the scenario oracle problem (9.2) for a finite scenario set $\tilde{\mathcal{Z}}$ and outputs a candidate solution $\tilde{y}$. Step (3) then solves the noise oracle problem (9.3) for the given candidate solution $\tilde{y}$ and outputs a most violated scenario $\tilde{z}$. If the optimal value of (9.3) exceeds ε, then the scenario $\tilde{z}$ is added to the scenario set $\tilde{\mathcal{Z}}$ and the process is repeated. Otherwise, $\tilde{y}$ from step (2) is returned as an ε-feasible solution to the semi-infinite program (9.1), that is, a solution $\tilde{y} \in \mathcal{Y}$ that satisfies $g_j(\tilde{y}, z_j) \leq \varepsilon$ for all $z_j \in \mathcal{Z}$ and $j \in [m]$.

Cutting-plane algorithms replace the sampling procedure of the scenario approach with a noise oracle, but they still require access to a scenario oracle that solves the master problems. In contrast to the scenario approach, however, the number of constraints in the master problem increases with each iteration, which can limit scalability. If the constraint functions $g_j(y, z_j)$ are Lipschitz-continuous in y, then Algorithm 2 terminates after $O(\delta^{-d_y})$ iterations, which, however, is exponential in the dimension of y (Mutapcic and Boyd 2009, § 5.2). Despite this, in practice, cutting-plane algorithms often converge to near-optimal solutions in very few iterations, which has contributed to their widespread use in robust optimization.

9.3. Online convex optimization algorithms

Cutting-plane algorithms can become computationally expensive due to their reliance on scenario and noise oracles for the solution of the master and sub-problems, respectively. In the following, we review how ideas from online convex optimization can help to alleviate these shortcomings (Shalev-Shwartz 2012, Hazan 2022).

In their seminal work on this topic, Ben-Tal, Hazan, Koren and Mannor (2015*b*) employ a bisection search to reduce the solution of problem (9.1) to the solution of a sequence of robust feasibility problems of the form

$$\inf\{0\colon\ y \in \mathcal{Y},\ f(y) \leq c,\ g_j(y, z_j) \leq 0\ \ \forall z_j \in \mathcal{Z},\ \forall j \in [m]\}. \tag{9.4}$$

More precisely, the following bisection algorithm can be used to solve (9.1).

Algorithm 3 (Bisection algorithm).

(1) *Initialization.* Find an interval $[a, b]$ that contains the optimal value of (9.1), and select an arbitrary feasible solution $\tilde{y}$.

(2) *Decision problem.* Set $c = (a + b)/2$, and check if (9.4) is feasible or not.

(3) *Update.* If (9.4) is feasible, update $\tilde{y}$ to a solution of the feasibility problem (9.4), and set $b \leftarrow c$; otherwise, set $a \leftarrow c$.

(4) *Termination.* If $b - a \leq \delta$, terminate with $\tilde{y}$ as an approximately optimal solution to (9.1). Otherwise, return to step (2).

Ben-Tal *et al.* (2015*b*) use techniques from online convex optimization to solve the robust feasibility problem (9.4). In particular, they develop a method similar to Algorithm 2 that approximately solves a nominal feasibility problem instead of calling the scenario oracle and that uses a first-order update rule instead of calling

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

the noise oracle. Accordingly, they require the constraint functions g_j, $j \in [m]$, to be differentiable. Their algorithm can be summarized as follows.

Algorithm 4 (Dual-subgradient meta algorithm).

(1) *Initialization.* Choose initial uncertainty realizations $z_j \in \mathcal{Z}$, $j \in [m]$.

(2) *Nominal problem.* Find $\tilde{y}$ that solves the *approximate* feasibility problem

$$\inf\{0\colon f(y) \leq c,\ g_j(y, z_j) \leq \varepsilon\ \ \forall j \in [m]\}$$

corresponding to the current uncertainty realizations and corresponding to some $\varepsilon > 0$. If no such $\tilde{y}$ exists, terminate and report that (9.4) is infeasible.

(3) *Update parameters.* Update z_j, $j \in [m]$, using the gradient rule

$$z_j \leftarrow \mathrm{Proj}_{\mathcal{Z}}[z_j + \eta \nabla_z g_j(\tilde{y}, z_j)] \quad \text{for all } j \in [m],$$

where η is a given stepsize and $\mathrm{Proj}_{\mathcal{Z}}$ denotes the Euclidean projection onto $\mathcal{Z}$.

(4) *Termination.* Once a termination condition is met, return the average of all candidate solutions $\tilde{y}$ found in step (2).

Algorithms 3 and 4 can be combined into a single algorithm that finds a δ-optimal and ε-feasible solution to the semi-infinite program (9.1) in $O(\varepsilon^{-2}\log(1/\delta))$ iterations. This convergence guarantee holds under the following assumptions. The feasible sets $\mathcal{Y}$ and $\mathcal{Z}$ are closed and convex, the objective function $f\colon \mathcal{Y} \to \mathbb{R}$ is convex and Lipschitz-continuous, and the constraint functions $g_j\colon \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$, $j \in [m]$, constitute saddle functions. Specifically, $g(y, z_j)$ is convex and Lipschitz-continuous in y as well as concave and upper semicontinuous in z_j for every $j \in [m]$. We refer to Ben-Tal *et al.* (2015*b*) for further implementation details.

Algorithm 4 still solves multiple nominal feasibility problems in step (2), which can be expensive. As a remedy, Ho-Nguyen and Kılınç-Karzan (2018) reduce the solution of the feasibility problem (9.4) to the verification of the inequality

$$\min_{y \in \mathcal{Y}_c} \max_{z \in \mathcal{Z}} \max_{j \in [m]} g_j(y, z) \leq \varepsilon \tag{9.5}$$

for a given tolerance $\varepsilon > 0$, where $\mathcal{Y}_c = \{y \in \mathcal{Y}\colon f(y) \leq c\}$. Checking (9.5) requires the solution of a saddle point problem. Note that the objective function of this saddle point problem is convex in y but but fails to be concave in z when $m > 1$. Therefore, standard primal–dual algorithms from online convex optimization do not apply. Nevertheless, Ho-Nguyen and Kılınç-Karzan (2018) construct an online algorithm that outputs a trajectory of candidate solutions $\tilde{y}$ and uncertainty realizations $\tilde{z}$ that converge to a saddle point. This method uses a first-order algorithm $\mathcal{A}_y$ for solving the (parametric) minimization problem $\min_{y \in \mathcal{Y}_c} \max_{j \in [m]} g_j(y, z)$ as well as a first-order algorithm $\mathcal{A}_j$ for solving the (parametric) maximization problem $\max_{z \in \mathcal{Z}} g_j(y, z)$ for each $j \in [m]$ as subroutines. Specifically, $\mathcal{A}_y$ is assumed to map any history of candidate solutions $\tilde{y}^1, \ldots, \tilde{y}^t$ and uncertainty realizations

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$z_j^1, \ldots z_j^t \in \mathcal{Z}$ for $j \in [m]$ and $t \in \mathbb{N}$ to a new candidate solution $\tilde{y}^{t+1}$ such that

$$\sum_{t\in[T]} \max_{j\in[m]} g_j(\tilde{y}^t, \tilde{z}_j^t) - \min_{y\in\mathcal{Y}_c} \sum_{t\in[T]} \max_{j\in[m]} g_j(y, \tilde{z}_j^t) \leq \mathcal{R}_y(T) \quad \text{for all } T \in \mathbb{N},$$

where $\mathcal{R}_y(T)$ is a sublinear regret bound. Similarly, it is assumed that $\mathcal{A}_j$ maps any history of candidate solutions and uncertainty realizations of length $t \in \mathbb{N}$ to a new uncertainty realization $\tilde{z}_j^{t+1}$ such that

$$\max_{z\in\mathcal{Z}} \sum_{t\in[T]} g_j(\tilde{y}^t, z) - \sum_{t\in[T]} g_j(\tilde{y}^t, \tilde{z}_j^t) \leq \mathcal{R}_j(T) \quad \text{for all } T \in \mathbb{N},$$

where $\mathcal{R}_j(T)$ is a sublinear regret bound for every $j \in [m]$. The algorithm by Ho-Nguyen and Kılınç-Karzan (2018) can now be summarized as follows.

Algorithm 5 (Online learning framework).

(1) *Initialization.* Initialize the solution history to $\mathcal{H} \leftarrow \emptyset$.

(2) *Find candidate solution.* Use algorithm $\mathcal{A}_y$ with input $\mathcal{H}$ to construct a new candidate solution, that is, set $\tilde{y} \leftarrow \mathcal{A}_y(\mathcal{H})$.

(3) *Find uncertainty realizations.* Use algorithm $\mathcal{A}_j$ with input $\mathcal{H}$ to construct a new uncertainty realization, that is, set $\tilde{z}_j \leftarrow \mathcal{A}_j(\mathcal{H})$ for all $j \in [m]$.

(4) *Update history.* Update the solution history to $\mathcal{H} \leftarrow \mathcal{H} \cup \{(\tilde{y}, (\tilde{z}_j)_j)\}$.

(5) *Termination.* Once a termination condition is met, return the average of all candidate solutions $\tilde{y}$ found in step (2).

Ho-Nguyen and Kılınç-Karzan (2018) show that Algorithm 5 solves the saddle point problem on the left-hand side of (9.5) approximately with regret guarantee

$$\sum_{t\in[T]} \max_{z\in\mathcal{Z}} \max_{j\in[m]} g_j(\tilde{y}^t, z) - \min_{y\in\mathcal{Y}_c} \max_{j\in[m]} g_j(y, \tilde{z}^t) \leq \mathcal{R}_y(T) + \max_{j\in[m]} \mathcal{R}_j(T)$$

for all $T \in \mathbb{N}$. The total regret bound in the above expression grows sublinearly with T. Under the usual convexity assumptions, Algorithms 3 and 5 can be combined into a joint algorithm that finds a δ-optimal and ε-feasible solution to the semi-infinite program (9.1) in $O(\varepsilon^{-2}\log(1/\delta))$ iterations. Thus the iteration complexity did not improve *vis-à-vis* the algorithm by Ben-Tal *et al.* (2015*b*). However, the computational effort per iteration is significantly lower for Algorithm 5 than for Algorithm 4. Indeed, Algorithm 4 solves a feasibility problem with m uncertainty realizations in each iteration, whereas Algorithm 5 only calls the algorithms $\mathcal{A}_y$ and $\mathcal{A}_j$, $j \in [m]$, which compute cheap first-order updates. For further details, we refer to Ho-Nguyen and Kılınç-Karzan (2018). In addition, Ho-Nguyen and Kılınç-Karzan (2019) exploit structural properties of the objective and constraint functions to reduce the overall iteration complexity to $O(\varepsilon^{-1}\log(1/\delta))$.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Up until now, all the algorithms discussed in this section relied on the bisection method to reduce the semi-infinite program (9.1) to a sequence of robust feasibility problems (9.4). This introduces unnecessary computational overhead. As a remedy, Postek and Shtern (2024) use primal–dual saddle point algorithms that address the following perspective reformulation of problem (9.1), which was initially introduced in Appendix A of the work by Ho-Nguyen and Kılınç-Karzan (2018):

$$\min_{y\in\mathcal{Y}} \max_{z\in\mathcal{Z},\lambda\in\Delta^m} f(y) + \sum_{j\in[m]} \lambda_j g_j(y, z/\lambda_j).$$

Here,

$$\Delta^m = \left\{ \lambda \in \mathbb{R}^m_+ : \sum_{j\in[m]} \lambda_j = 1 \right\},$$

and $0 \cdot g_j(y, z/0)$ is interpreted as the negative recession function of the convex function $-g_j(y,\cdot)$. By construction, the objective function of this reformulation is convex in y and jointly concave in Z and λ. While the primal–dual saddle point algorithm of Postek and Shtern (2024) typically enjoys an iteration complexity of $O(\varepsilon^{-2})$, where ε now represents the primal–dual gap in the saddle point formulation, it requires more sophisticated oracles than those used by Ho-Nguyen and Kılınç-Karzan (2018, 2019). This is primarily because the perspective transformation eliminates favourable properties such as strong convexity and smoothness, and it also significantly degrades the Lipschitz constants. To address this challenge while still relying on standard oracles as in Ho-Nguyen and Kılınç-Karzan (2018, 2019), Tu, Chen and Yue (2024) propose a two-layer algorithm based on the following Lagrangian formulation of (9.1):

$$\max_{\lambda\in\mathbb{R}^m_+} \min_{y\in\mathcal{Y}} \max_{z\in\mathcal{Z}} f(y) + \sum_{j\in[m]} \lambda_j g_j(y, z)$$

Tu *et al.* (2024) show that their algorithm has an iteration complexity of $O(\varepsilon^{-3})$ or $O(\varepsilon^{-2})$, depending on the smoothness of the objective and constraint functions.

9.4. Tailored numerical solution methods for specific ambiguity sets

With the exception of some of the online optimization algorithms, the numerical solution methods discussed thus far still rely on general-purpose solvers to solve auxiliary nominal, scenario, master and/or sub-problems. General-purpose solvers are typically based on second-order interior-point methods that may fail to offer scalability to large-scale problem instances. To alleviate this concern, several first-order methods have been developed for specific classes of ambiguity sets.

9.4.1. Gelbrich ambiguity sets

Gelbrich ambiguity sets naturally emerge in signal processing and control applications. The standard reformulations of DRO problems over Gelbrich ambiguity sets,

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

however, constitute semidefinite programs (see Theorem 7.10), which significantly limits their scalability. To circumvent this shortcoming, Shafieezadeh-Abadeh *et al.* (2018) develop a Frank–Wolfe algorithm whose direction-finding subproblem admits a quasi-closed-form solution. This algorithm enjoys a sublinear convergence rate. Leveraging the strong convexity of the Gelbrich ambiguity set, Nguyen *et al.* (2023*b*) improve this Frank–Wolfe algorithm to achieve a linear convergence rate whenever the loss function satisfies the Levitin–Polyak condition (Levitin and Polyak 1966). Using frequency-domain techniques, Kargin *et al.* (2024*b*,*c*) introduce a Frank–Wolfe algorithm for infinite-horizon robust control problems that involve infinite-dimensional moment matrices. Finally, McAllister and Mohajerin Esfahani (2024) propose a Newton method for solving a class of DRO problems over Gelbrich ambiguity sets that converges superlinearly in numerical experiments.

9.4.2. ϕ-divergence ambiguity sets

The existing literature largely focuses on DRO problems over the restricted ϕ-divergence ambiguity set (2.11), including the group DRO formulation introduced by Sagawa, Koh, Hashimoto and Liang (2020) as a special case. Unfortunately, stochastic gradient methods applied directly to the dual minimization problem (4.12) are known to be unstable. This challenge motivated Namkoong and Duchi (2016) to adopt a direct saddle point formulation of the DRO problem with a discrete reference distribution $\hat{\mathbb{P}}$, which they solve iteratively with a bandit mirror descent algorithm. Several other algorithms address the saddle point formulation, including customized multilevel Monte Carlo methods (Levy, Carmon, Duchi and Sidford 2020, Hu, Chen and He 2021, Hu, Wang, Chen and He 2024), accelerated methods that query ball optimization oracles (Carmon and Hausler 2022) and biased stochastic gradient methods (Ghosh, Squillante and Wollega 2021, Wang, Gao and Xie 2024*b*, Azizian, Iutzeler and Malick 2023*b*). Gürbüzbalaban, Ruszczyński and Zhu (2022) and Zhu, Gürbüzbalaban and Ruszczyński (2023) solve non-convex DRO problems over classes of ϕ-divergence ambiguity sets. Specifically, Gürbüzbalaban *et al.* (2022) introduce a subgradient algorithm for non-smooth and non-convex loss functions, while Zhu *et al.* (2023) establish convergence rates and finite-sample guarantees for a subgradient method targeted at weakly convex loss functions. Both works build on the foundational results of Ruszczyński (2021), which laid the groundwork for efficient first-order methods for multilevel optimization problems.

9.4.3. Optimal transport ambiguity sets

Li, Huang and So (2019*c*) develop a first-order iterative method for distributionally robust logistic regression problems over 1-Wasserstein balls. This method is based on a variant of the proximal alternating direction method of multipliers (ADMM). Numerical experiments demonstrate that the proposed algorithm is several orders of magnitude faster than general-purpose solvers. A similar conclusion is drawn by

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

Li, Chen and So (2020), who introduce epigraphical projection-based algorithms to solve distributionally robust support vector machine problems. When the loss function $\ell(x, z)$ is either convex–concave or convex–convex in x and z, respectively, the reformulation of the DRO problem (1.2) reveals a structure that is conducive to distributed implementation. Consequently, Cherukuri and Cortés (2019) use saddle point algorithms related to the augmented Lagrangian method to solve the reformulated problem over a network of agents. For convex–concave loss functions, Li and Martínez (2020) propose a hybrid algorithm that combines Frank–Wolfe and subgradient methods. For any fixed $x \in \mathcal{X}$, their approach solves the inner maximization problem in (1.2) with a variant of the Frank–Wolfe algorithm. The resulting maximizer is then used to construct an approximate subgradient for the outer minimization problem. All of these algorithms crucially rely on the reference distribution $\hat{\mathbb{P}}$ being discrete. Blanchet and Kang (2020) and Blanchet *et al.* (2022*c*) propose a stochastic gradient descent algorithm to solve DRO problems over optimal transport ambiguity sets with generic reference distributions. Other stochastic optimization schemes leverage variance reduction techniques (Yu, Lin, Mazumdar and Jordan 2022) and zeroth-order random reshuffling algorithms (Maheshwari *et al.* 2022). These works typically rely on the duality results introduced in Section 4, and subsequently apply stochastic subgradient descent, using subgradients of the regularized loss function $\ell_c(x, \hat{z}) = \sup_{z \in \mathcal{Z}} \ell(x, z) - \lambda c(z, \hat{z})$ with respect to x and λ. Ho-Nguyen and Wright (2023) extend this approach to non-convex robust binary classification problems. Sinha *et al.* (2018) examine relaxed distributionally robust neural network training problems, assuming that the required level of robustness against adversarial perturbations is sufficiently small. This is tantamount to forcing λ to exceed a sufficiently large threshold. If $c(z, \hat{z}) = \|z - \hat{z}\|_2^2$, this in turn ensures that the maximization problem over z that defines $\ell_c(x, \hat{z})$ has a strongly concave objective function and is thus efficiently solvable. Stochastic subgradients of $\ell_c(x, \hat{z})$ are therefore readily available thanks to Danskin's theorem. Shafiee *et al.* (2023) leverage non-convex duality theorems, such as Toland's duality principle, to solve distributionally robust portfolio selection problems. Algorithms that minimize the variation-regularized nominal loss, which is known to approximate the worst-case expected loss thanks to Theorem 8.7, are explored by Li *et al.* (2022) and Bai *et al.* (2023*a*). Finally, Wang *et al.* (2021, 2024*b*) and Azizian *et al.* (2023*b*) introduce entropy and ϕ-divergence regularizers to improve the efficiency of algorithms for Wasserstein DRO problems, and Vincent, Azizian, Malick and Iutzeler (2024) provide a Python library for training related distributionally robust machine learning models.

10. Statistical guarantees

Despite ample empirical evidence that distributionally robust decisions can outperform those provided by alternative methodologies for decision-making under uncertainty, the statistical properties of DRO remain under-explored. This section

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

aims to survey some of the key techniques and methods employed in the literature to analyse the statistical aspects of DRO, while at the same time acknowledging that numerous questions remain open in this domain.

The statistical guarantees of moment-based ambiguity sets are relatively weak in the sense that the optimal value of problem (1.2) under a moment-based ambiguity set $\mathcal{P}$ does not match the optimal value of the corresponding stochastic program (1.1) under the unknown true distribution $\mathbb{P} = \mathbb{P}_0$ even if $\mathbb{P}_0$ was known exactly when $\mathcal{P}$ is constructed. The reason for this is that exact knowledge of lower-order moments of $\mathbb{P}_0$ is not sufficient to uniquely characterize $\mathbb{P}_0$ itself. For this reason, our review focuses on ϕ-divergence and optimal transport ambiguity sets, which offer asymptotic consistency guarantees as the number of samples available from $\mathbb{P}_0$ grows, and we refer to Delage and Ye (2010) and Nguyen *et al.* (2021) for statistical analyses of Chebyshev and Gelbrich ambiguity sets, respectively.

Section 10.1 introduces the data-driven optimization framework that we will be interested in, as well as the two key performance criteria of *excess risk* and *out-of-sample disappointment*. Subsequently, Section 10.2 surveys asymptotic analyses, which are based on the laws of large numbers, the central limit theorem and the empirical likelihood approach, as well as the large and moderate deviations principles. Finally, Section 10.3 reviews non-asymptotic analyses, which rely on measure concentration bounds as well as generalization bounds.

Our review of the statistical properties of DRO omits several important topics. For example, we do not cover domain adaptation guarantees (Farnia and Tse 2016, Volpi *et al.* 2018, Lee and Raginsky 2018, Lee, Park and Shin 2020, Sutter, Krause and Kuhn 2021, Taşkesen *et al.* 2021, Rychener *et al.* 2024), which ensure that a DRO model trained on data from some source distribution generalizes to a different target distribution. We also omit discussions of adversarial generalization bounds (Sinha *et al.* 2018, Wang *et al.* 2019, Tu, Zhang and Tao 2019, Kwon, Kim, Won and Paik 2020, An and Gao 2021), which use DRO to analyse model robustness against adversarial perturbations, as well as applications in high-dimensional statistical learning (Aolaritei, Shafiee and Dörfler 2022*b*). Finally, we do not cover Bayesian guarantees (Gupta 2019, Shapiro, Zhou and Lin 2023, Liu, Su and Xu 2024*b*), which focus on average-case rather than worst-case performance guarantees.

10.1. Excess risk and out-of-sample disappointment

Consider the idealized scenario in which the uncertainty underlying a decision problem follows a known probability distribution $\mathbb{P}_0 \in \mathcal{P}(\mathcal{Z})$. In this case, we aim to determine a decision x_0 that minimizes the expected value of a loss function $\ell\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ with respect to $\mathbb{P}_0$. That is, we seek an element x_0 of

$$\mathcal{X}_0 = \operatorname*{arg\,min}_{x \in \mathcal{X}} \ \mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)]. \tag{10.1}$$

Note that problem (10.1) constitutes a classical stochastic program. While (10.1) is

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

theoretically sound, it faces two significant practical limitations. First, the distribution $\mathbb{P}_0$ underlying a decision problem is rarely known in practice. Second, even if $\mathbb{P}_0$ was known, evaluating the objective function of (10.1) requires the computation of an integral, which is intractable in high dimensions even for simple nonlinear loss functions (Dyer and Stougie 2006, Hanasusanto, Kuhn and Wiesemann 2016).

In practice, we often observe the true probability distribution $\mathbb{P}_0$ indirectly through historical data. From now on we thus assume we have access to N independent training samples from $\mathbb{P}_0$, denoted as $Z_1, \ldots, Z_N$. The goal of data-driven optimization is to construct a decision from the training samples. This decision should perform well not just on the training data but also on unseen test samples from $\mathbb{P}_0$. The performance of a data-driven decision on test data is also called its *out-of-sample performance*. Formally, data-driven optimization aims to learn a decision rule $\mathcal{T}_N : \mathcal{Z}^N \rightrightarrows \mathcal{X}$ that maps training samples from the product space $\mathcal{Z}^N$ to a set of candidate decisions in the decision space $\mathcal{X}$. Note that $\mathcal{T}_N$ constitutes a set-valued mapping because it is usually constructed as the set of minimizers of an optimization problem depending on the training samples. A data-driven decision is then any (Borel-measurable) function $\hat{X}_N$ of the training samples that satisfies

$$\hat{X}_N \in \mathcal{T}_N(Z_1, \ldots, Z_N).$$

Note that $\hat{X}_N$ inherits the randomness of the training samples and is therefore a random vector. However, we notationally suppress its dependence on the training samples in order to avoid clutter. Instead we use the superscript '$\hat{}$' together with the subscript 'N' to designate any random objects that are defined as functions of $Z_1, \ldots, Z_N$ and are thus governed by the product distribution $\mathbb{P}_0^N$.

Arguably the simplest approach to data-driven optimization is the sample average approximation (SAA), which is also known as empirical risk minimization in statistics. The idea of SAA is to replace the unobservable true distribution $\mathbb{P}_0$ in (10.1) with the observable empirical distribution

$$\hat{\mathbb{P}}_N = \frac{1}{N} \sum_{i \in [N]} \delta_{Z_i} \tag{10.2}$$

formed from the training samples $Z_1, \ldots, Z_N$ and to construct the decision rule

$$\mathcal{T}_N(Z_1, \ldots, Z_N) = \operatorname*{arg\,min}_{x \in \mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(x, Z)]. \tag{10.3}$$

As the empirical distribution is discrete, the SAA approach obviates the need to evaluate high-dimensional integrals and is thus computationally attractive. Nevertheless, the performance of its optimal solutions on test data can be disappointing even when the test data are independently sampled from the true distribution $\mathbb{P}_0$. This phenomenon has been observed across various application domains and has been given different names depending on the context. In finance, Michaud (1989) identifies this issue as the *error maximization effect* of portfolio optimization. Statistics and machine learning recognizes it as *overfitting*, a well-known challenge

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

where models perform well on training data but fail to generalize to new, unseen test data. In the stochastic programming literature, Shapiro (2003) refers to this phenomenon as the *optimization bias*, and in decision analysis the effect has been described as the *optimizer's curse* (Smith and Winkler 2006).

The disappointing out-of-sample performance of the SAA decisions prompted statisticians and machine learners to add a regularization term to the objective function in (10.3). The regularization term serves two purposes. It not only combats overfitting to the training data, but it also encourages simpler decisions. Such simplicity aligns with the principle of parsimony and reflects nature's inherent tendency towards simplicity. As Jeffreys and Wrinch (1921) aptly noted,

> The existence of simple laws is, then, apparently, to be regarded as a quality of nature; and accordingly we may infer that it is justifiable to prefer a simple law to a more complex one that fits our observations slightly better.

Formally, the regularized SAA approach provides the decision rule

$$\mathcal{T}_N(Z_1, \ldots, Z_N) = \operatorname*{arg\,min}_{x \in \mathcal{X}} \; \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(x, Z)] + R(x),$$

where the regularization function $R\colon \mathcal{X} \to \mathbb{R}_+$ penalizes the complexity of decision x. In the classical statistics literature, the regularization function is mostly *data-independent*, that is, it only depends on the decision x and not on the observed training data $Z_1, \ldots, Z_N$. The most prominent examples include norm regularization, where $R(x) = \|x\|$, and Tikhonov regularization, where $R(x) = \|x\|^2$. These regularization techniques balance the conflicting goals of computing decisions that are optimal for the observed training data and maintaining model simplicity, thereby improving the generalization capability of the derived decisions to unseen data.

Recall from Sections 6 and 8 that regularization and distributional robustness are closely intertwined. Assume that we use the empirical distribution $\hat{\mathbb{P}}_N$ as the centre of a ϕ-divergence ambiguity set (2.10) or optimal transport ambiguity set (2.27). Then the DRO approach provides the decision rule

$$\mathcal{T}_N(Z_1, \ldots, Z_N) = \operatorname*{arg\,min}_{x \in \mathcal{X}} \; \sup_{\mathbb{P} \in \hat{\mathcal{P}}_N} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)],$$

which can be viewed as a variant of the regularized SAA decision rule. The corresponding *data-dependent* regularization function is called the *DRO regularizer* and is given by

$$\hat{R}_N(x) = \sup_{\mathbb{P} \in \hat{\mathcal{P}}_N} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(x, Z)]. \tag{10.4}$$

Thus it depends on both the decision x and the observed training data $Z_1, \ldots, Z_N$. The regularizer (10.4) quantifies how much the worst-case expected loss across all distributions $\mathbb{P} \in \hat{\mathcal{P}}_N$ can exceed the in-sample expected loss $\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(x, Z)]$.

The performance of decision rules in data-driven optimization is primarily measured by two criteria, each of which is aligned with a different field of study and

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

addresses a different set of practical concerns. The first criterion, *excess risk*, is predominantly used in statistics. It quantifies the distance of a data-driven decision $\hat{X}_N$ to an optimal decision x_0. The second criterion, *out-of-sample disappointment*, is more commonly employed in operations research. It provides a measure of how much the out-of-sample risk of a data-driven decision $\hat{X}_N$ exceeds the in-sample risk of $\hat{X}_N$. In the following, we formally define both criteria.

Excess risk. Let $\eta \in (0,1)$ be a significance level, let $\mathcal{T}_N$ be a decision rule, and let $\Delta\colon \mathcal{X} \times \mathcal{X}_0 \to \mathbb{R}_+$ be a *performance function*. Suppose that $\hat{X}_N \in \mathcal{T}_N(Z_1, \ldots, Z_N)$. The excess risk criterion offers the guarantee that for any size $N \geq N(\mathcal{X}, \mathcal{Z}, \eta)$ of the training set, we have

$$\mathbb{P}_0^N [\Delta(\hat{X}_N, x_0) \leq \hat{\delta}_N] \geq 1 - \eta \tag{10.5}$$

for some (possibly data-dependent) error certificate $\hat{\delta}_N$. In statistical learning theory, performance functions often measure the regret in terms of the loss function ℓ under the true distribution $\mathbb{P}_0$. Specifically, for any feasible candidate decisions $x \in \mathcal{X}$ and any optimal decision $x_0 \in \mathcal{X}_0$, the *regret* takes the form

$$\Delta(x, x_0) = \mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)] - \mathbb{E}_{\mathbb{P}_0}[\ell(x_0, Z)] = \mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)] - \min_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)] \geq 0.$$

In compressed sensing and M-estimation problems with linear models, performance is often defined as the *estimation error* in the decision space, and it takes the form

$$\Delta(x, x_0) = \|x - x_0\|_2^2.$$

Here we assume for simplicity that the minimizer x_0 is unique. We refer to Mendelson (2003) and Bousquet, Boucheron and Lugosi (2004) for an introduction to statistical learning theory. For more advanced treatments, we refer to Anthony and Bartlett (1999), Koltchinskii (2011), Vapnik (2013), Shalev-Shwartz and Ben-David (2014), Vershynin (2018) and Wainwright (2019).

Out-of-sample disappointment. Let $\eta \in (0,1)$ be a significance level and let $\mathcal{T}_N$ be a decision rule. Suppose that $\hat{X}_N \in \mathcal{T}_N(Z_1, \ldots, Z_N)$. The out-of-sample disappointment criterion offers the guarantee that for any size $N \geq N(\mathcal{X}, \mathcal{Z}, \eta)$ of the training set, we have

$$\mathbb{P}_0^N \left[\mathbb{E}_{\mathbb{P}_0}[\ell(\hat{X}_N, Z)] \leq \hat{L}_N\right] \geq 1 - \eta \tag{10.6}$$

for some (possibly data-dependent) loss certificate $\hat{L}_N$. Alternatively, one can express (10.6) as a probabilistic bound on the difference between the out-of-sample performance and the in-sample performance,

$$\mathbb{P}_0^N \left[\mathbb{E}_{\mathbb{P}_0}[\ell(\hat{X}_N, Z)] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\hat{X}_N, Z)] \leq \hat{\delta}_N\right] \geq 1 - \eta,$$

for some error certificate $\hat{\delta}_N$. Both criteria become equivalent when we set $\hat{\delta}_N = \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\hat{X}_N, Z)] + \hat{L}_N$. Unlike the excess risk bound (10.5), the out-of-sample

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

disappointment bound (10.6) does not require explicit knowledge of an optimal decision x_0 and solely leverages the statistical properties of $\mathbb{P}_0$. As we will see in the following sections, $\hat{L}_N$ and $\hat{\delta}_N$ typically correspond to the optimal value of the DRO problem (1.2) and the DRO regularizer (10.4), respectively.

The next sections focus on ambiguity sets that are centred at the empirical distribution $\hat{\mathbb{P}}_N$ defined in (10.2). Specifically, we consider ambiguity sets constructed using a discrepancy measure $\mathrm{D}\colon \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty]$:

$$\hat{\mathcal{P}}_N = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z})\colon \mathrm{D}(\mathbb{P}, \hat{\mathbb{P}}_N) \leq r_N\}. \tag{10.7}$$

The discrepancy measure D could be a ϕ-divergence or a Wasserstein distance. We will explain how the radius r_N should scale with the training sample size N to obtain the least conservative statistical guarantees.

10.2. Asymptotic analyses

The laws of large numbers and the central limit theorem provide foundational insights into the statistical properties of the SAA approach. Under appropriate regularity conditions, the laws of large numbers guarantee that the empirical loss $\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(x, Z)]$ converges $\mathbb{P}_0$-almost surely to the true expected loss $\mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)]$, uniformly on $\mathcal{X}$ (see e.g. Shapiro *et al.* 2009, § 7.2.5). This implies that the optimal value and the set of optimal solutions of the SAA problem exhibit *asymptotic consistency*, that is, they both converge to their counterparts in the stochastic program under $\mathbb{P}_0$ as the sample size N approaches infinity. The central limit theorem, on the other hand hand, implies that the scaled difference between the empirical loss (under $\hat{\mathbb{P}}_N$) and true expected loss (under $\mathbb{P}_0$) converges weakly to a normal distribution with mean zero and variance equal to the true variance of the loss under $\mathbb{P}_0$ (see e.g. Shapiro *et al.* 2009, § 5.1.2). Thus the optimal value of the SAA problem also exhibits *asymptotic normality*. The asymptotic properties of the SAA decision rule have been studied extensively; see e.g. Cramér (1946), Huber (1967), Dupacová and Wets (1988), Shapiro (1989, 1990, 1991, 1993), King and Wets (1991), King and Rockafellar (1993), Van der Vaart (1998) and Lam (2021).

Building on these foundations, we will next review the asymptotic consistency and normality of DRO decision rules. While studying these asymptotic behaviours, different theoretical frameworks provide distinct insights. The central limit theorem and empirical likelihood approaches characterize the typical fluctuations around the mean under an appropriate scaling. The central limit theorem establishes Gaussian convergence, whereas the empirical likelihood theory provides a non-parametric framework for constructing likelihood ratio tests with asymptotic χ^2-limits, enabling hypothesis testing without specific parametric assumptions. In contrast, large deviations theory examines the tail behaviour of distribution sequences. Rather than focusing on typical fluctuations, it characterizes the exponential decay rate of probabilities associated with rare events far from the mean. Moderate deviations theory bridges the gap between the typical and rare event

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

analyses provided by the aforementioned frameworks. It studies the asymptotic behaviour of distribution sequences at intermediate scales, thus investigating larger deviations than the central limit theorem but smaller deviations than large deviations theory.

10.2.1. Asymptotic consistency and normality

Lam (2019, Theorem 6) establishes the asymptotic (uniform strong) consistency of the optimal value of DRO decision rules over likelihood ambiguity sets. The proof relies on the preservation theorem of Glivenko–Cantelli classes (Van Der Vaart and Wellner 2000, Theorem 3), which intuitively says that function classes maintain their uniform convergence properties when combined through continuous operations, assuming that the original classes are well-behaved. Duchi *et al.* (2021, Theorem 6) extend the analysis to more general ϕ-divergence ambiguity sets.

Mohajerin Esfahani and Kuhn (2018, Theorem 3.6) establish the asymptotic consistency of the optimal value and the optimal solutions of DRO decision rules over 1-Wasserstein balls of the form (10.7). Their proof combines the Borel–Cantelli lemma (Kallenberg 1997, Theorem 2.18) with measure concentration results by Fournier and Guillin (2015, Theorem 2). Intuitively, the Borel–Cantelli lemma asserts that if probabilities of an infinite sequence of events $(\mathcal{E}_N)_{N\in\mathbb{N}}$ have a finite sum, then the probability of infinitely many occurrences of these events is zero. Leveraging its contraposition, Mohajerin Esfahani and Kuhn (2018) consider the events $\mathcal{E}_N = \mathrm{W}_1(\mathbb{P}_0, \hat{\mathbb{P}}_N) \leq r_N$, where $\hat{\mathbb{P}}_N$ is the empirical distribution over N independent samples from $\mathbb{P}_0$; see (10.2). By selecting a converging sequence of radii $(r_N)_{N\in\mathbb{N}}$ that decay according to a scaling law informed by Fournier and Guillin (2015, Theorem 2), they prove that $\mathbb{P}_0^\infty(\lim_{N\to\infty} \mathrm{W}_1(\mathbb{P}_0, \hat{\mathbb{P}}_N) = 0) = 1$. This enables them to show that the optimal value of the DRO problem (1.2) over the 1-Wasserstein ball (10.7) converges asymptotically *from above* to the optimal value of the stochastic program (10.1). They also establish asymptotic convergence of the optimal solutions under an additional continuity assumption. This result can be extended to general p-Wasserstein ambiguity sets (Kuhn *et al.* 2019, Theorem 20). Similar asymptotic convergence results have been established by Gao *et al.* (2024*b*, Proposition 1), albeit through a different approach. Their proof does not rely on measure concentration results or an explicit characterization of the radius r_N. Instead, it leverages Theorem 4.18 together with the reverse Fatou lemma and the monotone convergence theorem. This approach, however, does not explicitly determine whether the asymptotic convergence occurs from above or below.

Lam (2019, Theorem 4) establishes the asymptotic normality of the optimal values of DRO problems over likelihood ambiguity sets. In a similar fashion, Duchi and Namkoong (2019, Theorem 10) establish the asymptotic normality of the optimal solutions of DRO problems over Pearson χ^2-divergence ambiguity sets. Duchi and Namkoong (2021, Theorem 11) extend this result to Cressie–Read ambiguity sets. The asymptotic normality of DRO decision rules over p-Wasserstein balls, finally, is established by Blanchet *et al.* (2019*b*, 2022*a*,*b*).

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

More recently, Blanchet and Shapiro (2024) have developed a comprehensive framework for analysing statistical limit theorems for DRO decision rules over both ϕ-divergences and Wasserstein ambiguity sets of the form (10.7). By connecting data-driven DRO formulations to their regularized counterparts (see Section 8), their framework provides insights into how DRO decision rules behave depending on the rate at which the radius r_N decreases with the sample size N. Specifically, Blanchet and Shapiro (2024, § 2.2) show that, under suitable regularity conditions, DRO formulations typically exhibit three distinct asymptotic behaviours.

(i) When r_N decreases faster than the critical statistical rate of $N^{-1/2}$, the DRO effect becomes negligible compared to the sampling error, and the asymptotic behaviour of DRO mirrors that of standard empirical risk minimization.

(ii) When r_N decreases at precisely the critical rate $N^{-1/2}$, the DRO effect manifests itself as a quantifiable asymptotic bias term that acts as a regularizer, and its interaction with the statistical noise results in a shifted normal limiting distribution.

(iii) When r_N decreases more slowly than $N^{-1/2}$, the DRO effect dominates the statistical noise, leading to a limiting behaviour governed primarily by the geometry of the ambiguity set.

The analysis employs the functional central limit theorem alongside careful Taylor expansions of the worst-case expectation akin to those presented in Section 8. In particular, Blanchet and Shapiro (2024) establish that, under appropriate regularity conditions, the limiting distributions are normal with explicitly characterized means and variances.

10.2.2. Empirical likelihood approach

The (generalized) empirical likelihood theory introduced by Owen (1988, 1990, 1991, 2001) provides a powerful non-parametric analogue to parametric maximum likelihood theory. At its core, the empirical distribution $\hat{\mathbb{P}}_N$ serves as a non-parametric maximum likelihood estimator for the unknown true distribution $\hat{\mathbb{P}}_0$, and statistically relies on empirical likelihood ratios. Under suitable conditions, the empirical likelihood ratio statistic converges to a χ^2-distribution. Unlike the central limit theorem, which yields normal approximations and thus symmetric confidence intervals, the empirical likelihood theory typically produces asymmetric confidence regions. A key advantage of this approach is that the resulting data-driven confidence regions automatically adapt to the geometry of the underlying distribution and naturally respect constraints such as boundedness or non-negativity, without requiring explicit transformations or variance estimation. However, this theoretical elegance and flexibility comes at the computational overhead of computing both the lower and upper bounds of the confidence interval separately.

In the following, we briefly review the empirical likelihood approach and its application to DRO decision rules. Let $Z_1, \ldots, Z_N$ be independent samples from

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$\mathbb{P}_0$, and let $\theta\colon \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$ be a statistical quantity of interest (e.g. the expected value of Z_i). Empirical likelihood confidence regions for $\theta(\mathbb{P}_0)$ can be constructed as

$$\hat{\mathcal{C}}_N = \left\{\theta(\mathbb{P})\colon \mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}_N) \leq \frac{r}{N}\right\} \tag{10.8}$$

for some $r \in \mathbb{R}_+$. Thus the set $\hat{\mathcal{C}}_N$ is the image of a ϕ-divergence neighbourhood around the empirical distribution $\hat{\mathbb{P}}_N$ under θ. The key tool for establishing probabilistic bounds is the so-called *profile divergence* $\pi_N\colon \mathbb{R} \to \mathbb{R}_+$, which is defined by

$$\pi_N(\tau) = \inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathrm{D}_\phi(\mathbb{P}, \hat{\mathbb{P}}_N)\colon \theta(\mathbb{P}) = \tau\}. \tag{10.9}$$

For a functional θ satisfying suitable smoothness conditions, the empirical likelihood method provides asymptotically exact coverage guarantees of the form

$$\lim_{N\to\infty} \mathbb{P}_0^N(\theta(\mathbb{P}_0) \in \hat{\mathcal{C}}_N) = \lim_{N\to\infty} \mathbb{P}_0^N\left(\pi_N(\theta(\mathbb{P}_0)) \leq \frac{r}{N}\right) = 1 - \eta,$$

where η represents a significance level determined by r and θ.

The classical empirical likelihood approach (Owen 1988, 2001) relies on the empirical likelihood divergence with entropy function $\phi(s) = -\log(s) + s - 1$ if $s \geq 0$ and $\phi(s) = \infty$ if $s < 0$ (see Table 2.1). In this case, π_N is called the *profile likelihood*. Assume that Z is a d-dimensional random vector that is governed by the distribution $\mathbb{P}_0$ and whose covariance matrix has rank $d_0 \leq d$. For the expected value $\theta(\mathbb{P}_0) := \mathbb{E}_{\mathbb{P}_0}[Z]$, Owen (1990) proves that, as $N \to \infty$, we have

$$\pi_N(\mathbb{E}_{\mathbb{P}_0}[Z]) \xrightarrow{d} \chi^2_{d_0},$$

where $\chi^2_{d_0}$ denotes the χ^2-distribution with d_0 degrees of freedom. Thus $\hat{\mathcal{C}}_N$ constitutes an asymptotically exact $(1-\eta)$-confidence interval for $\theta(\mathbb{P}_0)$ if we set r in (10.8) to the $(1-\eta)$-quantile of a χ^2-distribution with d_0 degrees of freedom.

In the context of stochastic programming problems, the statistical quantity of interest is typically the optimal value of the stochastic program, that is, $\theta(\mathbb{P}) = \inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)]$. In this case, the set $\hat{\mathcal{C}}_N$ becomes the interval

$$\hat{\mathcal{C}}_N = \left[\inf_{\mathbb{P}\in\hat{\mathcal{P}}_N} \inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)], \quad \sup_{\mathbb{P}\in\hat{\mathcal{P}}_N} \inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)]\right],$$

where $\hat{\mathcal{P}}_N$ is the ϕ-divergence ambiguity set of the form (10.7) around $\hat{\mathbb{P}}_N$. If $\hat{\mathcal{P}}_N$ is a likelihood ambiguity set, Lam (2019) investigates the asymptotic coverage probability of this interval by leveraging asymptotic guarantees for the SAA decision rule by Lam and Zhou (2017). In particular, he shows that if suitable regularity conditions hold and $r_N = r/N$, where r is the $(1-\eta)$-quantile of a χ^2-distribution with a single degree of freedom, then $\hat{\mathcal{C}}_N$ becomes an asymptotically exact $(1-\eta)$-confidence interval. One can thus show that the resulting confidence bounds

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

achieve the asymptotically exact coverage at the parametric rate $O(N^{-1/2})$. Duchi *et al.* (2021) further generalize these results to DRO decision rules over broader classes of ϕ-divergence ambiguity sets. Additionally, He and Lam (2021) examine higher-order coverage errors and introduce a correction term similar to the Bartlett correction. They derive higher-order correction terms for general von Mises differentiable functionals and thus move beyond the approximately smooth functions previously studied in the empirical likelihood literature.

In a parallel line of research, Blanchet *et al.* (2019*b*, 2022*a*,*b*), Blanchet and Kang (2021) and Lin, Blanchet, Glynn and Nguyen (2024) introduce the *Wasserstein profile* function as a Wasserstein analogue to the profile divergence (10.9). This approach replaces the ϕ-divergence with the 2-Wasserstein distance, and it offers a geometric perspective on uncertainty quantification. This approach yields confidence bounds that achieve asymptotic parametric rate $O(N^{-1/2})$. For more details, we direct the readers to the recent survey by Blanchet *et al.* (2021).

10.2.3. Large and moderate deviations principles

Unlike the central limit theorem and the empirical likelihood approach, which characterize limits of distribution sequences, the theories of large and moderate deviations study the asymptotic tail behaviour of distribution sequences. Specifically, they prove exponential decay rates of probabilities of rare events over sequences of random variables. The foundations of large deviations theory trace back to two seminal developments in physics and mathematics. The first is Boltzmann's groundbreaking works on statistical mechanics and entropy. The second is Cramér's pioneering paper on the asymptotic behaviour of sums of random variables (Cramér 1938). Despite these early advances, the field lacked a unified mathematical framework until Varadhan's seminal paper (Varadhan 1966), which introduces a formal definition of a *large deviation principle*. We refer to the textbooks by Ellis (2007) and Dembo and Zeitouni (2009) for a modern treatment of the topic.

Assume now that the unknown true distribution $\mathbb{P}_0$ is known to belong to a parametric distribution family $\{\mathbb{P}_\theta : \theta \in \Theta\} \subseteq \mathcal{P}(\mathcal{Z})$, where θ ranges over a prescribed parameter space Θ. In this case, estimating $\mathbb{P}_0$ is tantamount to estimating the unknown true parameter vector $\theta_0 \in \Theta$ that satisfies $\mathbb{P}_0 = \mathbb{P}_{\theta_0}$. A *statistic* $\hat{\theta}_N$ is a random variable valued in Θ and constructed from $(Z_1, \ldots, Z_N) \sim \mathbb{P}_\theta^N$ that converges in probability to θ as N grows, for any $\theta \in \Theta$. Formally, we say that the statistic $\hat{\theta}_N$ satisfies a *large deviations principle* with speed b_N and with lower semicontinuous rate function $I : \Theta \times \Theta \to [0, \infty]$ if

$$\begin{aligned} -\inf_{\theta' \in \mathrm{int}(\mathcal{B})} I(\theta', \theta) &\leq \liminf_{N\to\infty} \frac{1}{b_N} \log \mathbb{P}_\theta(\hat{\theta}_N \in \mathcal{B}) \\ &\leq \limsup_{N\to\infty} \frac{1}{b_N} \log \mathbb{P}_\theta(\hat{\theta}_N \in \mathcal{B}) \\ &\leq -\inf_{\theta' \in \mathrm{cl}(\mathcal{B})} I(\theta', \theta) \end{aligned} \tag{10.10}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

for all $\theta \in \Theta$ and for all Borel sets $\mathcal{B} \subseteq \Theta$. Here we assume that the sequence b_N, $N \in \mathbb{N}$, tends monotonically towards infinity. If (10.10) holds, one can show under mild conditions that $I(\theta, \theta) = 0$ because $\hat{\theta}_N$ converges to θ in probability under $\mathbb{P}_\theta$. It is therefore natural to interpret $I(\theta', \theta)$ as a discrepancy function that quantifies the dissimilarity between the estimator realization θ' and the probabilistic model θ. As I is lower semicontinuous, the minimization problems on the left- and right-hand sides of (10.10) share the same infimum $r = \inf_{\theta' \in \mathrm{int}(\mathcal{B})} I(\theta', \theta) = \inf_{\theta' \in \mathrm{cl}(\mathcal{B})} I(\theta', \theta)$ for most Borel sets $\mathcal{B}$ of interest. In these cases, the inequalities in (10.10) collapse to equalities, and (10.10) simplifies to the more intuitive statement

$$\mathbb{P}_\theta(\hat{\theta}_N \in \mathcal{B}) = \exp(-r b_N + o(b_N)).$$

That is, the probability of the estimator $\hat{\theta}_N$ falling into the set $\mathcal{B}$ decays exponentially at rate r with speed b_N, where r can be interpreted as the I-distance from θ to $\mathcal{B}$.

Several statistics of practical interest satisfy large deviations principles. For example, if $\mathcal{Z}$ is finite and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is the family of *all* distributions on $\mathcal{Z}$ encoded by the corresponding probability vectors $\theta \in \Theta$, where Θ is the probability simplex of appropriate dimension, then the empirical distribution $\hat{\mathbb{P}}_N$ corresponding to the empirical probability vector $\hat{\theta}_N$ is an estimator for the data-generating distribution $\mathbb{P}_\theta$. In this case, Sanov's theorem (Cover and Thomas 2006, Theorem 11.4.1) asserts that $\hat{\theta}_N$ satisfies a large deviations principle with rate function $I(\theta', \theta) = \mathrm{KL}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ and speed $b_N = N$. Similarly, if $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is any distribution family parametrized by its unknown mean vector $\theta = \mathbb{E}_{\mathbb{P}_\theta}[Z]$ and if the log-moment generating function $\Lambda_\theta(t) = \log(\mathbb{E}_{\mathbb{P}_\theta}[\exp(t^\top Z)])$ is finite for all $t, \theta \in \mathbb{R}^d$, then the sample mean $\hat{\theta}_N = \frac{1}{N} \sum_{i \in [N]} Z_i$ is an estimator for θ. In this case, Cramér's theorem (Cramér 1938) asserts that $\hat{\theta}_N$ satisfies a large deviations principle with rate function $I(\theta', \theta) = \Lambda_\theta^*(\theta')$ and speed $b_N = N$. Note that the log-moment generating function Λ_θ as well as its conjugate Λ_θ^* are both convex. We remark that a large deviations principle with sublinear speed ($\lim_{N \to \infty} b_N / N = 0$) is sometimes referred to as a *moderate deviations principle*. For an example of a moderate deviations principle we refer to Jongeneel, Sutter and Kuhn (2022).

Van Parys *et al.* (2021) leverage Sanov's theorem to show that the optimal value of the DRO problem with a likelihood ambiguity set of radius r around the empirical distribution $\hat{\mathbb{P}}_N$ yields the least conservative confidence bound on the optimal value of the true stochastic program, asymptotically as the sample size N grows large, with significance level η decaying exponentially as e^{-rN}. More generally, Sutter, Van Parys and Kuhn (2024) assume that $\mathbb{P}_0$ is known to belong to a parametric distribution family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and that θ admits an estimator $\hat{\theta}_N$ that satisfies a large deviations principle with rate function I and speed $b_N = N$. Under some regularity conditions, they then show that the optimal value of the DRO problem with ambiguity set $\hat{\mathcal{P}}_N = \{\mathbb{P}_\theta : \theta \in \Theta,\ I(\hat{\theta}_N, \theta) \leq r\}$ again yields the least conservative confidence bound on the optimal value of the true stochastic program with significance level $\eta \propto \mathrm{e}^{-rN}$. Similar statistical optimality results can sometimes be obtained even when the training samples are serially dependent, for

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

example when they are generated by a Markov process with unknown transition probability matrix or certain autoregressive processes (Sutter *et al.* 2024).

The DRO estimators by Van Parys *et al.* (2021) and Sutter *et al.* (2024) lack asymptotic consistency because they exploit large deviations principles with linear speed $b_N = N$. Bennouna and Van Parys (2021) show that asymptotic consistency can be recovered by relying on moderate deviations principles with sublinear speed. This line of research has seen significant recent developments. The use of large and moderate deviations principles has also been extended to various learning and control settings such as distributionally robust Markov decision processes (Li, Sutter and Kuhn 2021), bandit problems (Van Parys and Golrezaei 2024), bootstrap-based methods (Bertsimas and Van Parys 2022), optimal learning (Ganguly and Sutter 2023, Liu *et al.* 2023), control (Jongeneel, Sutter and Kuhn 2021, Jongeneel *et al.* 2022), contextual learning (Srivastava, Wang, Hanasusanto and Ho 2021) and robust statistics (Chan, Van Parys and Bennouna 2024).

10.3. Non-asymptotic analyses

Non-asymptotic statistics seeks finite-sample guarantees that hold regardless of the sample size. This is in contrast to the asymptotic methods described in Section 10.2, which rely on properties that emerge as sample size tends infinity. Non-asymptotic methods allow for a rigorous control over error rates, which makes them robust in situations where asymptotic approximations might produce misleading results. In the following, we review two major classes of non-asymptotic analyses, that is, measure concentration bounds and generalization bounds.

10.3.1. Measure concentration bounds

The most elementary approach to obtaining finite sample guarantees is to design the ambiguity set $\hat{\mathcal{P}}_N$ such that it contains the unknown true probability distribution $\mathbb{P}_0$ with high probability. This requires an analysis of the convergence rate of $\hat{\mathbb{P}}_N$ towards $\mathbb{P}_0$, and it leads to out-of-sample disappointment bounds that depend only on $\hat{\mathcal{P}}_N$ and not on the complexity of the loss function ℓ or the decision space $\mathcal{X}$.

Theorem 10.1 (Out-of-sample disappointment). Suppose that the ambiguity set $\hat{\mathcal{P}}_N$ defined in (10.7) satisfies

$$\mathbb{P}_0^N(\mathbb{P}_0 \in \hat{\mathcal{P}}_N) \geq 1 - \eta. \tag{10.11}$$

We then have

$$\mathbb{P}_0^N\left(\mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)] \leq \sup_{\mathbb{P}\in\hat{\mathcal{P}}_N} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] \ \ \forall x \in \mathcal{X}\right) \geq 1 - \eta. \tag{10.12a}$$

Moreover, if $\hat{X}_N$ is an optimizer of the distributionally robust decision problem with respect to the ambiguity set $\hat{\mathcal{P}}_N$, then we have

$$\mathbb{P}_0^N\left(\mathbb{E}_{\mathbb{P}_0}[\ell(\hat{X}_N, Z)] \leq \min_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\hat{\mathcal{P}}_N} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)]\right) \geq 1 - \eta. \tag{10.12b}$$

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

The proof of (10.12a) and (10.12b) readily follows from the measure concentration bound (10.11) and is therefore omitted. Theorem 10.1 asserts that the worst-case expected loss provides an upper confidence bound on the true expected loss under the unknown data-generating distribution uniformly across all loss functions. Moreover, it also asserts that the optimal value of the DRO problem (1.2) provides an upper confidence bound on the out-of-sample performance of its optimizers.

When using ϕ-divergences to construct $\hat{\mathcal{P}}_N$ as in (10.7), the probabilistic requirement (10.11) only applies to underlying distributions $\mathbb{P}_0$ that are discrete (Polyanskiy and Wu 2024, § 7). In contrast, the Wasserstein distance applies to generic distributions $\mathbb{P}_0$. This area of study has a rich history, with seminal contributions from Dudley (1969), Ajtai, Komlós and Tusnády (1984) and Dobrić and Yukich (1995). More recent advancements have been made by Bolley, Guillin and Villani (2007), Boissard and Le Gouic (2014), Dereich, Scheutzow and Schottstedt (2013) and Fournier and Guillin (2015). Of particular importance to our discussion is the following measure concentration result, which serves as the foundation for finite sample guarantees in DRO over p-Wasserstein ambiguity sets.

Theorem 10.2 (Measure concentration (Fournier and Guillin 2015, Thm 2)). Suppose that $\hat{\mathbb{P}}_N$ is the empirical distribution constructed from N independent samples from $\mathbb{P}_0$, $p \neq d/2$, and that $\mathbb{P}_0$ is light-tailed in the sense that there exist $\alpha > p$ and $A > 0$ such that $\mathbb{E}_{\mathbb{P}_0}(\exp(\|Z\|^\alpha)) \leq A$. Then there are constants $c_1, c_2 > 0$ that depend on $\mathbb{P}_0$ only through α, A, and d such that, for any $\eta \in (0, 1]$, the concentration inequality $\mathbb{P}_0^N(W_p(\mathbb{P}_0, \hat{\mathbb{P}}) \leq r_N) \geq 1-\eta$ holds whenever r exceeds

$$r(d, N, \eta) = \begin{cases} \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{\min\{1/d, 1/2\}} & \text{if } N \geq \dfrac{\log(c_1/\eta)}{c_2}, \\ \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{1/\alpha} & \text{if } N < \dfrac{\log(c_1/\eta)}{c_2}. \end{cases} \tag{10.13}$$

The result remains valid for $p = d/2$ but with a more complicated formula for $r(d, N, \eta)$ (Fournier and Guillin 2015, Theorem 2). Intuitively, Theorem 10.2 asserts that any p-Wasserstein ball $\hat{\mathcal{P}}_N$ of $r_N \geq r(d, N, \eta)$ around $\hat{\mathbb{P}}_N$ represents a $(1 - \eta)$-confidence set for the unknown data-generating distribution $\mathbb{P}_0$. For uncertainty dimensions $d > 2$, the critical radius $r(d, N, \eta)$ of this confidence set decays as $O(N^{-1/d})$. In other words, to reduce the critical radius by 50%, the sample size must increase by 2^d. Unfortunately, this curse of dimensionality is fundamental, and the decay rate of $r(d, N, \eta)$ is essentially optimal (Fournier and Guillin 2015, § 1.3). Explicit constants c_1 and c_2 are provided by Fournier (2023).

Generic measure concentration bounds suffer from a curse of dimensionality. Shafieezadeh-Abadeh *et al.* (2019) and Wu *et al.* (2022) show that this curse can be overcome in the context of linear prediction models by projecting Z to a one-dimensional random variable, yielding the parametric convergence rate $O(N^{-1/2})$. Nietert *et al.* (2024*a*) develop a similar approach for rank-k linear models, where

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

$2 < k < d$, and achieve an improved rate of $O(N^{-1/k})$ based on k-sliced Wasserstein distances. The 1-sliced Wasserstein distance is also used by Olea *et al.* (2022) to obtain the parametric rate $O(N^{-1/2})$ for a class of regression problems.

We conclude this section by highlighting that the DRO approach admits instance-dependent regret bounds, which essentially depend on no complexity measures of the decision space or the loss function. Instead, they only depend on the complexity of the optimal solution x_0 through the DRO regularizer $\hat{R}_N(x_0)$. Zeng and Lam (2022, Theorem 4.1) and Nietert *et al.* (2024*a*, Theorem 1) establish such bounds for DRO problems over the ambiguity set (10.7) when D is the maximum mean discrepancy and the (outlier-robust) Wasserstein distance, respectively. Similar instance-dependent guarantees for DRO problems with Wasserstein ambiguity sets have been developed by Hou *et al.* (2023).

10.3.2. Generalization bounds

An alternative approach to obtaining statistical guarantees leverages the union bound from probability theory and covering numbers or complexity measures from statistical learning theory. The first step consists in deriving an inequality of the form

$$\mathbb{P}_0^N(\mathbb{E}_{\mathbb{P}_0}[\ell(x, Z)] \leq \hat{L}_N(x)) \geq 1 - \eta \quad \text{for all } x \in \mathcal{X}, \tag{10.14}$$

where the loss certificate $\hat{L}_N(x)$ depends on the decision $x \in \mathcal{X}$. For example, a guarantee of the form (10.14) can be obtained by combining empirical Bernstein inequalities (Maurer and Pontil 2009) and a DRO model with a χ^2-divergence ambiguity set (Duchi and Namkoong 2019, Theorem 2). In this case, the certificate $\hat{L}_N(x)$ reduces to the sum of the expected loss under $\hat{\mathbb{P}}_N$ and a variance regularizer under $\mathbb{P}_0$. Alternatively, a guarantee of the form (10.14) can also be obtained by combining transport inequalities (Marton 1986, Talagrand 1996) and a DRO model with a Wasserstein ambiguity set (Gao 2023, Theorem 1). In this case, $\hat{L}_N(x)$ reduces to the sum of the expected loss under $\hat{\mathbb{P}}_N$ and a variation regularizer under $\mathbb{P}_0$. The second step consists in converting the individual guarantee (10.14) to a uniform guarantee. For example, if $\mathcal{X}$ is finite, this can easily be achieved by using the union bound. If $\mathcal{X}$ is uncountable, one may use one of several standard techniques. If the loss function is Lipschitz-continuous in $x \in \mathcal{X}$ uniformly across all $z \in \mathcal{Z}$ and $\mathcal{X}$ is compact, then one can discretize $\mathcal{X}$ by uniform gridding. In this case, the loss at an arbitrary point is uniformly approximated by the loss at the nearest grid point, and a uniform guarantee can again be obtained by using the union bound. However, the number of grid points needed for an ε-approximation is of the order $O((1/\varepsilon)^d)$, which is impractical in high dimensions d. A more sophisticated approach to discretizing $\mathcal{X}$ exploits structural knowledge of the loss function at multiple scales. However, obtaining tight approximation in high dimensions remains challenging. In order to mitigate the computational burden related to discretization, one may exploit several complexity measures that quantify the expressiveness of the functions $\ell(x, \cdot)$ for all $x \in \mathcal{X}$ such as the VC dimension or

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

the Rademacher complexity as well as its local version. Nonetheless, Rademacher complexities can be computationally challenging to compute. For full details we refer to Boucheron, Lugosi and Massart (2013), Vershynin (2018) and Wainwright (2019).

The last step consists in approximating the certificate $\hat{L}_N(x)$ by the worst-case expected loss over a data-driven ambiguity set $\hat{\mathcal{P}}_N$ based on the χ^2-divergence or a Wasserstein distance. The corresponding approximation error can be controlled by leveraging Taylor approximations as in Theorems 8.4 and 8.7 together with appropriate concentration inequalities. In summary, this procedure shows that the optimal value of a data-driven DRO problem over a χ^2-divergence or a Wasserstein ambiguity set provides a finite-sample upper confidence bound on the corresponding stochastic program under the unknown true distribution $\mathbb{P}_0$.

Duchi and Namkoong (2019) and Gao (2023) derive generalization bounds of this kind for χ^2-divergence and Wasserstein ambiguity sets, respectively, while Azizian, Iutzeler and Malick (2023*a*) extend their analysis to entropic regularized optimal transport ambiguity sets. All these bounds exhibit the parametric rate $O(N^{-1/2})$. In addition, Duchi and Namkoong (2019) demonstrate that, under certain curvature conditions, χ^2-divergence decision rules can achieve the fast rate $O(N^{-1})$.

Acknowledgements

This research was supported by the Swiss National Science Foundation under the NCCR Automation (grant agreement 51NF40_180545). The authors thank Nicolas Lanzetti, Mengmeng Li, Karthik Natarajan, Jakob Nylöf, Yves Rychener, Philipp Schneider, Buse Sen, Ehsan Sharifian, Bradley Sturt and Man-Chung Yue for their valuable feedback on the paper. We are responsible for all remaining errors.

References

C. Acerbi (2002), Spectral measures of risk: A coherent representation of subjective risk aversion, *J. Banking Finance* **26**, 1505–1518.

A. Ahmadi-Javid (2012), Entropic value-at-risk: A new coherent risk measure, *J. Optim. Theory Appl.* **155**, 1105–1123.

S. Ahmed (2006), Convexity and decomposition of mean-risk stochastic programs, *Math. Program.* **106**, 433–446.

M. Ajtai, J. Komlós and G. Tusnády (1984), On optimal matchings, *Combinatorica* **4**, 259–264.

F. Al Taha, S. Yan and E. Bitar (2023), A distributionally robust approach to regret optimal control using the Wasserstein distance, in *62nd IEEE Conference on Decision and Control (CDC)*, pp. 2768–2775.

S. M. Ali and S. D. Silvey (1966), A general class of coefficients of divergence of one distribution from another, *J. Royal Statist. Soc. Ser. B* **28**, 131–142.

J. M. Altschuler and E. Boix-Adsera (2023), Polynomial-time algorithms for multimarginal optimal transport problems with structure, *Math. Program.* **199**, 1107–1178.

L. Ambrosio, N. Gigli and G. Savaré (2008), *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Springer.

Y. An and R. Gao (2021), Generalization bounds for (Wasserstein) robust optimization, in *Advances in Neural Information Processing Systems 34* (M. Ranzato *et al.*, eds), Curran Associates, pp. 10382–10392.
B. Analui and G. C. Pflug (2014), On distributionally robust multiperiod stochastic optimization, *Comput. Manag. Sci.* **11**, 197–220.
M. Anthony and P. L. Bartlett (1999), *Neural Network Learning: Theoretical Foundations*, Cambridge University Press.
J. Anunrojwong, S. R. Balseiro and O. Besbes (2024), On the robustness of second-price auctions in prior-independent mechanism design, *Oper. Res.* Available at doi:10.1287/opre.2022.0428.
L. Aolaritei, N. Lanzetti, H. Chen and F. Dörfler (2022*a*), Uncertainty propagation via optimal transport ambiguity sets. Available at arXiv:2205.00343.
L. Aolaritei, S. Shafiee and F. Dörfler (2022*b*), Wasserstein distributionally robust estimation in high dimensions: Performance analysis and optimal hyperparameter tuning. Available at arXiv:2206.13269.
P. Artzner, F. Delbaen, J.-M. Eber and D. Heath (1999), Coherent measures of risk, *Math. Finance* **9**, 203–228.
C. Atkinson and A. F. Mitchell (1981), Rao's distance measure, *Sankhyā* **43**, 345–365.
W. Azizian, F. Iutzeler and J. Malick (2023*a*), Exact generalization guarantees for (regularized) Wasserstein distributionally robust models, in *Advances in Neural Information Processing Systems 36* (A. Oh *et al.*, eds), Curran Associates, pp. 14584–14596.
W. Azizian, F. Iutzeler and J. Malick (2023*b*), Regularization for Wasserstein distributionally robust optimization, *ESAIM Control Optim. Calc. Var.* **29**, 1–33.
F. Bach (2013), Learning with submodular functions: A convex optimization perspective, *Found. Trends Mach. Learn.* **6**, 145–373.
F. Bach (2019), Submodular functions: From discrete to continuous domains, *Math. Program.* **175**, 419–459.
X. Bai, G. He, Y. Jiang and J. Obloj (2023*a*), Wasserstein distributional robustness of neural networks, in *Advances in Neural Information Processing Systems 36* (A. Oh *et al.*, eds), Curran Associates, pp. 26322–26347.
Y. Bai, H. Lam and X. Zhang (2023*b*), A distributionally robust optimization framework for extreme event estimation. Available at arXiv:2301.01360.
R. Baire (1905), *Leçons sur les Fonctions Discontinues*, Gauthier-Villars.
S. Banach (1938), Über homogene Polynome in (L^2), *Studia Math.* **7**, 36–44.
C. Bandi and D. Bertsimas (2014), Optimal design for multi-item auctions: A robust optimization approach, *Math. Oper. Res.* **39**, 1012–1038.
D. Bartl, S. Drapeau, J. Oblój and J. Wiesel (2021), Sensitivity analysis of Wasserstein distributionally robust optimization problems, *Proc. Royal Soc. Ser. A* **477**, art. 20210176.
T. Başar (1977), Optimum Fisherian information for multivariate distributions, *Ann. Statist.* **5**, 1240–1244.
T. Başar (1983), The Gaussian test channel with an intelligent jammer, *IEEE Trans. Inform. Theory* **29**, 152–157.
T. Başar and T. Ü. Başar (1984), A bandwidth expanding scheme for communication channels with noiseless feedback in the presence of unknown jamming noise, *J. Franklin Institute* **317**, 73–88.
T. Başar and P. Bernhard (1995), *$\mathcal{H}_\infty$-optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Springer.

T. Başar and M. Max (1973), A multistage pursuit-evasion game that admits a Gaussian random process as a maximin control policy, *Stochastics* **1**, 25–69.
T. Başar and M. Mintz (1972), Minimax terminal state estimation for linear plants with unknown forcing functions, *Internat. J. Control* **16**, 49–69.
T. Başar and M. Mintz (1973), On a minimax estimate for the mean of a normal random vector under a generalized quadratic loss function, *Ann. Statist.* **1**, 127–134.
T. Başar and Y. W. Wu (1985), A complete characterization of minimax and maximin encoder-decoder policies for communication channels with incomplete statistical description, *IEEE Trans. Inform. Theory* **31**, 482–489.
T. Başar and Y. W. Wu (1986), Solutions to a class of minimax decision problems arising in communication systems, *J. Optim. Theory Appl.* **51**, 375–404.
T. Ü. Başar and T. Başar (1982), Optimum coding and decoding schemes for the transmission of a stochastic process over a continuous-time stochastic channel with partially unknown statisticst, *Stochastics* **8**, 213–237.
H. I. Bayrak, Ç. Koçyiğit, D. Kuhn and M. C. Pınar (2025), Distributionally robust optimal allocation with costly verification, *Oper. Res.* Available at doi:10.1287/opre.2022.0662.
G. Bayraksan and D. K. Love (2015), Data-driven stochastic programming using phi-divergences, *INFORMS TutORials in Operations Research*, pp. 1–19. Available at doi:10.1287/educ.2015.0134.
E. M. L. Beale (1955), On minimizing a convex function subject to linear inequalities, *J. Royal Statist. Soc. Ser. B* **17**, 173–184.
A. Beck and A. Ben-Tal (2009), Duality in robust optimization: Primal worst equals dual best, *Oper. Res. Lett.* **37**, 1–6.
R. Belbasi, A. Selvi and W. Wiesemann (2023), It's all in the mix: Wasserstein machine learning with mixed features. Available at arXiv:2312.12230.
A. Ben-Tal and E. Hochman (1972), More bounds on the expectation of a convex function of a random variable, *J. Appl. Probab.* **9**, 803–812.
A. Ben-Tal and A. Nemirovski (1998), Robust convex optimization, *Math. Oper. Res.* **23**, 769–805.
A. Ben-Tal and A. Nemirovski (1999*a*), Robust solutions of uncertain linear programs, *Oper. Res. Lett.* **25**, 1–13.
A. Ben-Tal and A. Nemirovski (1999*b*), Robust truss topology design via semidefinite programming, *SIAM J. Optim.* **7**, 991–1016.
A. Ben-Tal and A. Nemirovski (2000), Robust solutions of linear programming problems contaminated with uncertain data, *Math. Program.* **88**, 411–424.
A. Ben-Tal and A. Nemirovski (2001), *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, SIAM.
A. Ben-Tal and A. Nemirovski (2002), Robust optimization–methodology and applications, *Math. Program.* **92**, 453–480.
A. Ben-Tal and M. Teboulle (1986), Expected utility, penalty functions, and duality in stochastic nonlinear programming, *Manag. Sci.* **32**, 1445–1466.
A. Ben-Tal and M. Teboulle (2007), An old–new concept of convex risk measures: The optimized certainty equivalent, *Math. Finance* **17**, 449–476.
A. Ben-Tal, A. Ben-Israel and M. Teboulle (1991), Certainty equivalents and information measures: Duality and extremal principles, *J. Math. Anal. Appl.* **157**, 211–236.
A. Ben-Tal, D. den Hertog and J.-P. Vial (2015*a*), Deriving robust counterparts of nonlinear uncertain inequalities, *Math. Program.* **149**, 265–299.

A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg and G. Rennen (2013), Robust solutions of optimization problems affected by uncertain probabilities, *Manag. Sci.* **59**, 341–357.

A. Ben-Tal, L. El Ghaoui and A. Nemirovski (2009), *Robust Optimization*, Princeton University Press.

A. Ben-Tal, E. Hazan, T. Koren and S. Mannor (2015*b*), Oracle-based robust optimization via online learning, *Oper. Res.* **63**, 628–638.

A. Bennouna and B. P. G. Van Parys (2021), Learning and decision-making with data: Optimal formulations and phase transitions. Available at arXiv:2109.06911.

A. Bennouna and B. P. G. Van Parys (2023), Holistic robust data-driven decisions. Available at arXiv:2207.09560.

A. Bennouna, R. Lucas and B. P. G. Van Parys (2023), Certified robust neural networks: Generalization and corruption resistance, in *40th International Conference on Machine Learning*, Vol. 202 of Proceedings of Machine Learning Research, PMLR, pp. 2092–2112.

C. Berge (1963), *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*, Courier Corporation.

D. Bergemann and K. H. Schlag (2008), Pricing without priors, *J. Eur. Econom. Assoc.* **6**, 560–569.

D. S. Bernstein (2009), *Matrix Mathematics: Theory, Facts, and Formulas*, Princeton University Press.

D. Bertsimas and D. den Hertog (2022), *Robust and Adaptive Optimization*, Dynamic Ideas.

D. Bertsimas and I. Popescu (2002), On the relation between option and stock prices: A convex optimization approach, *Oper. Res.* **50**, 358–374.

D. Bertsimas and I. Popescu (2005), Optimal inequalities in probability theory: A convex optimization approach, *SIAM J. Optim.* **15**, 780–804.

D. Bertsimas and J. Sethuraman (2000), Moment problems and semidefinite optimization, in *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications* (H. Wolkowicz, R. Saigal and L. Vandenberghe, eds), Springer, pp. 469–509.

D. Bertsimas and M. Sim (2004), The price of robustness, *Oper. Res.* **52**, 35–53.

D. Bertsimas and B. P. G. Van Parys (2022), Bootstrap robust prescriptive analytics, *Math. Program.* **195**, 39–78.

D. Bertsimas, D. B. Brown and C. Caramanis (2011), Theory and applications of robust optimization, *SIAM Rev.* **53**, 464–501.

D. Bertsimas, D. den Hertog and J. Pauphilet (2021), Probabilistic guarantees in robust optimization, *SIAM J. Optim.* **31**, 2893–2920.

D. Bertsimas, X. V. Doan, K. Natarajan and C.-P. Teo (2010), Models for minimax stochastic linear optimization problems with risk aversion, *Math. Oper. Res.* **35**, 580–602.

D. Bertsimas, V. Gupta and N. Kallus (2018*a*), Data-driven robust optimization, *Math. Program.* **167**, 235–292.

D. Bertsimas, V. Gupta and N. Kallus (2018*b*), Robust sample average approximation, *Math. Program.* **171**, 217–282.

D. Bertsimas, K. Natarajan and C.-P. Teo (2004), Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds, *SIAM J. Optim.* **15**, 185–209.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

D. Bertsimas, K. Natarajan and C.-P. Teo (2006*a*), Persistence in discrete optimization under data uncertainty, *Math. Program.* **108**, 251–274.
D. Bertsimas, K. Natarajan and C.-P. Teo (2006*b*), Tight bounds on expected order statistics, *Probab. Engrg Inform. Sci.* **20**, 667–686.
D. Bertsimas, S. Shtern and B. Sturt (2022), Two-stage sample robust optimization, *Oper. Res.* **70**, 624–640.
D. Bertsimas, S. Shtern and B. Sturt (2023), A data-driven approach to multistage stochastic linear optimization, *Manag. Sci.* **69**, 51–74.
R. Bhatia, T. Jain and Y. Lim (2018), Strong convexity of sandwiched entropies and related optimization problems, *Rev. Math. Phys.* **30**, art. 1850014.
R. Bhatia, T. Jain and Y. Lim (2019), On the Bures–Wasserstein distance between positive definite matrices, *Expo. Math.* **37**, 165–191.
C. Bhattacharyya (2004), Second order cone programming formulations for feature selection, *J. Mach. Learn. Res.* **5**, 1417–1433.
P. Billingsley (2013), *Convergence of Probability Measures*, Wiley.
J. R. Birge and F. Louveaux (2011), *Introduction to Stochastic Programming*, Springer.
J. R. Birge and R. J.-B. Wets (1986), Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse, *Math. Program. Study* **27**, 54–102.
C. M. Bishop (2006), *Pattern Recognition and Machine Learning*, Springer.
J. Blanchet and Y. Kang (2020), Semi-supervised learning based on distributionally robust optimization, in *Data Analysis and Applications 3* (A. Makrides, A. Karagrigoriou and C. H. Skiadas, eds), Wiley, pp. 1–33.
J. Blanchet and Y. Kang (2021), Sample out-of-sample inference based on Wasserstein distance, *Oper. Res.* **69**, 985–1013.
J. Blanchet and K. Murthy (2019), Quantifying distributional model risk via optimal transport, *Math. Oper. Res.* **44**, 565–600.
J. Blanchet and A. Shapiro (2024), Statistical limit theorems in distributionally robust optimization, in *Proceedings of the Winter Simulation Conference (WSC '23)*, IEEE Press, pp. 31–45.
J. Blanchet, L. Chen and X. Y. Zhou (2022*a*), Distributionally robust mean–variance portfolio selection with Wasserstein distances, *Manag. Sci.* **68**, 6382–6410.
J. Blanchet, P. W. Glynn, J. Yan and Z. Zhou (2019*a*), Multivariate distributionally robust convex regression under absolute error loss, in *Advances in Neural Information Processing Systems 32* (H. Wallach *et al.*, eds), Curran Associates, pp. 11817–11826.
J. Blanchet, F. He and K. Murthy (2020), On distributionally robust extreme value analysis, *Extremes* **23**, 317–347.
J. Blanchet, Y. Kang and K. Murthy (2019*b*), Robust Wasserstein profile inference and applications to machine learning, *J. Appl. Probab.* **56**, 830–857.
J. Blanchet, D. Kuhn, J. Li and B. Taşkesen (2023), Unifying distributionally robust optimization via optimal transport theory. Available at arXiv:2308.05414.
J. Blanchet, H. Lam, Y. Liu and R. Wang (2024*a*), Convolution bounds on quantile aggregation, *Oper. Res.* Available at doi:10.1287/opre.2021.0765.
J. Blanchet, J. Li, S. Lin and X. Zhang (2024*b*), Distributionally robust optimization and robust statistics. Available at arXiv:2401.14655.
J. Blanchet, K. Murthy and V. A. Nguyen (2021), Statistical analysis of Wasserstein distributionally robust estimators, *INFORMS TutORials in Operations Research*, pp. 227–254. Available at doi:10.1287/educ.2021.0233.

J. Blanchet, K. Murthy and N. Si (2022*b*), Confidence regions in Wasserstein distributionally robust estimation, *Biometrika* **109**, 295–315.
J. Blanchet, K. Murthy and F. Zhang (2022*c*), Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes, *Math. Oper. Res.* **47**, 1500–1529.
N. E. Blankenstein, E. A. Crone, W. van den Bos and A. C. K. van Duijvenvoorde (2016), Adolescents display distinctive tolerance to ambiguity and to uncertainty during risky decision making, *Develop. Neuropsychol.* **41**, 77–92.
E. Boissard and T. Le Gouic (2014), On the mean speed of convergence of empirical and occupation measures in Wasserstein distance, *Ann. Inst. Henri Poincaré Probab. Statist.* **50**, 539–563.
F. Bolley, A. Guillin and C. Villani (2007), Quantitative concentration inequalities for empirical measures on non-compact spaces, *Probab. Theory Related Fields* **137**, 541–593.
G. Boole (1854), *An Investigation of the Laws of Thought*, Walton and Maberly.
S. Bose and A. Daripa (2009), A dynamic mechanism and surplus extraction under ambiguity, *J. Econom. Theory* **144**, 2084–2114.
D. Boskos, J. Cortés and S. Martínez (2020), Data-driven ambiguity sets with probabilistic guarantees for dynamic processes, *IEEE Trans. Automat. Control* **66**, 2991–3006.
P. Bossaerts, P. Ghirardato, S. Guarnaschelli and W. R. Zame (2010), Ambiguity in asset markets: Theory and experiment, *Rev. Financ. Stud.* **23**, 1325–1359.
S. Boucheron, G. Lugosi and P. Massart (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.
O. Bousquet, S. Boucheron and G. Lugosi (2004), Introduction to statistical learning theory, in *Advanced Lectures on Machine Learning* (O. Bousquet, U. von Luxburg and G. Rätsch, eds), Springer, pp. 169–207.
G. E. P. Box (1953), Non-normality and tests on variances, *Biometrika* **40**, 318–335.
G. E. P. Box (1979), Robustness in the strategy of scientific model building, in *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds), Academic Press, pp. 201–236.
Y. Brenier (1991), Polar factorization and monotone rearrangement of vector-valued functions, *Commun. Pure Appl. Math.* **44**, 375–417.
H. Brezis (2011), *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer.
J. Brugman, J. S. H. Van Leeuwaarden and C. Stegehuis (2022), Sharpest possible clustering bounds using robust random graph analysis, *Phys. Rev. E* **106**, art. 064311.
M. Buckert, C. Schwieren, B. M. Kudielka and C. J. Fiebach (2014), Acute stress affects risk taking but not ambiguity aversion, *Front. Neurosci.* **8**, art. 82.
N. Bui, D. Nguyen and V. A. Nguyen (2022), Counterfactual plans under distributional ambiguity, in *International Conference on Learning Representations (ICLR 2022)*.
L. Bungert, N. García Trillos and R. Murray (2023), The geometry of adversarial training in binary classification, *Inform. Inference* **12**, 921–968.
L. Bungert, T. Laux and K. Stinson (2024), A mean curvature flow arising in adversarial training, *J. Math. Pures Appl.* **192**, art. 103625.
L. Cabantous (2007), Ambiguity aversion in the field of insurance: Insurers' attitude to imprecise and conflicting probability estimates, *Theory Decis.* **62**, 219–240.
J. Cai, J. Y.-M. Li and T. Mao (2023), Distributionally robust optimization under distorted expectations, *Oper. Res.* **73**, 969–985.

G. C. Calafiore (2007), Ambiguous risk measures and optimal robust portfolios, *SIAM J. Optim.* **18**, 853–877.
G. C. Calafiore and M. C. Campi (2005), Uncertain convex programs: Randomized solutions and confidence levels, *Math. Program.* **102**, 25–46.
G. C. Calafiore and M. C. Campi (2006), The scenario approach to robust control design, *IEEE Trans. Automat. Control* **51**, 742–753.
G. C. Calafiore and L. El Ghaoui (2006), On distributionally robust chance-constrained linear programs, *J. Optim. Theory Appl.* **130**, 1–22.
G. C. Calafiore, F. Dabbene and R. Tempo (2011), Research on probabilistic methods for control system design, *Automatica* **47**, 1279–1293.
M. C. Campi and A. Caré (2013), Random convex programs with L_1-regularization: Sparsity and generalization, *SIAM J. Control Optim.* **51**, 3532–3557.
M. C. Campi and S. Garatti (2008), The exact feasibility of randomized solutions of uncertain convex programs, *SIAM J. Optim.* **19**, 1211–1230.
M. C. Campi and S. Garatti (2011), A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality, *J. Optim. Theory Appl.* **148**, 257–280.
M. C. Campi and S. Garatti (2018), Wait-and-judge scenario optimization, *Math. Program.* **167**, 155–189.
A. Caré, S. Garatti and M. C. Campi (2014), FAST: Fast algorithm for the scenario technique, *Oper. Res.* **62**, 662–671.
Y. Carmon and D. Hausler (2022), Distributionally robust optimization via ball oracle acceleration, in *Advances in Neural Information Processing Systems 35* (S. Koyejo *et al.*, eds), Curran Associates, pp. 35866–35879.
G. Carroll (2017), Robustness and separation in multidimensional screening, *Econometrica* **85**, 453–488.
T. Champion, L. De Pascale and P. Juutinen (2008), The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps, *SIAM J. Math. Anal.* **40**, 1–20.
G. Chan, B. Van Parys and A. Bennouna (2024), From distributional robustness to robust statistics: A confidence sets perspective. Available at arXiv:2410.14008.
P. Chebyshev (1874), Sur les valeurs limites des intégrales, *J. Math. Pures Appl.* **19**, 157–160.
L. Chen and M. Sim (2024), Robust CARA optimization, *Oper. Res.* Available at doi:10.1287/opre.2021.0654.
L. Chen, C. Fu, F. Si, M. Sim and P. Xiong (2024*a*), Robust optimization with moment-dispersion ambiguity, *Oper. Res.* Available at doi:10.1287/opre.2023.0579.
L. Chen, S. He and S. Zhang (2011), Tight bounds for some risk measures, with applications to robust portfolio selection, *Oper. Res.* **59**, 847–865.
L. Chen, W. Ma, K. Natarajan, D. Simchi-Levi and Z. Yan (2022), Distributionally robust linear and discrete optimization with marginals, *Oper. Res.* **70**, 1822–1834.
L. Chen, D. Padmanabhan, C. C. Lim and K. Natarajan (2020), Correlation robust influence maximization, in *Advances in Neural Information Processing Systems 33* (H. Larochelle *et al.*, eds), Curran Associates, pp. 7078–7089.
R. Chen and I. C. Paschalidis (2018), A robust learning approach for regression models based on distributionally robust optimization, *J. Mach. Learn. Res.* **19**, 517–564.
R. Chen and I. C. Paschalidis (2019), Selecting optimal decisions via distributionally robust nearest-neighbor regression, in *Advances in Neural Information Processing Systems 32* (H. Wallach *et al.*, eds), Curran Associates, pp. 749–759.

W. Chen, M. Sim, J. Sun and C.-P. Teo (2010), From CVaR to uncertainty set: Implications in joint chance-constrained optimization, *Oper. Res.* **58**, 470–485.
Z. Chen, Z. Hu and R. Wang (2024*b*), Screening with limited information: A dual perspective, *Oper. Res.* **72**, 1487–1504.
Z. Chen, D. Kuhn and W. Wiesemann (2024*c*), Data-driven chance constrained programs over Wasserstein balls, *Oper. Res.* **72**, 410–424.
Z. Chen, M. Sim and H. Xu (2019), Distributionally robust optimization with infinitely constrained ambiguity sets, *Oper. Res.* **67**, 1328–1344.
J. Cheng, E. Delage and A. Lisser (2014), Distributionally robust stochastic knapsack problem, *SIAM J. Optim.* **24**, 1485–1506.
A. Cherukuri and J. Cortés (2019), Cooperative data-driven distributionally robust optimization, *IEEE Trans. Automat. Control* **65**, 4400–4407.
L. Chizat (2022), Sparse optimization on measures with over-parameterized gradient descent, *Math. Program.* **194**, 487–532.
L. Chizat and F. Bach (2018), On the global convergence of gradient descent for over-parameterized models using optimal transport, in *Advances in Neural Information Processing Systems 31* (S. Bengio *et al.*, eds), Curran Associates, pp. 3040–3050.
P. Clément and W. Desch (2008), Wasserstein metric and subordination, *Studia Math.* **189**, 35–52.
J. Coulson, J. Lygeros and F. Dörfler (2021), Distributionally robust chance constrained data-enabled predictive control, *IEEE Trans. Automat. Control* **67**, 3289–3304.
T. Cover and J. Thomas (2006), *Elements of Information Theory*, Wiley.
H. Cramér (1938), Sur un nouveau théorème-limite de la théorie des probabilités, *Actualités Sci. Indust.* **736**, 5–23.
H. Cramér (1946), *Mathematical Methods of Statistics*, Princeton University Press.
I. Csiszár (1963), Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Publ. Math. Inst. Hungar. Acad. Sci.* **8**, 85–108.
I. Csiszár (1967), Information-type measures of difference of probability distributions and indirect observation, *Studia Sci. Math. Hungar.* **2**, 229–318.
G. B. Dantzig (1955), Linear programming under uncertainty, *Manag. Sci.* **1**, 197–206.
G. B. Dantzig (1956), *The Simplex Method*, RAND Corporation.
B. Das, A. Dhara and K. Natarajan (2021), On the heavy-tail behavior of the distributionally robust newsvendor, *Oper. Res.* **69**, 1077–1099.
D. P. De Farias and B. Van Roy (2004), On constraint sampling in the linear programming approach to approximate dynamic programming, *Math. Oper. Res.* **29**, 462–478.
E. Delage and D. A. Iancu (2015), Robust multistage decision making, *INFORMS TutORials in Operations Research*, pp. 20–46. Available at doi:10.1287/educ.2015.0139.
E. Delage and Y. Ye (2010), Distributionally robust optimization under moment uncertainty with application to data-driven problems, *Oper. Res.* **58**, 595–612.
E. Delage, D. Kuhn and W. Wiesemann (2019), 'Dice'-sion-making under uncertainty: When can a random decision reduce risk?, *Manag. Sci.* **65**, 3282–3301.
F. Delbaen (2002), Coherent risk measures on general probability spaces, in *Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann* (K. Sandmann and P. J. Schönbucher, eds), Springer, pp. 1–37.
A. Dembo and O. Zeitouni (2009), *Large Deviations Techniques and Applications*, Springer.

V. DeMiguel and F. J. Nogales (2009), Portfolio selection with robust estimation, *Oper. Res.* **57**, 560–577.
V. DeMiguel, L. Garlappi and R. Uppal (2009), Optimal versus naive diversification: How inefficient is the $1/n$ portfolio strategy?, *Rev. Financ. Stud.* **22**, 1915–1953.
A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru and F. Roli (2019), Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks, in *28th USENIX Security Symposium*, pp. 321–338.
S. Dereich, M. Scheutzow and R. Schottstedt (2013), Constructive quantization: Approximation by empirical measures, *Ann. Inst. Henri Poincaré Probab. Statist.* **49**, 1183–1203.
S. Dharmadhikari and K. Joag-Dev (1988), *Unimodality, Convexity, and Applications*, Elsevier.
I. Diakonikolas and D. M. Kane (2023), *Algorithmic High-Dimensional Robust Statistics*, Cambridge University Press.
I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra and A. Stewart (2019), Robust estimators in high-dimensions without the computational intractability, *SIAM J. Comput.* **48**, 742–864.
M. Z. Diao, K. Balasubramanian, S. Chewi and A. Salim (2023), Forward–backward Gaussian variational inference via JKO in the Bures–Wasserstein space, in *40th International Conference on Machine Learning*, Vol. 202 of Proceedings of Machine Learning Research, PMLR, pp. 7960–7991.
S. G. Dimmock, R. Kouwenberg and P. P. Wakker (2016), Ambiguity attitudes in a large representative sample, *Manag. Sci.* **62**, 1363–1380.
X. V. Doan and K. Natarajan (2012), On the complexity of nonoverlapping multivariate marginal bounds for probabilistic combinatorial optimization problems, *Oper. Res.* **60**, 138–149.
X. V. Doan, X. Li and K. Natarajan (2015), Robustness to dependency in portfolio optimization using overlapping marginals, *Oper. Res.* **63**, 1468–1488.
V. Dobrić and J. E. Yukich (1995), Asymptotics for transportation cost in high dimensions, *J. Theoret. Probab.* **8**, 97–118.
S. P. Dokov and D. P. Morton (2005), Second-order lower bounds on the expectation of a convex function, *Math. Oper. Res.* **30**, 662–677.
D. L. Donoho and R. C. Liu (1988), The 'automatic' robustness of minimum distance functionals, *Ann. Statist.* **16**, 552–586.
M. D. Donsker and S. R. S. Varadhan (1983), Asymptotic evaluation of certain Markov process expectations for large time IV, *Commun. Pure Appl. Math.* **36**, 183–212.
D. C. Dowson and B. V. Landau (1982), The Fréchet distance between multivariate normal distributions, *J. Multivariate Anal.* **12**, 450–455.
J. C. Doyle, K. Glover, P. Khargonekar and B. Francis (1989), Robust control of time-delay systems, *IEEE Trans. Automat. Control* **34**, 674–683.
J. C. Duchi and H. Namkoong (2019), Variance-based regularization with convex objectives, *J. Mach. Learn. Res.* **20**, 1–55.
J. C. Duchi and H. Namkoong (2021), Learning models with uniform performance via distributionally robust optimization, *Ann. Statist.* **49**, 1378–1406.
J. C. Duchi, P. W. Glynn and H. Namkoong (2021), Statistics of robust optimization: A generalized empirical likelihood approach, *Math. Oper. Res.* **46**, 946–969.
J. Duchi, T. Hashimoto and H. Namkoong (2023), Distributionally robust losses for latent covariate mixtures, *Oper. Res.* **71**, 649–664.

R. M. Dudley (1969), The speed of mean Glivenko–Cantelli convergence, *Ann. Math. Statist.* **40**, 40–50.
J. H. Dulá and R. V. Murthy (1992), A Tchebysheff-type bound on the expectation of sublinear polyhedral functions, *Oper. Res.* **40**, 914–922.
G. E. Dullerud and F. Paganini (2001), *A Course in Robust Control Theory: A Convex Approach*, Springer.
J. Dupačová (2006), Stress testing via contamination, in *Coping with Uncertainty: Modeling and Policy Issues* (K. Marti *et al.*, eds), Springer, pp. 29–46.
J. Dupacová and R. Wets (1988), Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems, *Ann. Statist.* **16**, 1517–1549.
J. Dupačová (1966), On minimax solutions of stochastic linear programming problems, *Časopis pro pěstování matematiky* **91**, 423–430.
J. Dupačová (1987), The minimax approach to stochastic programming and an illustrative application, *Stochastics* **20**, 73–88.
J. Dupačová (1994), Applications of stochastic programming under incomplete information, *J. Comput. Appl. Math.* **56**, 113–125.
P. Dupuis and Y. Mao (2022), Formulation and properties of a divergence used to compare probability measures without absolute continuity, *ESAIM Control Optim. Calc. Var.* **28**, art. 10.
D. Duque and D. P. Morton (2020), Distributionally robust stochastic dual dynamic programming, *SIAM J. Optim.* **30**, 2841–2865.
M. Dyer and L. Stougie (2006), Computational complexity of stochastic programming problems, *Math. Program.* **106**, 423–432.
H. P. Edmundson (1956), Bounds on the expectation of a convex function of a random variable. The Rand Corporation Paper 982, Santa Monica, CA.
L. El Ghaoui and H. Lebret (1998*a*), Robust optimization of control systems: A convex approach, *IEEE Trans. Automat. Control* **43**, 309–319.
L. El Ghaoui and H. Lebret (1998*b*), Robust solutions to least-squares problems with uncertain data, *SIAM J. Matrix Anal. Appl.* **18**, 1035–1064.
L. El Ghaoui, M. Oks and F. Oustry (2003), Worst-case value-at-risk and robust portfolio optimization: A conic programming approach, *Oper. Res.* **51**, 543–556.
L. El Ghaoui, F. Oustry and H. Lebret (1998), Robust solutions to uncertain semidefinite programs, *SIAM J. Optim.* **9**, 33–52.
R. S. Ellis (2007), *Entropy, Large Deviations, and Statistical Mechanics*, Springer.
D. Ellsberg (1961), Risk, ambiguity, and the Savage axioms, *Quart. J. Econom.* **75**, 643–669.
P. Embrechts and G. Puccetti (2006), Bounds for functions of multivariate risks, *J. Multivariate Anal.* **97**, 526–547.
L. G. Epstein and J. Miao (2003), A two-person dynamic equilibrium under ambiguity, *J. Econom. Dynam. Control* **27**, 1253–1288.
E. Erdoğan and G. Iyengar (2006), Ambiguous chance constrained problems and robust optimization, *Math. Program.* **107**, 37–61.
Y. Ermoliev, A. A. Gaivoronski and C. Nedeva (1985), Stochastic optimization problems with incomplete information on distribution functions, *SIAM J. Control Optim.* **23**, 697–716.
A. Esteban-Pérez and J. M. Morales (2022), Distributionally robust stochastic programs with side information based on trimmings, *Math. Program.* **195**, 1069–1105.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

F. Farnia and D. Tse (2016), A minimax approach to supervised learning, in *Advances in Neural Information Processing Systems 29* (D. Lee *et al.*, eds), Curran Associates, pp. 4240–4248.

W. Fenchel (1953), *Convex Cones, Sets, and Functions*, Princeton University Press.

C. Finlay and A. M. Oberman (2021), Scaleable input gradient regularization for adversarial robustness, *Mach. Learn. Appl.* **3**, art. 100017.

G. B. Folland (1999), *Real Analysis: Modern Techniques and Their Applications*, Wiley.

H. Föllmer and A. Schied (2008), *Stochastic Finance. An Introduction in Discrete Time*, de Gruyter.

N. Fournier (2023), Convergence of the empirical measure in expected Wasserstein distance: Non-asymptotic explicit bounds in $\mathbb{R}^d$, *ESAIM Probab. Statist.* **27**, 749–775.

N. Fournier and A. Guillin (2015), On the rate of convergence in Wasserstein distance of the empirical measure, *Probab. Theory Related Fields* **162**, 707–738.

N. Frank and J. Niles-Weed (2024*a*), The adversarial consistency of surrogate risks for binary classification, in *Advances in Neural Information Processing Systems 36* (A. Oh *et al.*, eds), Curran Associates, pp. 41343–41354.

N. S. Frank and J. Niles-Weed (2024*b*), Existence and minimax theorems for adversarial surrogate risks in binary classification, *J. Mach. Learn. Res.* **25**, 1–41.

K. Frauendorfer (1992), *Stochastic Two-Stage Programming*, Springer.

M. Fréchet (1935), Généralisation du théorème des probabilités totales, *Fund. Math.* **25**, 379–387.

A. A. Gaivoronski (1991), A numerical method for solving stochastic programming problems with moment constraints on a distribution function, *Ann. Oper. Res.* **31**, 347–370.

G. Gallego and I. Moon (1993), The distribution free newsboy problem: Review and extensions, *J. Oper. Res. Soc.* **44**, 825–834.

A. Ganguly and T. Sutter (2023), Optimal learning via moderate deviations theory. Available at arXiv:2305.14496.

R. Gao (2023), Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality, *Oper. Res.* **71**, 2291–2306.

R. Gao and A. J. Kleywegt (2023), Distributionally robust stochastic optimization with Wasserstein distance, *Math. Oper. Res.* **48**, 603–655.

R. Gao, R. Arora and Y. Huang (2024*a*), Data-driven multistage distributionally robust linear optimization with nested distance. Available at arXiv:2407.16346.

R. Gao, X. Chen and A. J. Kleywegt (2017), Wasserstein distributional robustness and regularization in statistical learning. Available at arXiv:1712.06050.

R. Gao, X. Chen and A. J. Kleywegt (2024*b*), Wasserstein distributionally robust optimization and variation regularization, *Oper. Res.* **72**, 1177–1191.

R. Gao, L. Xie, Y. Xie and H. Xu (2018), Robust hypothesis testing using Wasserstein uncertainty sets, in *Advances in Neural Information Processing Systems 31* (S. Bengio *et al.*, eds), Curran Associates, pp. 7902–7912.

C. A. García Trillos and N. García Trillos (2022), On the regularized risk of distributionally robust learning over deep neural networks, *Res. Math. Sci.* **9**, art. 54.

N. García Trillos and M. Jacobs (2023), An analytical and geometric perspective on adversarial robustness, *Notices Amer. Math. Soc.* **70**, 1193–1204.

N. García Trillos and R. Murray (2022), Adversarial classification: Necessary conditions and geometric flows, *J. Mach. Learn. Res.* **23**, 1–38.

N. García Trillos, M. Jacobs and J. Kim (2023), The multimarginal optimal transport formulation of adversarial multiclass classification, *J. Mach. Learn. Res.* **24**, 1–56.
H. Gassmann and W. T. Ziemba (1986), A tight upper bound for the expectation of a convex function of a multivariate random variable, in *Stochastic Programming 84 Part I* (A. Prékopa and R. J.-B. Wets, eds), Vol. 27 of Mathematical Programming Studies, Springer, pp. 39–53.
M. Gelbrich (1990), On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces, *Math. Nachr.* **147**, 185–203.
G. Georgakopoulos, D. Kavvadias and C. H. Papadimitriou (1988), Probabilistic satisfiability, *J. Complexity* **4**, 1–11.
R. Ghanem, D. Higdon and H. Owhadi (2017), *Handbook of Uncertainty Quantification*, Springer.
S. Ghosh, M. Squillante and E. Wollega (2021), Efficient stochastic gradient descent for learning with distributionally robust optimization, in *Advances in Neural Information Processing Systems 34* (M. Ranzato *et al.*, eds), Curran Associates, pp. 28310–28322.
I. Gilboa and D. Schmeidler (1989), Maxmin expected utility with a non-unique prior, *J. Math. Econom.* **18**, 141–153.
C. R. Givens and R. M. Shortt (1984), A class of Wasserstein metrics for probability distributions, *Michigan Math. J.* **31**, 231–240.
M. Goerigk and J. Kurtz (2023), Data-driven robust optimization using deep neural networks, *Comput. Oper. Res.* **151**, art. 106087.
I. J. Goodfellow, J. Shlens and C. Szegedy (2015), Explaining and harnessing adversarial examples, in *International Conference on Learning Representations (ICLR 2015)*.
J.-Y. Gotoh, M. J. Kim and A. E. Lim (2018), Robust empirical optimization is almost the same as mean–variance optimization, *Oper. Res. Lett.* **46**, 448–452.
J.-Y. Gotoh, M. J. Kim and A. E. Lim (2021), Calibration of distributionally robust empirical optimization models, *Oper. Res.* **69**, 1630–1650.
N. Gravin and P. Lu (2018), Separation in correlation-robust monopolist problem with budget, in *2018 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2069–2080.
M. Green and D. J. N. Limebeer (1995), H-infinity control theory: A tutorial, *Automatica* **31**, 213–222.
G. Gül and A. M. Zoubir (2017), Minimax robust hypothesis testing, *IEEE Trans. Inform. Theory* **63**, 5572–5587.
I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville (2017), Improved training of Wasserstein GANs, in *Advances in Neural Information Processing Systems 30* (I. Guyon *et al.*, eds), Curran Associates, pp. 5769–5779.
V. Gupta (2019), Near-optimal Bayesian ambiguity sets for distributionally robust optimization, *Manag. Sci.* **65**, 4242–4260.
M. Gürbüzbalaban, A. Ruszczyński and L. Zhu (2022), A stochastic subgradient method for distributionally robust non-convex and non-smooth learning, *J. Optim. Theory Appl.* **194**, 1014–1041.
J. Hajar, T. Kargin and B. Hassibi (2023), Wasserstein distributionally robust regret-optimal control under partial observability, in *59th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1–6.
A. Hakobyan and I. Yang (2024), Wasserstein distributionally robust control of partially observable linear stochastic systems, *IEEE Trans. Automat. Control* **69**, 6121–6136.

H. Hamburger (1920), Über eine Erweiterung des Stieltjesschen Momentenproblems, *Math. Ann.* **81**, 235–319.

F. R. Hampel (1968), Contributions to the theory of robust estimation. Technical report, University of California, Berkeley.

F. R. Hampel (1971), A general qualitative definition of robustness, *Ann. Math. Statist.* **42**, 1887–1896.

B. Han, C. Shang and D. Huang (2021), Multiple kernel learning-aided robust optimization: Learning algorithm, computational tractability, and usage in multi-stage decision-making, *European J. Oper. Res.* **292**, 1004–1018.

S. Han, M. Tao, U. Topcu, H. Owhadi and R. M. Murray (2015), Convex optimal uncertainty quantification, *SIAM J. Optim.* **25**, 1368–1387.

G. A. Hanasusanto and D. Kuhn (2013), Robust data-driven dynamic programming, in *Advances in Neural Information Processing Systems 26* (C. J. Burges *et al.*, eds), Curran Associates, pp. 827–835.

G. A. Hanasusanto and D. Kuhn (2018), Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls, *Oper. Res.* **66**, 849–869.

G. A. Hanasusanto, D. Kuhn and W. Wiesemann (2016), A comment on 'Computational complexity of stochastic programming problems', *Math. Program.* **159**, 557–569.

G. A. Hanasusanto, D. Kuhn, S. W. Wallace and S. Zymler (2015*a*), Distributionally robust multi-item newsvendor problems with multimodal demand distributions, *Math. Program.* **152**, 1–32.

G. A. Hanasusanto, V. Roitch, D. Kuhn and W. Wiesemann (2015*b*), A distributionally robust perspective on uncertainty quantification and chance constrained programming, *Math. Program.* **151**, 35–62.

L. P. Hansen and T. J. Sargent (2008), *Robustness*, Princeton University Press.

L. P. Hansen and T. J. Sargent (2010), Wanting robustness in macroeconomics, in *Handbook of Monetary Economics 3* (B. M. Friedman and M. Woodford, eds), Elsevier, pp. 1097–1157.

C. A. Hartley and L. H. Somerville (2015), The neuroscience of adolescent decision-making, *Current Opinion Behav. Sci.* **5**, 108–115.

J. Hartung (1982), An extension of Sion's minimax theorem with an application to a method for constrained games, *Pacific J. Math.* **103**, 401–408.

T. Hastie, R. Tibshirani and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

F. Hausdorff (1923), Momentprobleme für ein endliches Intervall, *Math. Zeitschr.* **16**, 220–248.

B. Hayden, S. Heilbronner and M. Platt (2010), Ambiguity aversion in rhesus macaques, *Front. Neurosci.* Available at doi:10.3389/fnins.2010.00166/full.

E. Hazan (2022), *Introduction to Online Convex Optimization*, MIT Press.

Q. He, G. Xue, C. Chen, Z. Lu, Q. Dong, X. Lei, N. Ding, J. Li, H. Li, C. Chen, J. Li, R. K. Moyzis and A. Bechara (2010), Serotonin transporter gene-linked polymorphic region (5-HTTLPR) influences decision making under ambiguity and risk in a large Chinese sample, *Neuropharmacol.* **59**, 518–526.

S. He and H. Lam (2021), Higher-order expansion and Bartlett correctability of distributionally robust optimization. Available at arXiv:2108.05908.

J. P. Hespanha (2019), *Linear Systems Theory*, Princeton University Press.

N. Ho-Nguyen and F. Kılınç-Karzan (2018), Online first-order framework for robust convex optimization, *Oper. Res.* **66**, 1670–1692.
N. Ho-Nguyen and F. Kılınç-Karzan (2019), Exploiting problem structure in optimization under uncertainty via online convex optimization, *Math. Program.* **177**, 113–147.
N. Ho-Nguyen and S. J. Wright (2023), Adversarial classification via distributional robustness with Wasserstein ambiguity, *Math. Program.* **198**, 1411–1447.
N. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz and D. Lee (2022), Distributionally robust chance-constrained programs with right-hand side uncertainty under Wasserstein ambiguity, *Math. Program.* **196**, 641–672.
P. Honeyman, R. E. Ladner and M. Yannakakis (1980), Testing the universal instance assumption, *Inform. Process. Lett.* **10**, 14–19.
L. J. Hong, Z. Huang and H. Lam (2020), Learning-based robust optimization: Procedures and statistical guarantees, *Manag. Sci.* **67**, 3447–3467.
R. A. Horn and C. R. Johnson (1985), H∞-optimal control and related minimax design problems, *IEEE Trans. Automat. Control* **30**, 1057–1069.
S. Hou, P. Kassraie, A. Kratsios, A. Krause and J. Rothfuss (2023), Instance-dependent generalization bounds via optimal transport, *J. Mach. Learn. Res.* **24**, 16815–16865.
M. Hsu, M. Bhatt, R. Adolphs, D. Tranel and C. F. Camerer (2005), Neural systems responding to degrees of uncertainty in human decision-making, *Science* **310**, 1680–1683.
Y. Hu, X. Chen and N. He (2021), On the bias–variance–cost tradeoff of stochastic optimization, in *Advances in Neural Information Processing Systems 34* (M. Ranzato *et al.*, eds), Curran Associates, pp. 22119–22131.
Y. Hu, J. Wang, X. Chen and N. He (2024), Multi-level Monte-Carlo gradient methods for stochastic optimization with biased oracles. Available at arXiv:2408.11084.
Z. Hu and L. J. Hong (2013), Kullback–Leibler divergence constrained distributionally robust optimization. Available at optimization-online.org:2012/11/3677.pdf.
Z. Hu, L. J. Hong and A. M.-C. So (2013), Ambiguous probabilistic programs. Available at optimization-online.org:2013/09/4039.pdf.
K. Huang, H. Yang, I. King, M. R. Lyu and L. Chan (2004), The minimum error minimax probability machine, *J. Mach. Learn. Res.* **5**, 1253–1286.
P. Huber (1981), *Robust Statistics*, Wiley.
P. J. Huber (1964), Robust estimation of a location parameter, *Ann. Math. Statist.* **35**, 73–101.
P. J. Huber (1967), The behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 221–233.
P. J. Huber (1968), Robust confidence limits, *Z. Wahrscheinlichkeitsth. verwandte Gebiete* **10**, 269–278.
H. Husain (2020), Distributional robustness with IPMs and links to regularization and GANs, in *Advances in Neural Information Processing Systems 33* (H. Larochelle *et al.*, eds), Curran Associates, pp. 11816–11827.
K. Isii (1960), The extrema of probability determined by generalized moments I: Bounded random variables, *Ann. Inst. Statist. Math.* **12**, 119–134.
K. Isii (1962), On sharpness of Tchebycheff-type inequalities, *Ann. Inst. Statist. Math.* **14**, 185–197.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

G. Iyengar, H. Lam and T. Wang (2023), Hedging against complexity: Distributionally robust optimization with parametric approximation, in *26th International Conference on Artificial Intelligence and Statistics*, Vol. 206 of Proceedings of Machine Learning Research, PMLR, pp. 9976–10011.

R. Jagannathan (1977), Minimax procedure for a class of linear programs under uncertainty, *Oper. Res.* **25**, 173–177.

D. Jakubovitz and R. Giryes (2018), Improving DNN robustness to adversarial attacks using Jacobian regularization, in *European Conference on Computer Vision*, pp. 514–529.

S. L. Janak, X. Lin and C. A. Floudas (2007), A new robust optimization approach for scheduling under uncertainty II: Uncertainty with known probability distribution, *Comput. Chem. Engrg* **31**, 171–195.

H. Jeffreys and D. Wrinch (1921), On certain fundamental principles of scientific enquiry, *Philos. Mag.* **42**, 269–298.

J. L. W. V. Jensen (1906), Sur les fonctions convexes et les inégalités entre les valeurs moyennes, *Acta Math.* **30**, 175–193.

N. Jiang and W. Xie (2024), Distributionally favorable optimization: A framework for data-driven decision-making with endogenous outliers, *SIAM J. Optim.* **34**, 419–458.

R. Jiang and Y. Guan (2016), Data-driven chance constrained stochastic program, *Math. Program.* **158**, 291–327.

R. Jiang and Y. Guan (2018), Risk-averse two-stage stochastic program with distributional ambiguity, *Oper. Res.* **66**, 1390–1405.

Y. Jiang and J. Obloj (2024), Sensitivity of causal distributionally robust optimization. Available at arXiv:2408.17109.

Y. Jiang, S. Chewi and A.-A. Pooladian (2024), Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space, in *37th Conference on Learning Theory*, Vol. 247 of Proceedings of Machine Learning Research, PMLR, pp. 2720–2721.

W. Jongeneel, T. Sutter and D. Kuhn (2021), Topological linear system identification via moderate deviations theory, *IEEE Control Systems Letters* **6**, 307–312.

W. Jongeneel, T. Sutter and D. Kuhn (2022), Efficient learning of a linear dynamical system with stability guarantees, *IEEE Trans. Automat. Control* **68**, 2790–2804.

H. Jylhä (2015), The L^∞ optimal transport: Infinite cyclical monotonicity and the existence of optimal transport maps, *Calc. Var. Partial Differential Equations* **52**, 303–326.

O. Kallenberg (1997), *Foundations of Modern Probability*, Springer.

T. Kargin, J. Hajar, V. Malik and B. Hassibi (2024*a*), The distributionally robust infinite-horizon LQR. Available at arXiv:2408.06230.

T. Kargin, J. Hajar, V. Malik and B. Hassibi (2024*b*), Distributionally robust Kalman filtering over finite and infinite horizon. Available at arXiv:2407.18837.

T. Kargin, J. Hajar, V. Malik and B. Hassibi (2024*c*), Infinite-horizon distributionally robust regret-optimal control, in *41st International Conference on Machine Learning*, pp. 23187–23214.

T. Kargin, J. Hajar, V. Malik and B. Hassibi (2024*d*), Wasserstein distributionally robust regret-optimal control over infinite-horizon, in *6th Annual Learning for Dynamics & Control Conference*, Vol. 242 of Proceedings of Machine Learning Research, PMLR, pp. 1688–1701.

S. Karlin and W. J. Studden (1966), *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience Publishers.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

N. Karmarkar (1984), A new polynomial-time algorithm for linear programming, *Combinatorica* **4**, 373–395.
J. E. Kelley Jr (1960), The cutting-plane method for solving convex programs, *J. Soc. Indust. Appl. Math.* **8**, 703–712.
C. Kent, J. Li, J. Blanchet and P. W. Glynn (2021), Modified Frank Wolfe in probability space, in *Advances in Neural Information Processing Systems 34* (M. Ranzato *et al.*, eds), Curran Associates, pp. 14448–14462.
J. M. Keynes (1921), *A Treatise on Probability*, Macmillan.
L. G. Khachiyan (1979), A polynomial algorithm in linear programming, *Dokl. Akad. Nauk* **244**, 1093–1096.
H. K. Khalil (1996), *Control System Analysis and Design with Advanced Design Tools*, Prentice Hall.
A. J. King and R. T. Rockafellar (1993), Asymptotic theory for solutions in statistical estimation and stochastic programming, *Math. Oper. Res.* **18**, 148–162.
A. J. King and R. J.-B. Wets (1991), Epi-consistency of convex stochastic programs, *Stoch. Stoch. Reports* **34**, 83–92.
D. Klabjan, D. Simchi-Levi and M. Song (2013), Robust stochastic lot-sizing by means of histograms, *Prod. Oper. Manag.* **22**, 691–710.
F. H. Knight (1921), *Risk, Uncertainty and Profit*, Houghton Mifflin.
Ç. Koçyiğit, G. Iyengar, D. Kuhn and W. Wiesemann (2020), Distributionally robust mechanism design, *Manag. Sci.* **66**, 159–189.
Ç. Koçyiğit, N. Rujeerapaiboon and D. Kuhn (2022), Robust multidimensional pricing: Separation without regret, *Math. Program.* **196**, 841–874.
V. Koltchinskii (2011), *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, Springer.
P. Kouvelis and G. Yu (1997), *Robust Discrete Optimization and its Applications*, Springer.
A. L. Krain, A. M. Wilson, R. Arbuckle, X. F. Castellanos and M. P. Milham (2006), Distinct neural mechanisms of risk and ambiguity: A meta-analysis of decision-making, *NeuroImage* **32**, 477–484.
S. G. Krantz and H. R. Parks (2002), *A Primer of Real Analytic Functions*, Springer.
D. Kuhn (2005), *Generalized Bounds for Convex Multistage Stochastic Programs*, Springer.
D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen and S. Shafieezadeh-Abadeh (2019), Wasserstein distributionally robust optimization: Theory and applications in machine learning, *INFORMS TutORials in Operations Research*, pp. 130–166. Available at doi:10.1287/educ.2019.0198.
S. Kullback (1959), *Information Theory and Statistics*, Wiley.
M. Kupper and W. Schachermayer (2009), Representation results for law invariant time consistent functions, *Math. Financ. Econom.* **2**, 189–210.
A. Kurakin, I. J. Goodfellow and S. Bengio (2017), Adversarial machine learning at scale, in *International Conference on Learning Representations (ICLR 2017)*.
S. Kusuoka (2001), On law invariant coherent risk measures, in *Advances in Mathematical Economics* (S. Kusuoka and T. Maruyama, eds), Springer, pp. 83–95.
Y. Kwon, W. Kim, J.-H. Won and M. C. Paik (2020), Principled learning method for Wasserstein distributionally robust optimization with local perturbations, in *37th International Conference on Machine Learning*, Vol. 119 of Proceedings of Machine Learning Research, PMLR, pp. 5567–5576.

D. N. Lal (1955), A note on a form of Tchebycheff's inequality for two or more variables, *Sankhyā* **15**, 317–320.
H. Lam (2016), Robust sensitivity analysis for stochastic systems, *Math. Oper. Res.* **41**, 1248–1275.
H. Lam (2018), Sensitivity to serial dependency of input processes: A robust approach, *Manag. Sci.* **64**, 1311–1327.
H. Lam (2019), Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization, *Oper. Res.* **67**, 1090–1105.
H. Lam (2021), On the impossibility of statistically improving empirical optimization: A second-order stochastic dominance perspective. Available at arXiv:2105.13419.
H. Lam and C. Mottet (2017), Tail analysis without parametric models: A worst-case perspective, *Oper. Res.* **65**, 1696–1711.
H. Lam and E. Zhou (2017), The empirical likelihood approach to quantifying uncertainty in sample average approximation, *Oper. Res. Lett.* **45**, 301–307.
H. Lam, Z. Liu and D. I. Singham (2024), Shape-constrained distributional optimization via importance-weighted sample average approximation. Available at arXiv:2406.07825.
H. Lam, Z. Liu and X. Zhang (2021), Orthounimodal distributionally robust optimization: Representation, computation and multivariate extreme event applications. Available at arXiv:2111.07894.
M. Lambert, S. Chewi, F. Bach, S. Bonnabel and P. Rigollet (2022), Variational inference via Wasserstein gradient flows, in *Advances in Neural Information Processing Systems 35* (S. Koyejo *et al.*, eds), Curran Associates, pp. 14434–14447.
G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya and M. I. Jordan (2001), Minimax probability machine, in *Advances in Neural Information Processing Systems 14* (T. Dietterich *et al.*, eds), MIT Press, pp. 801–807.
G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya and M. I. Jordan (2002), A robust minimax approach to classification, *J. Mach. Learn. Res.* **3**, 555–582.
N. Lanzetti, S. Bolognani and F. Dörfler (2022), First-order conditions for optimization in the Wasserstein space. Available at arXiv:2209.12197.
N. Lanzetti, A. Terpin and F. Dörfler (2024), Variational analysis in the Wasserstein space. Available at arXiv:2406.10676.
J. B. Lasserre (2001), Global optimization with polynomials and the problem of moments, *SIAM J. Optim.* **11**, 796–817.
J. B. Lasserre (2002), Bounds on measures satisfying moment conditions, *Ann. Appl. Probab.* **12**, 1114–1137.
J. B. Lasserre (2008), A semidefinite programming approach to the generalized problem of moments, *Math. Program.* **112**, 65–92.
J. B. Lasserre (2009), *Moments, Positive Polynomials and Their Applications*, World Scientific.
J. B. Lasserre and T. Weisser (2021), Distributionally robust polynomial chance-constraints under mixture ambiguity sets, *Math. Program.* **185**, 409–453.
T. T.-K. Lau and H. Liu (2022), Wasserstein distributionally robust optimization with Wasserstein barycenters. Available at arXiv:2203.12136.
J. Lee and M. Raginsky (2018), Minimax statistical learning with Wasserstein distances, in *Advances in Neural Information Processing Systems 31* (S. Bengio *et al.*, eds), Curran Associates, pp. 2687–2696.

J. Lee, S. Park and J. Shin (2020), Learning bounds for risk-sensitive learning, in *Advances in Neural Information Processing Systems 33* (H. Larochelle *et al.*, eds), Curran Associates, pp. 13867–13879.
E. L. Lehmann and G. Casella (2006), *Theory of Point Estimation*, Springer.
E. S. Levitin and B. T. Polyak (1966), Constrained minimization methods, *USSR Comput. Math. Math. Phys.* **6**, 1–50.
B. C. Levy (2008), Robust hypothesis testing with a relative entropy tolerance, *IEEE Trans. Inform. Theory* **55**, 413–421.
B. C. Levy and R. Nikoukhah (2004), Robust least-squares estimation with a relative entropy constraint, *IEEE Trans. Inform. Theory* **50**, 89–104.
B. C. Levy and R. Nikoukhah (2012), Robust state space filtering under incremental model perturbations subject to a relative entropy tolerance, *IEEE Trans. Automat. Control* **58**, 682–695.
D. Levy, Y. Carmon, J. C. Duchi and A. Sidford (2020), Large-scale methods for distributionally robust optimization, in *Advances in Neural Information Processing Systems 33* (H. Larochelle *et al.*, eds), Curran Associates, pp. 8847–8860.
B. Li, R. Jiang and J. L. Mathieu (2016), Distributionally robust risk-constrained optimal power flow using moment and unimodality information, in *55th IEEE Conference on Decision and Control (CDC)*, pp. 2425–2430.
B. Li, R. Jiang and J. L. Mathieu (2019*a*), Ambiguous risk constraints with moment and unimodality information, *Math. Program.* **173**, 151–192.
C. Li, U. Turmunkh and P. P. Wakker (2019*b*), Trust as a decision under ambiguity, *Exp. Econom.* **22**, 51–75.
D. Li and S. Martínez (2020), Data assimilation and online optimization with performance guarantees, *IEEE Trans. Automat. Control* **66**, 2115–2129.
J. Li, C. Chen and A. M.-C. So (2020), Fast epigraphical projection-based incremental algorithms for Wasserstein distributionally robust support vector machine, in *Advances in Neural Information Processing Systems 33* (H. Larochelle *et al.*, eds), Curran Associates, pp. 4029–4039.
J. Li, S. Huang and A. M.-C. So (2019*c*), A first-order algorithmic framework for Wasserstein distributionally robust logistic regression, in *Advances in Neural Information Processing Systems 32* (H. Wallach *et al.*, eds), Curran Associates, pp. 3937–3947.
J. Li, S. Lin, J. Blanchet and V. A. Nguyen (2022), Tikhonov regularization is optimal transport robust under martingale constraints, in *Advances in Neural Information Processing Systems 35* (S. Koyejo *et al.*, eds), Curran Associates, pp. 17677–17689.
J. Y.-M. Li (2018), Closed-form solutions for worst-case law invariant risk measures with application to robust portfolio optimization, *Oper. Res.* **66**, 1533–1541.
J. Y.-M. Li and T. Mao (2022), A general Wasserstein framework for data-driven distributionally robust optimization: Tractability and applications. Available at arXiv:2207.09403.
M. Li, T. Sutter and D. Kuhn (2021), Distributionally robust optimization with Markovian data, in *38th International Conference on Machine Learning*, Vol. 139 of Proceedings of Machine Learning Research, PMLR, pp. 6493–6503.
Z. Li, R. Ding and C. A. Floudas (2011), A comparative theoretical and computational study on robust counterpart optimization I: Robust linear optimization and robust mixed integer linear optimization, *Indust. Engrg Chem. Res.* **50**, 10567–10603.
F. Liese and I. Vajda (1987), *Convex Statistical Distances*, Teubner.

S. Lin, J. Blanchet, P. Glynn and V. A. Nguyen (2024), Small sample behavior of Wasserstein projections, connections to empirical likelihood, and other applications. Available at arXiv:2408.11753.

F. Liu, Z. Chen, R. Wang and S. Wang (2024*a*), Newsvendor under mean–variance ambiguity and misspecification. Available at arXiv:2405.07008.

J. Liu, Z. Su and H. Xu (2024*b*), Bayesian distributionally robust Nash equilibrium and its application. Available at arXiv:2410.20364.

Z. Liu and P.-L. Loh (2023), Robust W-GAN-based estimation under Wasserstein contamination, *Inform. Inference* **12**, 312–362.

Z. Liu, B. P. G. Van Parys and H. Lam (2023), Smoothed f-divergence distributionally robust optimization: Exponential rate efficiency and complexity-free calibration. Available at arXiv:2306.14041.

D. Z. Long, J. Qi and A. Zhang (2024), Supermodularity in two-stage distributionally robust optimization, *Manag. Sci.* **70**, 1394–1409.

C. Lyu, K. Huang and H.-N. Liang (2015), A unified gradient regularization family for adversarial examples, in *IEEE International Conference on Data Mining*, pp. 301–309.

A. Madansky (1959), Bounds on the expectation of a convex function of a multivariate random variable, *Ann. Math. Statist.* **30**, 743–746.

A. Mądry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu (2018), Towards deep learning models resistant to adversarial attacks, in *International Conference on Learning Representations (ICLR 2018)*.

C. Maheshwari, C.-Y. Chiu, E. Mazumdar, S. Sastry and L. Ratliff (2022), Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization, in *25th International Conference on Artificial Intelligence and Statistics*, Vol. 151 of Proceedings of Machine Learning Research, PMLR, pp. 6702–6734.

H.-Y. Mak, Y. Rong and J. Zhang (2015), Appointment scheduling with limited distributional information, *Manag. Sci.* **61**, 316–334.

A. Markov (1884), On certain applications of algebraic continued fractions (in Russian). PhD thesis, St Petersburg.

A. W. Marshall and I. Olkin (1960), A one-sided inequality of the Chebyshev type, *Ann. Math. Statist.* **31**, 488–491.

K. Marton (1986), A simple proof of the blowing-up lemma, *IEEE Trans. Inform. Theory* **32**, 445–446.

A. Maurer and M. Pontil (2009), Empirical Bernstein bounds and sample variance penalization, in *22nd Conference on Learning Theory (COLT 2009)*. Available at https://www.cs.mcgill.ca/~colt2009/papers/012.pdf#page=1.

R. D. McAllister and P. Mohajerin Esfahani (2024), Distributionally robust model predictive control: Closed-loop guarantees and scalable algorithms, *IEEE Trans. Automat. Control*. Available at doi:10.1109/TAC.2024.3498702.

A. McNeil, R. Frey and P. Embrechts (2015), *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press.

S. Mendelson (2003), A few notes on statistical learning theory, in *Advanced Lectures on Machine Learning* (S. Mendelson and A. J. Smola, eds), Springer, pp. 1–40.

R. O. Michaud (1989), The Markowitz optimization enigma: Is 'optimized' optimal?, *Financ. Anal. J.* **45**, 31–42.

J. Milz and M. Ulbrich (2020), An approximation scheme for distributionally robust nonlinear optimization, *SIAM J. Optim.* **30**, 1996–2025.

J. Milz and M. Ulbrich (2022), An approximation scheme for distributionally robust PDE-constrained optimization, *SIAM J. Control Optim.* **60**, 1410–1435.
V. K. Mishra, K. Natarajan, D. Padmanabhan, C.-P. Teo and X. Li (2014), On theoretical and empirical aspects of marginal distribution choice models, *Manag. Sci.* **60**, 1511–1531.
V. K. Mishra, K. Natarajan, H. Tao and C.-P. Teo (2012), Choice prediction with semi-definite optimization when utilities are correlated, *IEEE Trans. Automat. Control* **57**, 2450–2463.
P. Mohajerin Esfahani and D. Kuhn (2018), Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations, *Math. Program.* **171**, 115–166.
P. Mohajerin Esfahani, S. Shafieezadeh-Abadeh, G. A. Hanasusanto and D. Kuhn (2018), Data-driven inverse optimization with imperfect information, *Math. Program.* **167**, 191–234.
P. Mohajerin Esfahani, T. Sutter and J. Lygeros (2015), Performance bounds for the scenario approach and an extension to a class of non-convex programs, *IEEE Trans. Automat. Control* **60**, 46–58.
J. R. Munkres (2000), *Topology*, Prentice Hall.
A. Mutapcic and S. Boyd (2009), Cutting-set methods for robust convex optimization with pessimizing oracles, *Optim. Methods Softw.* **24**, 381–406.
V. Nagarajan and J. Z. Kolter (2017), Gradient descent GAN optimization is locally stable, in *Advances in Neural Information Processing Systems 30* (I. Guyon *et al.*, eds), Curran Associates, pp. 5591–5600.
H. Nakao, R. Jiang and S. Shen (2021), Distributionally robust partially observable Markov decision process with moment-based ambiguity, *SIAM J. Optim.* **31**, 461–488.
H. Namkoong and J. C. Duchi (2016), Stochastic gradient methods for distributionally robust optimization with f-divergences, in *Advances in Neural Information Processing Systems 29* (D. Lee *et al.*, eds), Curran Associates, pp. 2216–2224.
K. Natarajan (2021), *Optimization with Marginals and Moments*, Dynamic Ideas.
K. Natarajan and Z. Linyi (2007), A mean–variance bound for a three-piece linear function, *Probab. Engrg Inform. Sci.* **21**, 611–621.
K. Natarajan, D. Pachamanova and M. Sim (2009*a*), Constructing risk measures from uncertainty sets, *Oper. Res.* **57**, 1129–1141.
K. Natarajan, D. Padmanabhan and A. Ramachandra (2023), Distributionally robust optimization through the lens of submodularity. Available at arXiv:2312.04890.
K. Natarajan, M. Sim and J. Uichanco (2010), Tractable robust expected utility and risk models for portfolio optimization, *Math. Finance* **20**, 695–731.
K. Natarajan, M. Sim and J. Uichanco (2018), Asymmetry and ambiguity in newsvendor models, *Manag. Sci.* **64**, 3146–3167.
K. Natarajan, M. Song and C.-P. Teo (2009*b*), Persistency model and its applications in choice modeling, *Manag. Sci.* **55**, 453–469.
K. Natarajan, C. P. Teo and Z. Zheng (2011), Mixed 0-1 linear programs under objective uncertainty: A completely positive representation, *Oper. Res.* **59**, 713–728.
A. Nemirovski and A. Shapiro (2007), Convex approximations of chance constrained programs, *SIAM J. Optim.* **17**, 969–996.
Y. Nesterov and A. Nemirovskii (1994), *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

D. Nguyen, N. Bui and V. A. Nguyen (2023*a*), Distributionally robust recourse action, in *International Conference on Learning Representations (ICLR 2023)*.

V. A. Nguyen, D. Kuhn and P. Mohajerin Esfahani (2022), Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator, *Oper. Res.* **70**, 490–515.

V. A. Nguyen, S. Shafiee, D. Filipović and D. Kuhn (2021), Mean–covariance robust risk measurement. Available at arXiv:2112.09959.

V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn and P. Mohajerin Esfahani (2023*b*), Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization, *Math. Oper. Res.* **48**, 1–37.

V. A. Nguyen, S. Shafieezadeh-Abadeh, M.-C. Yue, D. Kuhn and W. Wiesemann (2019), Optimistic distributionally robust optimization for nonparametric likelihood approximation, in *Advances in Neural Information Processing Systems 32* (H. Wallach *et al.*, eds), Curran Associates, pp. 15872–15882.

V. A. Nguyen, F. Zhang, J. Blanchet, E. Delage and Y. Ye (2020), Distributionally robust local non-parametric conditional estimation, in *Advances in Neural Information Processing Systems 33* (H. Larochelle *et al.*, eds), Curran Associates, pp. 15232–15242.

V. A. Nguyen, F. Zhang, S. Wang, J. Blanchet, E. Delage and Y. Ye (2024), Robustifying conditional portfolio decisions via optimal transport, *Oper. Res.* Available at doi:10.1287/opre.2021.0243.

S. Nietert, Z. Goldfeld and S. Shafiee (2024*a*), Outlier-robust Wasserstein DRO, in *Advances in Neural Information Processing Systems 36* (A. Oh *et al.*, eds), Curran Associates, pp. 62792–62820.

S. Nietert, Z. Goldfeld and S. Shafiee (2024*b*), Robust distribution learning with local and global adversarial corruptions, in *37th Conference on Learning Theory*, Vol. 247 of Proceedings of Machine Learning Research, PMLR, pp. 4007–4008.

K. G. Nishimura and H. Ozaki (2004), Search and Knightian uncertainty, *J. Econom. Theory* **119**, 299–333.

K. G. Nishimura and H. Ozaki (2006), An axiomatic approach to-contamination, *Econom. Theory* **27**, 333–340.

J. L. M. Olea, C. Rush, A. Velez and J. Wiesel (2022), The out-of-sample prediction error of the square-root-LASSO and related estimators. Available at arXiv:2211.07608.

I. Olkin and F. Pukelsheim (1982), The distance between two random vectors with given dispersion matrices, *Linear Algebra Appl.* **48**, 257–263.

C. Ordoudis, V. A. Nguyen, D. Kuhn and P. Pinson (2021), Energy and reserve dispatch with distributionally robust joint chance constraints, *Oper. Res. Lett.* **49**, 291–299.

A. B. Owen (1988), Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* **75**, 237–249.

A. B. Owen (1990), Empirical likelihood ratio confidence regions, *Ann. Statist.* **18**, 90–120.

A. B. Owen (1991), Empirical likelihood for linear models, *Ann. Statist.* **19**, 1725–1747.

A. B. Owen (2001), *Empirical Likelihood*, Chapman & Hall.

H. Owhadi and C. Scovel (2017), Extreme points of a ball about a measure with finite support, *Commun. Math. Sci.* **15**, 77–96.

H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns and M. Ortiz (2013), Optimal uncertainty quantification, *SIAM Rev.* **55**, 271–345.

V. M. Panaretos and Y. Zemel (2020), *An Invitation to Statistics in Wasserstein Space*, Springer.

P. A. Parrilo (2000), Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. PhD thesis, California Institute of Technology.
P. A. Parrilo (2003), Semidefinite programming relaxations for semialgebraic problems, *Math. Program.* **96**, 293–320.
B. Pass (2015), Multi-marginal optimal transport: Theory and applications, *ESAIM Math. Model. Numer. Anal.* **49**, 1771–1790.
S. Peng (1997), Backward SDE and related G-expectation, in *Backward Stochastic Differential Equations in Finance* (N. El Karoui, S. Peng and M. C. Quenez, eds), Wiley, pp. 141–160.
S. Peng (2007*a*), G-Brownian motion and dynamic risk measure under volatility uncertainty. Available at arXiv:0711.2834.
S. Peng (2007*b*), G-expectation, G-Brownian motion and related stochastic calculus of Itô type, in *Stochastic Analysis and Applications* (F. E. Benth *et al.*, eds), Springer, pp. 541–567.
S. Peng (2019), *Nonlinear Expectations and Stochastic Calculus under Uncertainty: With Robust CLT and G-Brownian Motion*, Springer.
S. Peng (2023), G-Gaussian processes under sublinear expectations and q-Brownian motion in quantum mechanics, *Numer. Algebra Control Optim.* **13**, 583–603.
G. Perakis and G. Roels (2008), Regret in the newsvendor model with partial information, *Oper. Res.* **56**, 188–203.
S. Pesenti, Q. Wang and R. Wang (2024), Optimizing distortion riskmetrics with distributional uncertainty, *Math. Program.* Available at doi:10.1007/s10107-024-02128-6.
G. C. Pflug and A. Pichler (2014), *Multistage Stochastic Optimization*, Springer.
G. C. Pflug and D. Wozabal (2007), Ambiguity in portfolio selection, *Quant. Finance* **7**, 435–442.
G. C. Pflug, A. Pichler and D. Wozabal (2012), The $1/N$ investment strategy is optimal under high model ambiguity, *J. Banking Finance* **36**, 410–417.
R. R. Phelps (1965), *Lectures on Choquet's Theorem*, Van Nostrand Mathematical Studies.
A. B. Philpott, V. L. de Matos and L. Kapelevich (2018), Distributionally robust SDDP, *Comput. Manag. Sci.* **15**, 431–454.
A. Pichler (2013), Evaluations of risk measures for different probability measures, *SIAM J. Optim.* **23**, 530–551.
I. Pinelis (2016), On the extreme points of moments sets, *Math. Methods Oper. Res.* **83**, 325–349.
I. Pólik and T. Terlaky (2007), A survey of the S-lemma, *SIAM Rev.* **49**, 371–418.
Y. Polyanskiy and Y. Wu (2024), *Information Theory: From Coding to Learning*, Cambridge University Press.
I. Popescu (2005), A semidefinite programming approach to optimal-moment bounds for convex classes of distributions, *Math. Oper. Res.* **30**, 632–657.
I. Popescu (2007), Robust mean–covariance solutions for stochastic optimization, *Oper. Res.* **55**, 98–112.
K. Postek and S. Shtern (2024), First-order algorithms for robust optimization problems via convex-concave saddle-point Lagrangian reformulation, *INFORMS J. Comput.* Available at doi:10.1287/ijoc.2022.0200.
K. Postek, A. Ben-Tal, D. den Hertog and B. Melenberg (2018), Robust optimization with ambiguous stochastic constraints under mean and dispersion information, *Oper. Res.* **66**, 814–833.

K. Postek, D. den Hertog and B. Melenberg (2016), Computationally tractable counterparts of distributionally robust constraints on risk measures, *SIAM Rev.* **58**, 603–650.

K. Postek, W. Romeijnders, D. den Hertog and M. H. van der Vlerk (2019), An approximation framework for two-stage ambiguous stochastic integer programs under mean–MAD information, *European J. Oper. Res.* **274**, 432–444.

G. Puccetti and L. Rüschendorf (2013), Sharp bounds for sums of dependent risks, *J. Appl. Probab.* **50**, 42–53.

M. S. Pydi and V. Jog (2021), Adversarial risk via optimal transport and optimal couplings, *IEEE Trans. Inform. Theory* **67**, 6031–6052.

M. S. Pydi and V. Jog (2024), The many faces of adversarial risk: An expanded study, *IEEE Trans. Inform. Theory* **70**, 550–570.

H. Rahimian and S. Mehrotra (2022), Frameworks and results in distributionally robust optimization, *Open J. Math. Optim.* **3**, 1–85.

H. Rahimian, G. Bayraksan and T. Homem-de-Mello (2019*a*), Controlling risk and demand ambiguity in newsvendor models, *European J. Oper. Res.* **279**, 854–868.

H. Rahimian, G. Bayraksan and T. Homem-de-Mello (2019*b*), Identifying effective scenarios in distributionally robust stochastic programs with total variation distance, *Math. Program.* **173**, 393–430.

H. Rahimian, G. Bayraksan and T. Homem-de-Mello (2022), Effective scenarios in multistage distributionally robust optimization with a focus on total variation distance, *SIAM J. Optim.* **32**, 1698–1727.

M. D. Reid and R. C. Williamson (2011), Information, divergence and risk for binary experiments, *J. Mach. Learn. Res.* **12**, 731–817.

H. Richter (1957), Parameterfreie Abschätzung und Realisierung von Erwartungswerten, *Blätter der DGVFM* **3**, 147–162.

R. T. Rockafellar (1970), *Convex Analysis*, Princeton University Press.

R. T. Rockafellar (1974), *Conjugate Duality and Optimization*, SIAM.

R. T. Rockafellar and J. O. Royset (2013), Superquantiles and their applications to risk, random variables, and regression, *INFORMS TutORials in Operations Research*, pp. 151–167. Available at doi:10.1287/educ.2013.0111.

R. T. Rockafellar and J. O. Royset (2014), Random variables, monotone relations, and convex analysis, *Math. Program.* **148**, 297–331.

R. T. Rockafellar and J. O. Royset (2015), Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity, *SIAM J. Optim.* **25**, 1179–1208.

R. T. Rockafellar and S. Uryasev (2000), Optimization of conditional value-at-risk, *J. Risk* **2**, 21–41.

R. T. Rockafellar and S. Uryasev (2002), Conditional value-at-risk for general loss distributions, *J. Banking Finance* **26**, 1443–1471.

R. T. Rockafellar and S. Uryasev (2013), The fundamental risk quadrangle in risk management, optimization and statistical estimation, *Surv. Oper. Res. Manag. Sci.* **18**, 33–53.

R. T. Rockafellar and R. J.-B. Wets (2009), *Variational Analysis*, Springer.

R. T. Rockafellar, S. Uryasev and M. Zabarankin (2006), Generalized deviations in risk analysis, *Finance Stoch.* **10**, 51–74.

R. T. Rockafellar, S. Uryasev and M. Zabarankin (2008), Risk tuning with generalized linear regression, *Math. Oper. Res.* **33**, 712–729.

W. W. Rogosinski (1958), Moments of non-negative mass, *Proc. Royal Soc. London Ser. A.* **245**, 1–27.
N. Rontsis, M. A. Osborne and P. J. Goulart (2020), Distributionally ambiguous optimization for batch Bayesian optimization, *J. Mach. Learn. Res.* **21**, 1–26.
K. Roth, A. Lucchi, S. Nowozin and T. Hofmann (2017), Stabilizing training of generative adversarial networks through regularization, in *Advances in Neural Information Processing Systems 30* (I. Guyon *et al.*, eds), Curran Associates, pp. 2018–2028.
J. O. Royset (2022), Risk-adaptive approaches to learning and decision making: A survey. Available at arXiv:2212.00856.
Y. Ruan, X. Li, K. Murthy and K. Natarajan (2023), A nonparametric approach with marginals for modeling consumer choice, in *24th ACM Conference on Economics and Computation*, ACM, p. 1078.
N. Rujeerapaiboon, D. Kuhn and W. Wiesemann (2016), Robust growth-optimal portfolios, *Manag. Sci.* **62**, 2090–2109.
N. Rujeerapaiboon, D. Kuhn and W. Wiesemann (2018), Chebyshev inequalities for products of random variables, *Math. Oper. Res.* **43**, 887–918.
L. Rüschendorf (1983), Solution of a statistical optimization problem by rearrangement methods, *Metrika* **30**, 55–61.
L. Rüschendorf (1991), Fréchet-bounds and their applications, in *Advances in Probability Distributions with Given Marginals: Beyond the Copulas* (G. Dall'Aglio, S. Kotz and G. Salinetti, eds), Springer, pp. 151–187.
L. Rüschendorf (2013), *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios*, Springer.
B. Rustem and M. Howe (2009), *Algorithms for Worst-Case Design and Applications to Risk Management*, Princeton University Press.
A. Ruszczyński (2021), A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization, *SIAM J. Control Optim.* **59**, 2301–2320.
A. Ruszczyński and A. Shapiro (2006), Optimization of convex risk functions, *Math. Oper. Res.* **31**, 433–452.
Y. Rychener, A. Esteban-Pérez, J. M. Morales and D. Kuhn (2024), Wasserstein distributionally robust optimization with heterogeneous data sources. Available at arXiv:2407.13582.
U. Sadana, E. Delage and A. Georghiou (2024), Data-driven decision-making under uncertainty with entropic risk measure. Available at arXiv:2409.19926.
S. Sagawa, P. W. Koh, T. B. Hashimoto and P. Liang (2020), Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, in *International Conference on Learning Representations (ICLR 2020)*.
A. A. Salo and M. Weber (1995), Ambiguity aversion in first-price sealed-bid auctions, *J. Risk Uncertain.* **11**, 123–137.
N. Sauldubois and N. Touzi (2024), First order martingale model risk and semi-static hedging. Available at arXiv:2410.06906.
S. L. Savage (2012), *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*, Wiley.
S. L. Savage, S. Scholtes and D. Zweidler (2006), Probability management, *OR/MS Today*. Available at https://www.wiley.com/en-us/The+Flaw+of+Averages%3A+Why+We+Underestimate+Risk+in+the+Face+of+Uncertainty-p-9781118073759.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

H. E. Scarf (1958), A min-max solution to an inventory problem, in *Studies in Mathematical Theory of Inventory and Production* (K. J. Arrow, S. Karlin and H. E. Scarf, eds), Stanford University Press, pp. 201–209.
G. Schildbach, L. Fagiano and M. Morari (2013), Randomized solutions to convex programs with multiple chance constraints, *SIAM J. Optim.* **23**, 2479–2501.
A. Selvi, M. R. Belbasi, M. Haugh and W. Wiesemann (2022), Wasserstein logistic regression with mixed features, in *Advances in Neural Information Processing Systems 35* (S. Koyejo *et al.*, eds), Curran Associates, pp. 16691–16704.
S. Shafiee and D. Kuhn (2024), Minimax theorems and Nash equilibria in distributionally robust optimization problems. Working paper.
S. Shafiee, L. Aolaritei, F. Dörfler and D. Kuhn (2023), New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. Available at arXiv:2303.03900.
S. Shafieezadeh-Abadeh, D. Kuhn and P. Mohajerin Esfahani (2019), Regularization via mass transportation, *J. Mach. Learn. Res.* **20**, 1–68.
S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani and D. Kuhn (2015), Distributionally robust logistic regression, in *Advances in Neural Information Processing Systems 28* (C. Cortes *et al.*, eds), Curran Associates, pp. 1576–1584.
S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn and P. Mohajerin Esfahani (2018), Wasserstein distributionally robust Kalman filtering, in *Advances in Neural Information Processing Systems 31* (S. Bengio *et al.*, eds), Curran Associates, pp. 8474–8483.
S. Shalev-Shwartz (2012), Online learning and online convex optimization, *Found. Trends Mach. Learn.* **4**, 107–194.
S. Shalev-Shwartz and S. Ben-David (2014), *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press.
A. Shapiro (1989), Asymptotic properties of statistical estimators in stochastic programming, *Ann. Statist.* **17**, 841–858.
A. Shapiro (1990), On differential stability in stochastic programming, *Math. Program.* **47**, 107–116.
A. Shapiro (1991), Asymptotic analysis of stochastic programs, *Ann. Oper. Res.* **30**, 169–186.
A. Shapiro (1993), Asymptotic behavior of optimal solutions in stochastic programming, *Math. Oper. Res.* **18**, 829–845.
A. Shapiro (2001), On duality theory of conic linear problems, in *Semi-Infinite Programming* (M. Á. Goberna and M. A. López, eds), Kluwer Academic, pp. 135–165.
A. Shapiro (2003), Monte Carlo sampling methods, in *Stochastic Programming* (A. Ruszczyński and A. Shapiro, eds), Elsevier, pp. 353–425.
A. Shapiro (2013), On Kusuoka representation of law invariant risk measures, *Math. Oper. Res.* **38**, 142–152.
A. Shapiro (2017), Distributionally robust stochastic programming, *SIAM J. Optim.* **27**, 2258–2275.
A. Shapiro and A. Kleywegt (2002), Minimax analysis of stochastic problems, *Optim. Methods Softw.* **17**, 523–542.
A. Shapiro, D. Dentcheva and A. Ruszczyński (2009), *Lectures on Stochastic Programming: Modeling and Theory*, SIAM.
A. Shapiro, E. Zhou and Y. Lin (2023), Bayesian distributionally robust optimization, *SIAM J. Optim.* **33**, 1279–1304.

K. S. Shehadeh (2023), Distributionally robust optimization approaches for a stochastic mobile facility fleet sizing, routing, and scheduling problem, *Transport. Sci.* **57**, 197–229.

K. S. Shehadeh, A. E. M. Cohn and R. Jiang (2020), A distributionally robust optimization approach for outpatient colonoscopy scheduling, *European J. Oper. Res.* **283**, 549–561.

H. Shen and R. Jiang (2023), Chance-constrained set covering with Wasserstein ambiguity, *Math. Program.* **198**, 621–674.

M. R. Sheriff and P. Mohajerin Esfahani (2024), Nonlinear distributionally robust optimization, *Math. Program.* Available at doi:10.1007/s10107-024-02151-7.

J. A. Shohat and J. D. Tamarkin (1950), *The Problem of Moments*, American Mathematical Society.

A. Sinha, H. Namkoong and J. Duchi (2018), Certifying some distributional robustness with principled adversarial training, in *International Conference on Learning Representations (ICLR 2018)*.

M. Sion (1958), On general minimax theorems, *Pacific J. Math.* **8**, 171–176.

J. E. Smith and R. L. Winkler (2006), The optimizer's curse: Skepticism and postdecision surprise in decision analysis, *Manag. Sci.* **52**, 311–322.

A. L. Soyster (1973), Convex programming with set-inclusive constraints and applications to inexact linear programming, *Oper. Res.* **21**, 1154–1157.

P. R. Srivastava, Y. Wang, G. A. Hanasusanto and C. P. Ho (2021), On data-driven prescriptive analytics with side information: A regularized Nadaraya–Watson approach. Available at arXiv:2110.04855.

M. Staib and S. Jegelka (2019), Distributionally robust optimization and generalization in kernel methods, in *Advances in Neural Information Processing Systems 32* (H. Wallach *et al.*, eds), Curran Associates, pp. 9134–9144.

T.-J. Stieltjes (1894), Recherches sur les fractions continues, *Ann. Fac. Sci. Toulouse Math. (6)* **8**, 1–122.

V. Strassen (1965), The existence of probability measures with given marginals, *Ann. Math. Statist.* **36**, 423–439.

T. Strohmann and G. Z. Grudic (2002), A formulation for minimax probability machine regression, in *Advances in Neural Information Processing Systems 15* (S. Becker *et al.*, eds), MIT Press, pp. 785–792.

K. R. Stromberg (2015), *An Introduction to Classical Real Analysis*, American Mathematical Society.

L. Sun, W. Xie and T. Witten (2023), Distributionally robust fair transit resource allocation during a pandemic, *Transport. Sci.* **57**, 954–978.

T. Sutter, A. Krause and D. Kuhn (2021), Robust generalization despite distribution shift via minimum discriminating information, in *Advances in Neural Information Processing Systems 34* (M. Ranzato *et al.*, eds), Curran Associates, pp. 29754–29767.

T. Sutter, B. P. G. Van Parys and D. Kuhn (2024), A Pareto dominance principle for data-driven optimization, *Oper. Res.* **72**, 1976–1999.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow and R. Fergus (2014), Intriguing properties of neural networks, in *International Conference on Learning Representations (ICLR 2014)*.

M. Talagrand (1996), Transportation cost for Gaussian and other product measures, *Geom. Funct. Anal.* **6**, 587–600.

B. Taşkesen, D. Iancu, Ç. Koçyiğit and D. Kuhn (2024), Distributionally robust linear quadratic control, in *Advances in Neural Information Processing Systems 36* (A. Oh *et al.*, eds), Curran Associates, pp. 18613–18632.

B. Taşkesen, S. Shafieezadeh-Abadeh and D. Kuhn (2023*a*), Semi-discrete optimal transport: Hardness, regularization and numerical solution, *Math. Program.* **199**, 1033–1106.

B. Taşkesen, S. Shafieezadeh-Abadeh, D. Kuhn and K. Natarajan (2023*b*), Discrete optimal transport with independent marginals is #P-hard, *SIAM J. Optim.* **33**, 589–614.

B. Taşkesen, M.-C. Yue, J. Blanchet, D. Kuhn and V. A. Nguyen (2021), Sequential domain adaptation by synthesizing distributionally robust experts, in *38th International Conference on Machine Learning*, Vol. 139 of Proceedings of Machine Learning Research, PMLR, pp. 10162–10172.

A. H. Tchen (1980), Inequalities for distributions with given marginals, *Ann. Probab.* **8**, 814–827.

A. Terpin, N. Lanzetti and F. Dörfler (2024), Dynamic programming in probability spaces via optimal transport, *SIAM J. Control Optim.* **62**, 1183–1206.

A. Terpin, N. Lanzetti, B. Yardim, F. Dörfler and G. Ramponi (2022), Trust region policy optimization with optimal transport discrepancies: Duality and algorithm for continuous actions, in *Advances in Neural Information Processing Systems 35* (S. Koyejo *et al.*, eds), Curran Associates, pp. 19786–19797.

Y. L. Tong (1980), *Probability Inequalities in Multivariate Distributions*, Academic Press.

F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh and P. McDaniel (2017), The space of transferable adversarial examples. Available at arXiv:1704.03453.

M. Y. Tsang and K. S. Shehadeh (2024), On the trade-off between distributional belief and ambiguity: Conservatism, finite-sample guarantees, and asymptotic properties. Available at arXiv:2410.19234.

K. Tu, Z. Chen and M.-C. Yue (2024), A max-min-max algorithm for large-scale robust optimization. Available at arXiv:2404.05377.

Z. Tu, J. Zhang and D. Tao (2019), Theoretical analysis of adversarial learning: A minimax approach, in *Advances in Neural Information Processing Systems 32* (H. Wallach *et al.*, eds), Curran Associates, pp. 12280–12290.

A. Van Der Vaart and J. A. Wellner (2000), Preservation theorems for Glivenko–Cantelli and uniform Glivenko–Cantelli classes, in *High Dimensional Probability II* (E. Giné, D. M. Mason and J. A. Wellner, eds), Springer, pp. 115–133.

A. W. Van der Vaart (1998), *Asymptotic Statistics*, Cambridge University Press.

W. J. E. C. van Eekelen, D. den Hertog and J. S. H. van Leeuwaarden (2022), MAD dispersion measure makes extremal queue analysis simple, *INFORMS J. Comput.* **34**, 1681–1692.

W. J. van Eekelen, G. A. Hanasusanto, J. J. Hasenbein and J. S. van Leeuwaarden (2025), Second-order bounds for the M/M/s queue with random arrival rate, *Queueing Syst.* **109**, art. 3.

J. S. H. Van Leeuwaarden and C. Stegehuis (2021), Robust subgraph counting with distribution-free random graph analysis, *Phys. Rev. E* **104**, art. 044313.

B. P. G. Van Parys (2024), Efficient data-driven optimization with noisy data, *Oper. Res. Lett.* **54**, art. 107089.

B. P. G. Van Parys and N. Golrezaei (2024), Optimal learning for structured bandits, *Manag. Sci.* **70**, 3951–3998.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press

B. P. G. Van Parys, P. J. Goulart and P. Embrechts (2016*a*), Fréchet inequalities via convex optimization. Available at optimization-online.org:2016/07/5536.pdf.
B. P. G. Van Parys, P. J. Goulart and D. Kuhn (2016*b*), Generalized Gauss inequalities via semidefinite programming, *Math. Program.* **156**, 271–302.
B. P. G. Van Parys, P. J. Goulart and M. Morari (2019), Distributionally robust expectation inequalities for structured distributions, *Math. Program.* **173**, 251–280.
B. P. G. Van Parys, D. Kuhn, P. J. Goulart and M. Morari (2015), Distributionally robust control of constrained stochastic systems, *IEEE Trans. Automat. Control* **61**, 430–442.
B. P. G. Van Parys, P. Mohajerin Esfahani and D. Kuhn (2021), From data to decisions: Distributionally robust optimization is optimal, *Manag. Sci.* **67**, 3387–3402.
V. Vapnik (2013), *The Nature of Statistical Learning Theory*, Springer.
S. R. S. Varadhan (1966), Asymptotic probabilities and differential equations, *Commun. Pure Appl. Math.* **19**, 261–286.
R. Vershynin (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press.
C. Villani (2003), *Topics in Optimal Transportation*, American Mathematical Society.
C. Villani (2008), *Optimal Transport: Old and New*, Springer.
F. Vincent, W. Azizian, J. Malick and F. Iutzeler (2024), `skwdro`: A library for Wasserstein distributionally robust machine learning. Available at arXiv:2410.21231.
R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino and S. Savarese (2018), Generalizing to unseen domains via adversarial data augmentation, in *Advances in Neural Information Processing Systems 31* (S. Bengio *et al.*, eds), Curran Associates, pp. 5339–5349.
H. Vu, T. Tran, M.-C. Yue and V. A. Nguyen (2022), Distributionally robust fair principal components via geodesic descents, in *International Conference on Learning Representations (ICLR 2022)*.
M. J. Wainwright (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press.
B. Wang and R. Wang (2011), The complete mixability and convex minimization problems with monotone marginal densities, *J. Multivariate Anal.* **102**, 1344–1360.
C. Wang, R. Gao, W. Wei, M. Shafie-khah, T. Bi and J. P. Catalao (2018), Risk-based distributionally robust optimal gas-power flow with Wasserstein distance, *IEEE Trans. Power Syst.* **34**, 2190–2204.
I. Wang, C. Becker, B. Van Parys and B. Stellato (2024*a*), Mean robust optimization, *Math. Program.* Available at doi:10.1007/s10107-024-02170-4.
I. Wang, C. Becker, B. P. G. Van Parys and B. Stellato (2023), Learning decision-focused uncertainty sets in robust optimization. Available at arXiv:2305.19225.
J. Wang, R. Gao and Y. Xie (2021), Sinkhorn distributionally robust optimization. Available at arXiv:2109.11926.
J. Wang, R. Gao and Y. Xie (2024*b*), Regularization for adversarial robust learning. Available at arXiv:2408.09672.
R. Wang, L. Peng and J. Yang (2013), Bounds for the sum of dependent risks and worst value-at-risk with monotone marginal densities, *Finance Stoch.* **17**, 395–417.
S. Wang (2024), The power of simple menus in robust selling mechanisms, *Manag. Sci.* Available at doi:10.1287/mnsc.2023.03738.
S. Wang, Z. Chen and T. Liu (2020), Distributionally robust hub location, *Transport. Sci.* **54**, 1189–1210.

S. Wang, S. Liu and J. Zhang (2024*c*), Minimax regret robust screening with moment information, *Manuf. Service Oper. Manag.* **26**, 992–1012.

Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou and Q. Gu (2019), On the convergence and robustness of adversarial training, in *36th International Conference on Machine Learning*, Vol. 97 of Proceedings of Machine Learning Research, PMLR, pp. 6586–6595.

Y. Wang, V. A. Nguyen and G. A. Hanasusanto (2024*d*), Wasserstein robust classification with fairness constraints, *Manuf. Service Oper. Manag.* **26**, 1567–1585.

Y. Wang, M. N. Prasad, G. A. Hanasusanto and J. J. Hasenbein (2024*e*), Distributionally robust observable strategic queues, *Stoch. Syst.* **14**, 229–361.

Z. Wang, P. W. Glynn and Y. Ye (2016), Likelihood robust optimization for data-driven problems, *Comput. Manag. Sci.* **13**, 241–261.

J. Weed and F. Bach (2019), Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance, *Bernoulli* **25**, 2620–2648.

P. Whittle (1990), *Risk-Sensitive Optimal Control*, Wiley.

W. Wiesemann, D. Kuhn and B. Rustem (2013), Robust Markov decision processes, *Math. Oper. Res.* **38**, 153–183.

W. Wiesemann, D. Kuhn and M. Sim (2014), Distributionally robust convex optimization, *Oper. Res.* **62**, 1358–1376.

D. Wozabal (2012), A framework for optimization under ambiguity, *Ann. Oper. Res.* **193**, 21–47.

D. Wozabal (2014), Robustifying convex risk measures for linear portfolios: A nonparametric approach, *Oper. Res.* **62**, 1302–1315.

Q. Wu, J. Y.-M. Li and T. Mao (2022), On generalization and regularization via Wasserstein distributionally robust optimization. Available at arXiv:2212.05716.

S. Wu, S. Sun, J. A. Camilleri, S. B. Eickhoff and R. Yu (2021), Better the devil you know than the devil you don't: Neural processing of risk and ambiguity, *NeuroImage* **236**, art. 118109.

W. Xie (2020), Tractable reformulations of distributionally robust two-stage stochastic programs over the type-∞ Wasserstein ball, *Oper. Res. Lett.* **48**, 513–523.

W. Xie (2021), On distributionally robust chance constrained programs with Wasserstein distance, *Math. Program.* **186**, 115–155.

W. Xie, S. Ahmed and R. Jiang (2022), Optimized Bonferroni approximations of distributionally robust joint chance constraints, *Math. Program.* **191**, 79–112.

W. Xie and S. Ahmed (2017), Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation, *IEEE Trans. Power Syst.* **33**, 1860–1867.

L. Xin and D. A. Goldberg (2021), Time (in)consistency of multistage distributionally robust inventory models with moment constraints, *European J. Oper. Res.* **289**, 1127–1141.

L. Xin and D. A. Goldberg (2022), Distributionally robust inventory control when demand is a martingale, *Math. Oper. Res.* **47**, 2387–2414.

C. Xu, J. Lee, X. Cheng and Y. Xie (2024), Flow-based distributionally robust optimization, *IEEE J. Select. Areas Inform. Theory* **5**, 62–77.

H. Xu, C. Caramanis and S. Mannor (2009), Robustness and regularization of support vector machines, *J. Mach. Learn. Res.* **10**, 1485–1510.

H. Xu, C. Caramanis and S. Mannor (2012*a*), A distributional interpretation of robust optimization, *Math. Oper. Res.* **37**, 95–110.

H. Xu, C. Caramanis and S. Mannor (2012*b*), Optimization under probabilistic envelope constraints, *Oper. Res.* **60**, 682–699.
V. A. Yakubovich (1971), S-procedure in nonlinear control theory (in Russian), *Vestnik Leninggradskogo Universiteta* pp. 62–77.
I. Yang (2018), A dynamic game approach to distributionally robust safety specifications for stochastic systems, *Automatica* **94**, 94–101.
I. Yang (2020), Wasserstein distributionally robust stochastic control: A data-driven approach, *IEEE Trans. Automat. Control* **66**, 3863–3870.
J. Yang, L. Zhang, N. Chen, R. Gao and M. Hu (2022), Decision-making with side information: A causal transport robust approach. Available at optimization-online.org:2022/10/DRO_with_side_info.pdf.
P. Yang and B. Chen (2018), Robust Kullback–Leibler divergence and universal hypothesis testing for continuous distributions, *IEEE Trans. Inform. Theory* **65**, 2360–2373.
W. Yang and H. Xu (2016), Distributionally robust chance constraints for non-linear uncertainties, *Math. Program.* **155**, 231–265.
I. Yanıkoğlu, B. L. Gorissen and D. den Hertog (2019), A survey of adjustable robust optimization, *European J. Oper. Res.* **277**, 799–813.
Y.-L. Yu, Y. Li, D. Schuurmans and C. Szepesvári (2009), A general projection property for distribution families, in *Advances in Neural Information Processing Systems 22* (Y. Bengio *et al.*, eds), Curran Associates, pp. 2232–2240.
Y. Yu, T. Lin, E. V. Mazumdar and M. Jordan (2022), Fast distributionally robust learning with variance-reduced min-max optimization, in *25th International Conference on Artificial Intelligence and Statistics*, Vol. 151 of Proceedings of Machine Learning Research, PMLR, pp. 1219–1250.
J. Yue, B. Chen and M.-C. Wang (2006), Expected value of distribution information for the newsvendor problem, *Oper. Res.* **54**, 1128–1136.
M.-C. Yue, D. Kuhn and W. Wiesemann (2022), On linear optimization over Wasserstein balls, *Math. Program.* **195**, 1107–1122.
G. Zames (1966), Robust control theory, *Proc. IEEE* **54**, 1442–1451.
O. Zeitouni and M. Gutman (1991), On universal hypotheses testing via large deviations, *IEEE Trans. Inform. Theory* **37**, 285–290.
Y. Zeng and H. Lam (2022), Generalization bounds with minimal dependency on hypothesis class via distributionally robust optimization, in *Advances in Neural Information Processing Systems 35* (S. Koyejo *et al.*, eds), Curran Associates, pp. 27576–27590.
A. Y. Zhang and H. H. Zhou (2020), Theoretical and computational guarantees of mean field variational inference for community detection, *Ann. Statist.* **48**, 2575–2598.
L. Zhang, J. Yang and R. Gao (2024*a*), Optimal robust policy for feature-based newsvendor, *Manag. Sci.* **70**, 2315–2329.
L. Zhang, J. Yang and R. Gao (2024*b*), A short and general duality proof for Wasserstein distributionally robust optimization, *Oper. Res.* Available at doi:10.1287/opre.2023.0135.
Y. Zhang, R. Jiang and S. Shen (2018), Ambiguous chance-constrained binary programs under mean–covariance information, *SIAM J. Optim.* **28**, 2922–2944.
C. Zhao and Y. Guan (2018), Data-driven risk-averse stochastic optimization with Wasserstein metric, *Oper. Res. Lett.* **46**, 262–267.
C. Zhao and R. Jiang (2017), Distributionally robust contingency-constrained unit commitment, *IEEE Trans. Power Syst.* **33**, 94–102.

J. Zhen, D. Kuhn and W. Wiesemann (2023), A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle, *Oper. Res.* **73**, 862–878.

K. Zhou and J. C. Doyle (1999), *Essentials of Robust Control*, Prentice Hall.

K. Zhou, J. C. Doyle and K. Glover (1996), *Robust and Optimal Control*, Prentice Hall.

B. Zhu, J. Jiao and J. Steinhardt (2022*a*), Generalized resilience and robust statistics, *Ann. Statist.* **50**, 2256–2283.

J.-J. Zhu, W. Jitkrittum, M. Diehl and B. Schölkopf (2020), Worst-case risk quantification under distributional ambiguity using kernel mean embedding in moment problem, in *59th IEEE Conference on Decision and Control (CDC)*, pp. 3457–3463.

J.-J. Zhu, W. Jitkrittum, M. Diehl and B. Schölkopf (2021), Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation, in *24th International Conference on Artificial Intelligence and Statistics*, Vol. 130 of Proceedings of Machine Learning Research, PMLR, pp. 280–288.

L. Zhu, M. Gürbüzbalaban and A. Ruszczyński (2023), Distributionally robust learning with weakly convex losses: Convergence rates and finite-sample guarantees. Available at arXiv:2301.06619.

S. Zhu, L. Xie, M. Zhang, R. Gao and Y. Xie (2022*b*), Distributionally robust weighted k-nearest neighbors, in *Advances in Neural Information Processing Systems 35* (S. Koyejo *et al.*, eds), Curran Associates, pp. 29088–29100.

M. Zorzi (2014), Multivariate spectral estimation based on the concept of optimal prediction, *IEEE Trans. Automat. Control* **60**, 1647–1652.

M. Zorzi (2016), Robust Kalman filtering under model perturbations, *IEEE Trans. Automat. Control* **62**, 2902–2907.

M. Zorzi (2017*a*), Convergence analysis of a family of robust Kalman filters based on the contraction principle, *SIAM J. Control Optim.* **55**, 3116–3131.

M. Zorzi (2017*b*), On the robustness of the Bayes and Wiener estimators under model uncertainty, *Automatica* **83**, 133–140.

L. F. Zuluaga and J. F. Pena (2005), A conic programming approach to generalized Tchebycheff inequalities, *Math. Oper. Res.* **30**, 369–388.

S. Zymler, D. Kuhn and B. Rustem (2013*a*), Distributionally robust joint chance constraints with second-order moment information, *Math. Program.* **137**, 167–198.

S. Zymler, D. Kuhn and B. Rustem (2013*b*), Worst-case value at risk of nonlinear portfolios, *Manag. Sci.* **59**, 172–188.

https://doi.org/10.1017/S0962492924000084 Published online by Cambridge University Press