Adv. Appl. Prob. **47,** 292–305 (2015)
Printed in Northern Ireland

ON THE RANDOM SAMPLING OF PAIRS, WITH PEDESTRIAN EXAMPLES

RICHARD ARRATIA,* *University of Southern California*
STEPHEN DESALVO,** *University of California, Los Angeles*

Abstract

For a collection of objects such as socks, which can be matched according to a characteristic such as color, we study the innocent phrase 'the distribution of the color of a matching pair' by looking at two methods for selecting socks. One method is memoryless and effectively samples socks with replacement, while the other samples socks sequentially, with memory, until the same color has been seen twice. We prove that these two methods yield the same distribution on colors if and only if the initial distribution of colors is a uniform distribution. We conjecture a nontrivial maximum value for the total variation distance of these distributions in all other cases.

Keywords: Total variation distance; random sampling; computational algebraic geometry; Poisson process; pair-derived distribution

2010 Mathematics Subject Classification: Primary 65C50
Secondary 60C05

1. Motivation

The problem that inspires us is the following. Suppose that a drawer has 12 white and 4 black socks. How many socks must one remove to ensure a pair of matching color? The answer, 3, illustrates the pigeon-hole principle. The statement of detailed counts, 12 and 4, was arbitrary, but leads to the problem that we address in this paper: what is the distribution of the color of a matching pair?

To simplify, we take the limit as the number of socks in the drawer goes to infinity while the proportions remain constant, e.g. 75 percent white and 25 percent black.

We consider two sensible methods for choosing 'a matching pair'.

(M1) Select objects two at a time until a pair of the same color is selected in a single round.

(M2) Select objects one at a time until the first pair of the same color is found.

For a second example, if there are 365 equally likely colors for socks then, under method (M2), the maximum number of socks inspected is 366, but the expected number is 23.6166 . . .. In contrast, the expected number of pairs inspected under method (M1) is exactly 365; hence, the expected number of socks inspected is 730. However, our focus is not on the *number* of socks inspected, but rather on the distribution of the *color* of the matching pair.

In our first example, under method (M1) the odds for a white pair over a black pair are $(\frac{12}{16})^2$ to $(\frac{4}{16})^2$; equivalently, 12^2 to 4^2, or 3^2 to 1^2, so $\frac{9}{10}$th of the time the pair is white,

Received 1 October 2013; revision received 8 April 2014.
* Postal address: Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA.
** Postal address: Department of Mathematics, University of California, Los Angeles, CA 90095, USA.
Email address: stephen.desalvo@gmail.com

and $\frac{1}{10}$th of the time it is black. Under method (M2), the outcomes resulting in a white pair correspond to ww, bww, and wbw, with total probability $(0.75)^2+2(0.75)^2(0.25)^2 = \frac{27}{32}$, and the outcomes resulting in a black pair correspond to bb, wbb, and bwb, with total probability $(0.25)^2 + 2(0.75)(0.25)^2 = \frac{5}{32}$.

To summarize, the input is a distribution on colors, $\boldsymbol{p} = (0.75, 0.25)$, and there are two outputs: under method (M1), the color of a pair is white with probability 0.9 and black with probability 0.1, while under method (M2), the color of a pair is white with probability $\frac{27}{32}$ and black with probability $\frac{5}{32}$:

$$\boldsymbol{p} = (0.75, 0.25), \qquad \mathrm{M1}(\boldsymbol{p}) = (0.9, 0.1), \qquad \mathrm{M2}(\boldsymbol{p}) = (0.843\,75, 0.156\,25).$$

Some natural questions that arise are: for an arbitrary discrete distribution $\boldsymbol{p}$, for the color of a single sock,

(Q1) when does $\mathrm{M1}(\boldsymbol{p}) = \mathrm{M2}(\boldsymbol{p})$?

(Q2) how far apart can $\mathrm{M1}(\boldsymbol{p})$ and $\mathrm{M2}(\boldsymbol{p})$ be from each other?

There are practical algorithms [1] for sampling, exploiting the birthday paradox, that require getting a matching pair whose color has the distribution (M1), but, under a naive *opportunistic* implementation, would find only a pair whose color is distributed according to (M2). Question (Q2) concerns quantifying the error that would result from using the opportunistic implementation.

2. Pair-derived distributions

In general, we write S for the random color of a single sock, and describe the initial distribution of colors with

$$p_i := \mathbb{P}(S = i).$$

When the number of colors is finite, say $n+1$, then we let the colors be $0, 1, 2, \ldots, n$, and the distribution of S is given by $\boldsymbol{p} = (p_0, p_1, \ldots, p_n)$. Our initial example had $n+1 = 2$ and $\boldsymbol{p} = (p_0, p_1) = (0.75, 0.25)$. When the number of colors is infinite, we take the colors to be $0, 1, 2, \ldots$, and then $\boldsymbol{p} = (p_0, p_1, p_2, \ldots)$.

Method (M1) may be described as the color X of a pair of randomly chosen socks, conditional on getting a match. More precisely, the two chosen socks have colors S and S', and are independent and identically distributed. We write

$$f_2 := \mathbb{P}(S = S') = \sum_i \mathbb{P}(S = S' = i) = \sum_i p_i^2 \tag{1}$$

for the probability that two randomly chosen socks match, so

$$\mathbb{P}(X = i) = \mathbb{P}(S = i \mid S = S') = \frac{p_i^2}{f_2}. \tag{2}$$

Method (M2) involves a sequential procedure: pick socks one at a time until a duplicate color is found. Suppose that when this duplicate is found, there have been k *other* colors, with $k = 0, 1, 2, \ldots$. Write i for the duplicate color, and $J = \{j_1, \ldots, j_k\}$ for the single colors, so $i \notin J$ and $|J| = k$. The second occurrence of color i is at time $k+2$, and, for the first

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

$k+1$ socks, any permutation of the colors in $\{i\} \cup J$ is valid. Hence, the color Y of the matching pair found by method (M2) has distribution given by

$$\mathbb{P}(Y = i) = p_i^2 \sum_k (k+1)! \sum_J p_{j_1} \cdots p_{j_k}. \tag{3}$$

In the sum above, $|J| = k$ and $i \notin J$.

3. When are the two pair-picking methods the same?

A discrete distribution is said to be *uniform* if it has finite support, say of size $n+1$, and, for each color i in the support, $p_i = 1/(n+1)$.

Proposition 1. *If* $\boldsymbol{p}$ *is uniform then* $\mathrm{M1}(\boldsymbol{p}) = \mathrm{M2}(\boldsymbol{p})$.

Proof. If $\boldsymbol{p}$ is a uniform distribution then both $\mathrm{M1}(\boldsymbol{p})$ and $\mathrm{M2}(\boldsymbol{p})$ are equal to the original uniform distribution—by the principle of ignorance, all possible colors are alike, and, hence, equally likely under each of the derived methods. One could alternatively calculate the point probabilities of X and Y.

The converse is true, but not so easy to prove; we will first prove an ancillary result in Lemma 1 below and then summarize in Theorem 1 below.

Lemma 1. *Under method (M2), as specified by (3), if*

$$p_i \geq p_j > 0 \quad \textit{then} \quad \frac{\mathbb{P}(Y=i)}{p_i^2} \leq \frac{\mathbb{P}(Y=j)}{p_j^2}; \tag{4}$$

hence, if

$$p_i = p_j > 0 \quad \textit{then} \quad \mathbb{P}(Y=i) = \mathbb{P}(Y=j).$$

Also, if

$$p_i > p_j > 0 \quad \textit{then} \quad \frac{\mathbb{P}(Y=i)}{p_i^2} < \frac{\mathbb{P}(Y=j)}{p_j^2}. \tag{5}$$

Proof. Assume that $p_i \geq p_j > 0$. Define $t(i,k)$ to be the inner sum of (3), so

$$\frac{\mathbb{P}(Y=i)}{p_i^2} = \sum_k (k+1)!\, t(i,k).$$

To prove (4), it suffices to show that if $p_i \geq p_j > 0$ then $t(i,k) \leq t(j,k)$ for all k, and to further prove (5), it suffices to show that if $p_i > p_j$ then $t(i,k) < t(j,k)$ for at least one k. With sums always taken over sets of size k,

$$t(i,k) = \sum_{i \notin J} p_{i_1} \cdots p_{i_k} = \sum_{i \notin J,\, j \in J} p_{i_1} \cdots p_{i_k} + \sum_{i,j \notin J} p_{i_1} \cdots p_{i_k},$$

i.e. in the sum over sets J excluding i, we take cases according to whether or not $j \in J$. With a similar decomposition of $t(j,k)$, taking the difference yields

$$t(i,k) - t(j,k) = k(p_j - p_i) \sum_{J' :\, i,j \notin J'} p_{i_1} \cdots p_{i_{k-1}},$$

where the set $J' = \{i_1, \ldots, i_{k-1}\}$ has size $|J'| = k-1$, and corresponds to $|J|$ with the element i or j removed.

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

Theorem 1. *Over all discrete distributions* $\boldsymbol{p}$*, the derived distributions of* X *and* Y*, given by (2) and (3), are equal if and only if* $\boldsymbol{p}$ *is a uniform distribution.*

Proof. Suppose that $\boldsymbol{p}$ is *not* a uniform distribution. Then we can fix i, j with $p_i > p_j > 0$. From (5) we obtain

$$\frac{\mathbb{P}(Y = i)}{p_i^2} < \frac{\mathbb{P}(Y = j)}{p_j^2},$$

and dividing by f_2 to relate with (2), and rearranging,

$$\frac{\mathbb{P}(X = i)}{\mathbb{P}(X = j)} > \frac{\mathbb{P}(Y = i)}{\mathbb{P}(Y = j)}, \tag{6}$$

which implies that X and Y have different distributions.

Theorem 1 gives a complete answer to our first question: when are the two pair-picking methods the same? Next we turn to the second question: when the two methods are different, how different can they be?

4. Total variation distance

We wish to quantify the following. Given a probability distribution $\boldsymbol{p}$, with the matching pair chosen by method (M1) or method (M2), how far apart are the two distributions with respect to the color of the matching pair?

A metric on the space of all probability measures is the *total variation distance*.

Definition 1. For two real-valued random variables X and Y, the total variation distance between the *laws* of X and Y is defined as

$$d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_{A \subseteq \mathbb{R}} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|,$$

where the sup is taken over all Borel sets $A \subseteq \mathbb{R}$. When there is no confusion, we write $d_{\mathrm{TV}}(X, Y)$ instead of $d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y))$.

This choice of definition is useful for probability, with the desirable property that $d_{\mathrm{TV}}(X, Y) \leq 1$, and it equals $\sup_{\{f\colon \mathbb{R} \to [0,1]\}} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$.

When X and Y are discrete random variables, an equivalent definition is

$$d_{\mathrm{TV}}(X, Y) = \frac{1}{2} \sum_k |\mathbb{P}(X = k) - \mathbb{P}(Y = k)|.$$

Furthermore, since $\sum_k \mathbb{P}(X = k) = \sum_k \mathbb{P}(Y = k)$, we can divide the summands into positive and negative parts to obtain two more equivalent definitions.

Lemma 2. *For each* $t \in \mathbb{R}$*, let* $t^+ = \max(0, t)$ *and* $t^- = \max(0, -t)$ *denote the positive part and negative parts of* t*, respectively; hence,* $|t| = t^+ + t^-$ *and* $t = t^+ - t^-$*. Then*

$$d_{\mathrm{TV}}(X, Y) = \sum_k (\mathbb{P}(X = k) - \mathbb{P}(Y = k))^+ = \sum_k (\mathbb{P}(X = k) - \mathbb{P}(Y = k))^-. \tag{7}$$

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

For example, when X is a Bernoulli random variable with parameter θ, so that $\mathbb{P}(X = 1) = \theta = 1 - \mathbb{P}(X = 0)$, and Y is Bernoulli with parameter θ', the total variation distance is $|\theta - \theta'|$.

Since our sample space is discrete, and the labels of the socks have no intrinsic meaning, it does not make sense to consider metrics such as the Wasserstein distance, which assign a metric on the sample space. A popular alternative is the Kullbach–Liebler divergence, or relative entropy, which has the undesirable property of being asymmetric.

Definition 2. Given a discrete probability distribution $\boldsymbol{p}$, let X have the method (M1) distribution given by (2), let Y have the method (M2) distribution given by (3), and define the *discrepancy* of $\boldsymbol{p}$ by

$$D(\boldsymbol{p}) = d_{\mathrm{TV}}(X(\boldsymbol{p}), Y(\boldsymbol{p})).$$

We could have written $D(\boldsymbol{p}) = d_{\mathrm{TV}}(X, Y)$ as above, but we preferred $d_{\mathrm{TV}}(X(\boldsymbol{p}), Y(\boldsymbol{p}))$, to emphasize that $D(\boldsymbol{p})$ is the total variation distance between two probability laws, with each law being a function of a third underlying law $\boldsymbol{p}$.

5. Special cases

5.1. Dimension $n = 1$: two colors of socks

In the $n = 1$ case, we write $\boldsymbol{p} = (p_0, p_1) = (x, 1-x)$. The discrepancy $D(\boldsymbol{p}) = d_{\mathrm{TV}}(X, Y)$ simplifies, via Lemma 2, to $|d_1|$, where

$$d_1(x) = \mathbb{P}(X = 0) - \mathbb{P}(Y = 0) = \frac{x^2}{x^2 + (1-x)^2} - (x^2 + 2(1-x)x^2).$$

The expression $|d_1(x)|$ is plotted in Figure 1.

Since d_1 is a rational function in one variable, it is easily optimized over $x \in [0, 1]$. We obtain five critical numbers: 0, 1, $\frac{1}{2}$,

$$x_1 := \frac{1}{6}\left(3 + \sqrt{3(-3 + 2\sqrt{3})}\right) \doteq 0.696\,660,$$

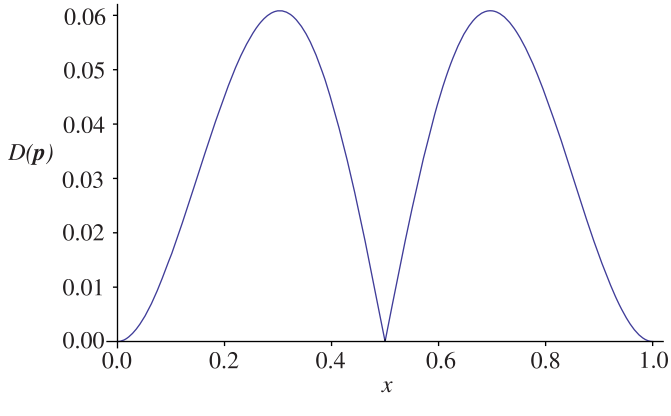

FIGURE 1: Plot of $D(\boldsymbol{p})$ for $\boldsymbol{p} = (x, 1 - x)$, as a function of $x \in [0, 1]$.

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

and the conjugate, $1 - x_1$. The cusp for $|d_1(x)|$ at $x = \frac{1}{2}$ is also critical, with $|d_1(\frac{1}{2}) = 0|$ corresponding to the uniform case. Evaluating $|d_1(x)|$ at these five critical numbers exhausts all possible extremes, and the maximum value is $d_1(x_1) = 1/\sqrt{135 + 78\sqrt{3}} \doteq 0.060\,846\,8$.

5.2. Dimension $n = 2$: three colors of socks

The $n = 2$ case can be set up similarly to the $n = 1$ case, but now we have three cases of possible signs underlying absolute values. Each case is a smooth, two-dimensional surface, and we find extremes by checking all critical values arising from points where the gradient vanishes and are on the boundary. To avoid subscripts, we switch notation from $\boldsymbol{p} = (p_0, p_1, p_2)$ to $\boldsymbol{p} = (a, b, c)$, and define

$$f(a, b, c) := a^2(1 + 2(b + c) + 6bc), \qquad T(a, b, c) = \frac{a^2}{a^2 + b^2 + c^2} - f(a, b, c),$$

so, when $\boldsymbol{p} = (a, b, c)$, with a being the probability that a single sock has color 0, $T(a, b, c) = \mathbb{P}(X = 0) - \mathbb{P}(Y = 0)$. By symmetries involving b and c, we have

$$2D(\boldsymbol{p}) = |T(a, b, c)| + |T(b, a, c)| + |T(c, a, b)|.$$

A major obstacle to this approach is the boundary, which is complicated, so instead we parameterize in terms of $(x, y) \in [0, 1]^2$ as follows:

$$\boldsymbol{p}(x, y) = (a, b, c), \quad \text{where} \quad t = 1 + x + y, \qquad a = \frac{1}{t}, \qquad b = \frac{x}{t}, \qquad c = \frac{y}{t}.$$

Now taking $a = a(x, y)$ and so on, we have three functions defined on $[0, 1]^2$, namely,

$$T_1(x, y) := T(a, b, c), \qquad T_2(x, y) := T(b, a, c), \qquad T_3(x, y) := T(c, a, b),$$

and so the total variation distance is given by

$$2d_{\mathrm{TV}}(X, Y) = |T_1(x, y)| + |T_2(x, y)| + |T_3(x, y)|.$$

Since $1 \geq x, y$, we have $a \geq b, c$, and, since the largest mass is at 1, we know that, for all $x, y \in [0, 1]$, $T_1(x, y) \geq 0$. We can eliminate the case in which $T_1 \geq 0$, $T_2 \geq 0$, and $T_3 \geq 0$, as this implies that $T_1 = T_2 + T_3 = 0$ since $T_1 + T_2 + T_3 = 0$. By Lemma 2, this case gives $D(\boldsymbol{p}) = 0$, not of interest in the search for the maximum value. There are three remaining cases of sign to consider. Let

$$\begin{aligned} d_1(x, y) &= T_1(x, y) + T_2(x, y) - T_3(x, y), \\ d_2(x, y) &= T_1(x, y) - T_2(x, y) + T_3(x, y), \\ d_3(x, y) &= T_1(x, y) - T_2(x, y) - T_3(x, y). \end{aligned}$$

Then $\max d_{\mathrm{TV}}(X, Y) = \max(d_1, d_2, d_3)$, and so it suffices to check the maximum values of each of these rational functions.

Let us consider $g(x, y) := d_1(x, y)$. Since g is a rational function in two variables, Bezout's theorem (see, for example, Theorem 10 of [3, page 420]) guarantees in this case a total of $7 \times 7 = 49$ complex solutions, although some of these may be 'at infinity'. MATHEMATICA® produces a set of 19 unique, easily verified solutions; when including multiplicities, this accounts for 39 of the total solutions. By hand we can find 10 solutions at ∞, so all 49 solutions have been addressed.

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

The term d_2 becomes d_1 under the interchange of x and y, so no further work is required for d_2. For d_3, the corresponding variety has $6 \times 6 = 36$ solutions, and a similar calculation accounts for the 36 solutions guaranteed by Bezout's theorem.

We obtain the largest value of d_{TV} from the point (x, y) given by

$$1 + 4x - 14x^2 - 4x^3 - 34x^4 + 20x^5 = 0, \qquad y = x, \quad \text{for } x \in (0, 1),$$

with $2d_{TV}$ given by the value $z \in (0, 0.2)$ that solves

$$32\,000 + 168\,192z - 4\,557\,600z^2 + 14\,567\,472z^3 - 821\,583z^4 + 314\,928z^5 = 0. \quad (8)$$

This solution is of the form $\boldsymbol{p} = (x_2, (1 - x_2)/2, (1 - x_2)/2)$ for the value of $x_2 \in [0.5, 0.6]$ that solves $-5 + 42x_2 - 114x_2^2 + 168x_2^3 - 153x_2^4 + 54x_2^5 = 0$, with

$$x_2 \doteq 0.582\,011, \qquad D(\boldsymbol{p}) \doteq 0.084\,294\,2;$$

the exact value of $D(\boldsymbol{p})$ is given by (8).

6. Conjectures about the largest possible discrepancy

The weakest conjecture is that there is some nontrivial upper bound on the discrepancy. Formally, we define the universal constant for the pair discrepancy by

$$\ell_0 := \sup_{\boldsymbol{p}} D(\boldsymbol{p}), \quad (9)$$

where the supremum is over all distributions $\boldsymbol{p}$ on a finite or countable set of colors. Since the total variation distance is always less than or equal to 1, trivially $\ell_0 \leq 1$, and the conjecture is as follows.

Conjecture 1. *The constant defined by (9) is* strictly *less than* 1, *i.e.* $\ell_0 < 1$.

6.1. Conjectures for a finite number of colors

If there are a finite number of colors, say $n + 1$ with $n \geq 0$, then we can relabel the colors as $0, 1, \ldots, n$ so that $\boldsymbol{p} = (p_0, \ldots, p_n)$ with

$$p_0 \geq p_1 \geq \cdots \geq p_n \geq 0, \qquad p_0 + p_1 + \cdots + p_n = 1. \quad (10)$$

Given $n > 0$ and $x \in [1/(n + 1), 1)$, let

$$\boldsymbol{p}(n, x) = \left(x, \frac{1 - x}{n}, \ldots, \frac{1 - x}{n}\right), \quad (11)$$

which, since to $x \in [1/(n + 1), 1)$, satisfies (10).

For each $n > 0$, (11) defines a *one-parameter family* of probability distributions. At the endpoint $x = 1/(n+1)$, $\boldsymbol{p}(n, x)$ is a uniform distribution. Now suppose that $x \in (1/(n+1), 1)$, so that $\boldsymbol{p}(n, x)$ has $p_0 > p_1 = p_2 = \cdots = p_n > 0$. It is obvious from (2) that $\mathbb{P}(X = 0) > \mathbb{P}(X = 1) = \cdots = \mathbb{P}(X = n) > 0$, and Lemma 1 implies that $\mathbb{P}(Y = 0) > \mathbb{P}(Y = 1) = \cdots = \mathbb{P}(Y = n) > 0$. That is, both X and Y have distributions in the same one-parameter family. Finally, (6) implies that $\mathbb{P}(X = 0) > \mathbb{P}(Y = 0)$, while, for $i = 1$ to n, $\mathbb{P}(X = i) < \mathbb{P}(Y = i)$,

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

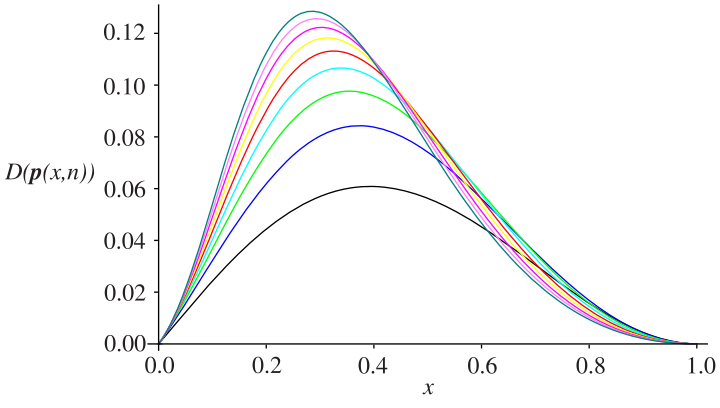

FIGURE 2: The discrepancy $D(\boldsymbol{p})$ for the one-parameter families (11), $n = 1$ to 9 (in order from top to bottom at $x = 0.2$). For each n, we plot $((n+1)/n)x - 1/n$ versus $D(\boldsymbol{p}(x,n))$ so that all nine graphs have domain [0,1].

and, hence, using (7), for each $n > 0$ and $x \in (1/(n+1), 1)$, $\boldsymbol{p} = \boldsymbol{p}(n,x)$ has the following simplified expression for its discrepancy:

$$D(\boldsymbol{p}) = \mathbb{P}(X=0) - \mathbb{P}(Y=0) = \frac{x^2}{x^2 + (1-x)^2/n} - x^2 \sum_{k=0}^{n} (k+1)! \binom{n}{k} \left(\frac{1-x}{n}\right)^k. \quad (12)$$

We present plots of these functions, appropriately scaled, in Figure 2.

Conjecture 2. *For every nonnegative integer n, among all probability distributions on $n+1$ colors, the maximum value of $D(\boldsymbol{p})$ is achieved by a distribution of the form $\boldsymbol{p}(n, x_n)$.*

A slightly stronger conjecture is the following.

Conjecture 3. *For every nonnegative integer n, among all probability distributions on $n+1$ colors, the maximum value of $D(\boldsymbol{p})$ is achieved* uniquely *by $\boldsymbol{p}(n, x_n)$, where $x_n = \operatorname{argmax}_x D(\boldsymbol{p}(n,x))$.*

We cannot prove Conjecture 2, but we believe it to be true for the following reasons.

- It is true trivially for $n = 0$ and $n = 1$, and, by Section 5.2, for $n = 2$.
- By broad analogy, many symmetric payoff functions achieve their extreme values at points with lots of symmetry. Indeed, Theorem 1 asserts that, for each n, $D(\boldsymbol{p})$ achieves its *minimum* value, 0, at the uniform distribution, corresponding to the maximum conceivable symmetry in $\boldsymbol{p}$, while the family in (11) corresponds to *breaking* symmetry somewhat, but as little as possible.
- The one-parameter family (11) arises in other extremal problems which share the feature that the *labels* on the colors are irrelevant, and only the values of the probabilities matter. In particular, in information theory, the one-parameter families show that 'Fano's inequality is sharp'; see Cover and Thomas [2, Equation (2.135), page 40].

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

- For the moderate values $n = 3, 4, \ldots, 8$, when generating a million random points from the n-dimensional region specified by (10), the largest observed $D(\boldsymbol{p})$ in the sample came from a $\boldsymbol{p}$ that was close, by eye, to the form of (11).

7. Limit analysis of the one-parameter family

Theorem 2 describes a limit function $\ell(c)$, and a plot of this function is shown in Figure 3.

Theorem 2. *For $c \in (0, \infty)$, define*

$$\ell(c) = \frac{c^2}{1+c^2} - \int_0^\infty c^2 t \mathrm{e}^{-ct-t^2/2}\,\mathrm{d}t. \tag{13}$$

For any $c \in (0, \infty)$ and $n > 1/c^2$, let $\boldsymbol{p}^{(n)} = \boldsymbol{p}(n, c/\sqrt{n})$ be the distribution governed by (11) with $x = c/\sqrt{n}$. Then

$$\lim_{n\to\infty} D(\boldsymbol{p}^{(n)}) = \ell(c), \tag{14}$$

where ℓ is defined by (13).

Proof. Extend method (M2) beyond the time of the first matching pair, i.e. pick socks forever. For each color i, let N_i be the number of sock picks needed to get the second sock of color i. As the color varies, these random variables are *dependent*, since, for any two distinct colors i, j and time $n \geq 2$, $0 = \mathbb{P}(N_i = N_j = n) < \mathbb{P}(N_i = n)\,\mathbb{P}(N_j = n)$. There is a standard technique to deal with this dependence, used in Markov chains (see, for example, [6]), which is to take a sequence of independent, exponentially distributed holding times $Y_1, Y_2, \ldots$, with $\mathbb{P}(Y_n > t) = \mathrm{e}^{-t}$, and declare that the nth sock arrives at time $Y_1 + Y_2 + \cdots + Y_n$. The number of socks picked by time t is thus Poisson distributed, with mean t. Let $C_i(t)$ denote the number of socks of color i chosen by time t. As i varies, the counts $C_i(t)$ are mutually independent; this observation is known as *Poissonization* (see, for example, Exercise XII.6.3 of [4]). With values in $(0, \infty)$, the time T_i at which color i is first seen for the second time can be expressed as $T_i = Y_1 + \cdots + Y_{N_i}$. The distribution of the color of the first matching pair found, initially specified by (3), can also be expressed as

$$\mathbb{P}(Y = i) = \mathbb{P}\Big(T_i < \min_{j\neq i} T_j\Big).$$

For each color i, the times at which socks of color i arrive form a Poisson arrivals process with rate p_i, and, as the color varies, these processes are mutually independent; in particular, the second arrival times T_i are mutually independent.

We are considering socks distributed according to $\boldsymbol{p}(n, c/\sqrt{n})$, that is, with $y := (1 - c/\sqrt{n})$,

$$p_0 = \frac{c}{\sqrt{n}}, \qquad p_1 = \frac{y}{n}, \qquad p_2 = \frac{y}{n}, \ldots, p_n = \frac{y}{n}. \tag{15}$$

Speed up time by a factor of $\sqrt{n}$; now socks of color 0 arrive at rate c, and, for each other color $i = 1$ to n, socks of color i arrive at rate $p_i\sqrt{n} = y/\sqrt{n}$. For $t > 0$ and each $i = 1$ to n, the number Z of socks of color i collected by time t is Poisson with parameter $\lambda = ty/\sqrt{n}$, and the event $\{T_i > t\}$ is the event $\{Z < 2\} = \{Z = 0 \text{ or } 1\}$, with probability

$$\mathbb{P}(T_i > t) = \exp\left(-\frac{ty}{\sqrt{n}}\right)\left(1 + \frac{ty}{\sqrt{n}}\right) = 1 - \frac{t^2y^2}{2n} + O(n^{-3/2}).$$

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

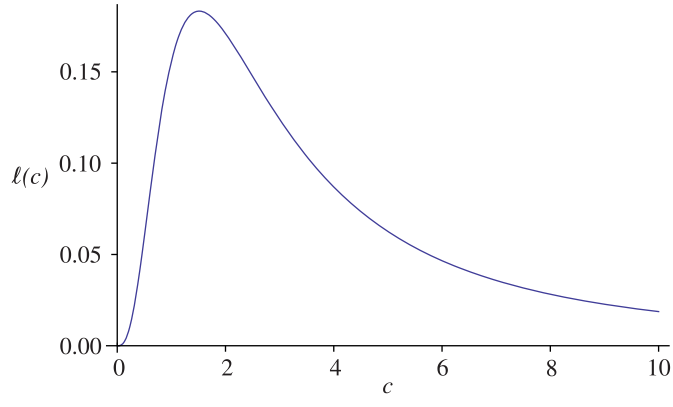

FIGURE 3: Plot of c versus $\ell(c)$ for $c = 0$ to 10. The maximum occurs at $c_0 \doteq 1.514$ and has the value $\ell(c_0) \doteq 0.183\,20$.

The event $\{\min(T_1, \ldots, T_n) > t\}$ is the intersection of the events $\{T_i > t\}$, so, using the mutual independence, together with $y \to 1$, for each $t > 0$,

$$\mathbb{P}(\min(T_1, \ldots, T_n) > t) = \left(1 - \frac{t^2 y^2}{2n} + O(n^{-3/2})\right)^n \to \exp\left(-\frac{t^2}{2}\right).$$

Finally, we argue that the density of T_0, the second arrival time in a Poisson process with rate c, is given by $f(t) = c^2 t \mathrm{e}^{-ct}$. This is a standard fact, known to some as the density of the gamma distribution with shape parameter 2 and scale parameter c. Using the independence of T_0 and $\min(T_1, \ldots, T_n)$, we can condition on the value t for T_0 to obtain

$$\begin{aligned}\mathbb{P}_n(Y = 0) &= \mathbb{P}(\min(T_1, \ldots, T_n) > T_0) \\ &= \int_0^\infty \mathbb{P}(\min(T_1, \ldots, T_n) > t) c^2 t \mathrm{e}^{-ct}\, \mathrm{d}t \\ &\to \int_0^\infty c^2 t \mathrm{e}^{-ct} \mathrm{e}^{-t^2/2}\, \mathrm{d}t. \end{aligned} \tag{16}$$

The above amounts to a calculation of the limit, as $n \to \infty$, of $\mathbb{P}_n(Y = 0)$, corresponding to method (M2) when the underlying colors come from (15).

Of course, we must justify the passage to the limit in (16). Here we have $f_n(t) := \mathbb{P}(\min(T_1, \ldots, T_n)) \to \exp(-t^2/2) =: f(t)$ pointwise for each $t > 0$, but we claim in (16) that the integrals also converge. Interpreting the (improper) integral as the Lebesgue integral, or alternatively as the Riemann integral, proving convergence in either case is a straightforward exercise; see, for example, Exercise 15 of [5, Chapter 2] for the Lebesgue integral and Exercise 12 of [7, Chapter 7] for the Riemann integral.

For method (M1) the calculation is easier: using (1), we have $f_2 = p_0^2 + p_1^2 + \cdots + p_n^2 = (c/\sqrt{n})^2 + n(y/n)^2 = c^2/n + y^2/n$ and

$$\mathbb{P}_n(X = 0) = \frac{p_0^2}{f_2} = \frac{c^2/n}{c^2/n + y^2/n} = \frac{c^2}{c^2 + y^2} \to \frac{c^2}{c^2 + 1}.$$

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

At (12) we had already argued that once n is large enough that $p_0 > p_1$, we have the simplification, for our one-parameter family, that $D(\boldsymbol{p}^{(n)}) = \mathbb{P}_n(X = 0) - \mathbb{P}_n(Y = 0)$. Combining this calculation of $D(\boldsymbol{p}^{(n)})$ with the limit values derived for $\mathbb{P}_n(Y = 0)$ and $\mathbb{P}_n(X = 0)$, (14) follows.

We note that instead of invoking Poissonization, as in the above proof, we can argue directly with the explicit expression in (12), to show that, under $x = c/\sqrt{n}$ and $k = t\sqrt{n}$, the sum in (12) is a Riemann approximation for $\int_0^\infty c^2 t \mathrm{e}^{-ct} \mathrm{e}^{-t^2/2}\, \mathrm{d}t$.

8. Discussion

If Conjecture 2 is true, it will follow that Conjecture 1 is also true, with the value of the universal constant for a pair of socks given by

$$\ell_0 = \sup_c \ell(c) = 0.183\,200\,062\,408\,710\,6\ldots. \tag{17}$$

The argument requires two parts. The first part is to show that ℓ_0, defined in (9) as the sup of $D(\boldsymbol{p})$ over *all discrete* distributions, is equal to the sup over distributions with *finite* support. This is a 'soft' analysis, showing first that $\boldsymbol{p} \mapsto D(\boldsymbol{p})$ is continuous; hence, given $\boldsymbol{p}$ with discrepancy greater than $\ell_0 - \varepsilon$, we can find a nearby distribution $\boldsymbol{p}'$ with finite support, close enough to $\boldsymbol{p}$ to guarantee that its discrepancy is greater than $\ell_0 - 2\varepsilon$. The second part, giving the concrete value for ℓ_0, uses compactness: given distributions $\boldsymbol{p}^{(n)} = \boldsymbol{p}(n, x_n)$ with discrepancies converging to ℓ_0, the values $c_n := x_n\sqrt{n} \in [0, \infty]$, $n \geq 1$, lie in a compact set, and, hence, there must be convergent subsequences. If $c_{n_k} \to c_0$ and $c_0 \in (0, \infty)$, then the proof of Theorem 2 already shows that the associated discrepancies converge to $\ell(c_0)$. If $c_{n_k} \to c_0$ with $c_0 = 0$ or $c_0 = \infty$, a small extension of the proof of Theorem 2 would show that the associated discrepancies would converge to 0. So, indeed, $c_n \to c_0$ and $D(\boldsymbol{p}^{(n)}) \to \ell(c_0)$.

9. Shoes instead of socks: a matching left–right pair

Suppose that instead of wanting to collect a pair of matching socks we want a pair of matching shoes. Naturally, this means one left shoe and one right shoe, both of the same color. There are two reasonable ways to extend our study to this situation.

9.1. One distribution for left colors, another distribution for right colors

The setup here involves two discrete probability distributions, say $\boldsymbol{p}$ for the color S of a left shoe and $\boldsymbol{q}$ for the color S' of a right shoe. The analog of (1) is

$$f_2 = \mathbb{P}(S = S') = \sum_i \mathbb{P}(S = S' = i) = \sum_i p_i q_i$$

for the probability that a random left shoe and a random right shoe match. We require that, for at least one value i, $p_i q_i > 0$. The analog of (2) is the method (M1) distribution for the color $X = X(\boldsymbol{p}, \boldsymbol{q})$ of a matching left–right pair:

$$\mathbb{P}(X = i) = \mathbb{P}(S = i \mid S = S') = \frac{p_i q_i}{f_2}.$$

For method (M2), we assume that at times 1, 3, 5, . . ., one left shoe is collected, and at times 2, 4, 6, . . ., one right shoe is collected. Suppose that at time $k-1$ there is not yet a matching left–right pair, but at time k, there is; then $Y = Y(\boldsymbol{p}, \boldsymbol{q})$ is the color of the shoe collected at time k.

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

(There are other sensible ways to determine the matching color under sequential collection of shoes, for example, selecting one left and one right shoe each at time 1, 2, 3, ... and breaking ties via a coin flip. Even here, choices remain. For example, if the outcome is $L_1 =$ red, $R_1 =$ blue, $L_2 =$ red, $R_2 =$ white, $L_3 =$ white, and $R_3 =$ red, then the tiebreak might be specified as equal odds for white versus red, or, since the available matches at time 3 are (L_1, R_3), (L_2, R_3), and (L_3, R_2), as 2 to 1 in favor of red over white. For this outcome, our specification in the main text is white, since the earliest match occurs at time 5, when $L_3 =$ white is observed.)

The analog of discrepancy is now

$$D(\boldsymbol{p}, \boldsymbol{q}) = d_{\mathrm{TV}}(X(\boldsymbol{p}, \boldsymbol{q}), Y(\boldsymbol{p}, \boldsymbol{q})).$$

It is fairly easy to see that, for this situation, the analog of Conjecture 1 is *false*; that is, the supremum of the discrepancy over all pairs of distributions is no smaller than the trivial upper bound on the total variation distance:

$$1 = \sup_{\boldsymbol{p}, \boldsymbol{q}} D(\boldsymbol{p}, \boldsymbol{q}). \tag{18}$$

We give a brief sketch of a proof of (18): with $a = a(n) = n^{-1/4}$ and $b = b(n) = n^{-2/3}$, let $\boldsymbol{p} = \boldsymbol{p}(n, a)$ and $\boldsymbol{q} = \boldsymbol{p}(n, b)$; in other words, $p_0 = \mathbb{P}(S = 0) = a$, $q_0 = \mathbb{P}(S' = 0) = b$, and, for $i = 1$ to n, $p_i = \mathbb{P}(S = i) = (1 - a)/n$ and $q_i = \mathbb{P}(S' = i) = (1 - b)/n$, with $a = n^{-1/4}$ and $b = n^{-2/3}$. We have $p_0 q_0 = n^{-11/12}$ and

$$\sum_{i=1}^{n} p_i q_i = n \frac{1 - a}{n} \frac{1 - b}{n} \sim \frac{1}{n} = o(p_0 q_0),$$

so the method (M1) distribution converges to a point mass at color 0, i.e. $\mathbb{P}_n(X = 0) \to 1$. To see that the method (M2) distribution has, in the limit, probability 0 of getting color 0, consider collecting alternately left and right shoes forever. At time $m = 2n^{5/8}$, we will have collected $n^{5/8}$ left and $n^{5/8}$ right shoes. Thanks to the small value $q_0 = b = n^{-2/3}$, we expect only $n^{-1/24}$ left shoes of color 0 at time m, so, with high probability, we do not yet have a matching pair of color 0. But, at time m, for *each* color $i = 1$ to n, the number of left shoes of color i is Binomial$(m, (1 - a)/n)$, and, hence, is greater than 0 with probability asymptotic to $m/n \sim n^{-3/8}$. Independently, the number of right shoes of color i is greater than 0 with probability asymptotic to $n^{-3/8}$; hence, the probability of at least one pair of color i is asymptotic to $n^{-3/4}$. The number W of colors $i > 0$ for which we have a pair has $\mathbb{E}W \sim n^{1/4}$, and the n events are negatively correlated with each other, so $\operatorname{var} W < \mathbb{E}W$. By Chebyshev's inequality, $\mathbb{P}(W = 0) \leq \operatorname{var} W/(\mathbb{E}W)^2 = O(n^{-1/4})$. So at time m we are unlikely to have any pair of color 0, and unlikely not to have at least one pair of some other color; hence, $\mathbb{P}_n(Y = 0) \to 0$.

9.2. With the constraint $\boldsymbol{p} = \boldsymbol{q}$

Now suppose that we declare that the distribution $\boldsymbol{p}$ for left shoes and the distribution $\boldsymbol{q}$ for right shoes must be equal. This does not reduce consideration of the distribution of a matching pair to the situation for socks; under the alternating left–right procedure, if we get a blue left shoe at time 1, a red right shoe at time 2, and another blue left shoe at time 3, then we still have not collected a matching pair.

The analog of Conjecture 1 for the situation of a matching left–right pair of shoes under the constraint of equal distributions is plausible.

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

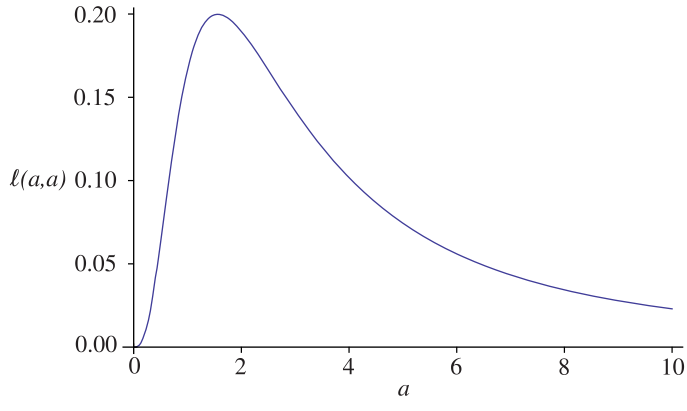

FIGURE 4: Plot of $\ell(a, a)$, the limit discrepancy $D(\boldsymbol{p}, \boldsymbol{q})$ when $\boldsymbol{p} = \boldsymbol{q} = \boldsymbol{p}(n, a/\sqrt{n})$. The maximum value 0.199 808 67 . . . occurs at $a = 1.562\,239\ldots$.

Conjecture 4. *It holds that*

$$\sup_{\boldsymbol{p}} D(\boldsymbol{p}, \boldsymbol{p}) < 1. \tag{19}$$

Furthermore, we can even propose a value for the universal constant for shoes, given by the left-hand side of (19). It comes from an analog of Theorem 2. This analog of Theorem 2 is easiest to understand without the constraint $\boldsymbol{p} = \boldsymbol{q}$.

Theorem 3. *For $a, b \in (0, \infty)$, define*

$$\ell(a, b) = \frac{ab}{1 + ab} - \int_0^\infty (a\mathrm{e}^{-at} + b\mathrm{e}^{-bt} - (a + b)\mathrm{e}^{-(a+b)t})\mathrm{e}^{-t^2}\,\mathrm{d}t.$$

For $a, b > 0$ and sufficiently large n, let

$$\boldsymbol{p}^{(n)} = \boldsymbol{p}\left(n, \frac{a}{\sqrt{n}}\right), \qquad \boldsymbol{q}^{(n)} = \boldsymbol{q}\left(n, \frac{b}{\sqrt{n}}\right), \tag{20}$$

as in (11). Then

$$\lim_{n\to\infty} D(\boldsymbol{p}^{(n)}, \boldsymbol{q}^{(n)}) = \ell(a, b).$$

Proof. The argument closely follows the proof of Theorem 2. We omit details, apart from sketching the main differences: under the distributions in (20), collecting left–right pairs with mean $1/\sqrt{n}$ holding times between pairs, the left shoes of color 0 form a rate-a Poisson process and the right shoes of color 0 form a rate-b Poisson process; $\mathbb{P}$(no left shoe of color 0 by time t) $= \mathrm{e}^{-at}$, $\mathbb{P}$(no right shoe of color 0 by time t) $= \mathrm{e}^{-bt}$, and *in the limit*, the two processes are independent, so $\mathbb{P}$(no left shoe of color 0 and no right shoe of color 0 by time t) $= \mathrm{e}^{-(a+b)t}$. Inclusion–exclusion and differentiation leads to the limit density of the time T_0 at which a left–right pair of color 0 is found, $f(t) = (a\mathrm{e}^{-at} + b\mathrm{e}^{-bt} - (a + b)\mathrm{e}^{-(a+b)t})$, instead of the $c^2 t\mathrm{e}^{-ct}$ of Theorem 2. At time t, for each of the n other colors, we expect, asymptotically, $t/\sqrt{n}$ instances on the left shoe and $t/\sqrt{n}$ instances on the right shoe, with t^2/n for the asymptotic chance of having a pair. This leads to $\mathbb{P}(\min(T_1, \ldots, T_n) > t) \to \exp(-t^2)$, instead of the $\exp(-t^2/2)$ of Theorem 2.

https://doi.org/10.1239/aap/1427814592 Published online by Cambridge University Press

While we do not have evidence for the analog of Conjecture 2—indeed, it seems daunting to deal with the analog of Section 5.2 for left–right pairs under an equal distribution for left and right—the analog of Conjecture 1 *combined with* (17) is the following plausible conjecture. See Figure 4 for the source of the constant 0.1998 . . . .

Conjecture 5. *It holds that*

$$\sup_{\boldsymbol{p}} D(\boldsymbol{p}, \boldsymbol{p}) = \max_{a} \ell(a, a) \doteq 0.199\,808\,674\,053.$$

Acknowledgement

SD was supported by a Dana and David Dornsife final-year dissertation fellowship.

References

[1] Arratia, R. and DeSalvo, S. (2011). Probabilistic divide-and-conquer: a new exact simulation method, with integer partitions as an example. Preprint. Available at http://uk.arxiv.org/abs/1110.3856v5
[2] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley, New York.
[3] Cox, D., Little, J. and O'Shea, D. (2007). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 3rd edn. Springer, New York.
[4] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd edn. John Wiley, New York.
[5] Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*, 2nd edn. John Wiley, New York.
[6] Lawler, G. F. (1995). *Introduction to Stochastic Processes*. Chapman & Hall, New York.
[7] Rudin, W. (1976). *Principles of Mathematical Analysis*, 3rd edn. McGraw-Hill, New York.