14

Amplitude amplification and estimation

Quantum amplitude amplification and estimation provide means to boost or extract the amplitude of a marked quantum state that is produced in superposition with orthogonal states by a unitary matrix. They are among the most widely used quantum primitives, providing quadratic speedups over classical algorithms in many settings.

The authors are grateful to Patrick Rall for reviewing this chapter.

14.1 Amplitude amplification

Rough overview (in words)

Given a quantum subroutine that succeeds with a probability less than one, amplitude amplification can be used to boost the success probability to 1 by making repeated calls to the subroutine and to a unitary that determines if the subroutine has succeeded. Amplitude amplification can be viewed as a generalization of Grover's search algorithm [464] and offers a quadratic speedup compared to classical methods in many instances.

Rough overview (in math)

We are given an initial state $|\psi_0\rangle$, a target ("good") state $|\psi_g\rangle$ that we can mark (i.e., the ability to reflect about the state), and a unitary U (and its inverse $U^\dagger$) such that

$$U|\psi_0\rangle = |\psi\rangle = a|\psi_g\rangle + b|\psi_b\rangle\,,$$

where $|\psi_b\rangle$ is a ("bad") state orthogonal to the target state. In other words, $|a|^2$ is the probability of success of applying U and measuring $|\psi_g\rangle$. In addition,

we are given the ability to implement the reflection operator around the initial state $R_{\psi_0} = I - 2|\psi_0\rangle\langle\psi_0|$ and an operation that, when restricted to the subspace spanned by $\{|\psi_g\rangle, |\psi_b\rangle\}$, acts as the reflection around the target state $R_{\psi_g} = I - 2|\psi_g\rangle\langle\psi_g|$.

Then, amplitude amplification allows us to boost the success probability to 1 through repeated calls to an operator $W = -UR_{\psi_0}U^\dagger R_{\psi_g}$, from the initial state $U|\psi_0\rangle = |\psi\rangle$. The standard analysis [186] proceeds by letting $a = \sin(\theta)$ and $b = \cos(\theta)$, and showing that the 2D subspace spanned by $|\psi_g\rangle, |\psi_b\rangle$ is invariant under W, which acts as a rotation operator such that $|\psi_g\rangle\langle\psi_g|W^m|\psi\rangle = \sin((2m+1)\theta)|\psi_g\rangle$.

The algorithm can also be viewed through the lens of quantum singular value transformation (QSVT) whereby U provides a generalized block-encoding (known as a projected unitary encoding) of the amplitude a. We can see this from $|\psi_g\rangle\langle\psi_g|U|\psi_0\rangle\langle\psi_0| = a|\psi_g\rangle\langle\psi_0|$. We choose to apply a polynomial $f(\cdot)$ satisfying the quantum signal processing conditions and $f(a) = 1$ to the block-encoded amplitude [429, Theorem 27 & 28]. For example, the textbook version of amplitude amplification is recovered by setting the QSVT rotation angles to $\pm\pi/2$.[1] This QSVT circuit applies a degree $2m+1$ Chebyshev polynomial of the first kind T_{2m+1} to the amplitude a, such that $|\psi_g\rangle\langle\psi_g|W^m|\psi\rangle = T_{2m+1}(a)|\psi_g\rangle = (-1)^m \sin((2m+1)\theta)|\psi_g\rangle$ for $a = \sin(\theta)$.

Dominant resource cost (gates/qubits)

The number of calls to W is

$$m = \frac{\pi}{4\arcsin(a)} - \frac{1}{2} = O\left(a^{-1}\right)$$

for small a. Each call to W requires a call to each of $U, U^\dagger, R_{\psi_0}, R_{\psi_g}$. Often we have $|\psi_0\rangle = |0^{n+k}\rangle$, and U acts on n register qubits and k ancilla qubits such that $U|0^{n+k}\rangle = a|\psi_g\rangle_n|0^k\rangle_k + b|\perp\rangle_{n,k}$, where $|\perp\rangle_{n,k}$ denotes a state orthogonal to $|0^k\rangle$ on the ancilla register. In this case the reflection operators are simple to implement using multicontrolled Toffoli gates.

Caveats

The textbook version of amplitude amplification assumes that the success amplitude a exactly equals $\sin(\pi/(4m+2))$ for an integer m. If this is not the case (e.g., when $a = 1/\sqrt{2}$), we can introduce a new qubit in $|0\rangle$ and apply an $R_y(2\phi)$ gate (i.e., a rotation about Y by angle 2ϕ) to reduce the success probability (now defined by measuring $|\psi_g\rangle|0\rangle$) to $a\cos(\phi) = \sin(\pi/(4m'+2))$ for an integer m'.

[1] These rotation angles enable a gate compilation that removes the need for the QSVT ancilla qubit.

In cases where we can only *lower bound* the success amplitude $a \geq a_0$, it is common to use fixed-point amplitude amplification [1067]. This is best understood through QSVT [429, Theorem 27], where the reflection operators are replaced by parameterized phase operators $e^{i\theta|\psi_g\rangle\langle\psi_g|}$ and $e^{i\phi|\psi_0\rangle\langle\psi_0|}$.[2] The QSVT rotation angles are chosen to implement a polynomial that maps *all* amplitudes taking value at least a_0 to at least $(1 - \epsilon)$. The fixed-point amplitude amplification circuit uses a QSVT circuit that makes $O(a_0^{-1} \log(\epsilon^{-1}))$ calls to $U, U^\dagger, e^{i\theta|\psi_g\rangle\langle\psi_g|}$, and $e^{i\phi|\psi_0\rangle\langle\psi_0|}$.

Example use cases

- Combinatorial optimization.
- Convex optimization via "minimum finding" subroutine (see [48, Appendix C]).
- Weakening cryptosystems.
- Tensor principal component analysis.
- Hamiltonian simulation using linear combinations of unitaries.

Further reading

- Both amplitude amplification and Grover search can be viewed through the lens of quantum walks on suitably constructed graphs. The quantum walks also take the form of a product of two reflections and more generally can be understood as quantizing a Markov chain describing a classical random walk [974]. We refer the interested reader to [276, 733, 52, 427].
- Oblivious amplitude amplification: Amplitude amplification can be extended to the case of oblivious amplitude amplification (OAA) [135]. The original formulation considered a setting where one is given unitary U such that for any state $|\psi\rangle$, we have

$$U|0^m\rangle|\psi\rangle = a|0^m\rangle V|\psi\rangle + b|0_\perp^m\phi\rangle$$

for a unitary operator V. The goal is to amplify the probability for the state $|0^m\rangle V|\psi\rangle$ to 1. This is achieved through $O(a^{-1})$ applications of an operator $W = U(I - 2|0^m\rangle\langle 0^m|)U^\dagger(I - 2|0^m\rangle\langle 0^m|)$ applied to $U|0^m\rangle|\psi\rangle$. We see that W does not require reflections around the initial state $|\psi\rangle$. We can recognize U as an m-qubit block-encoding of the operator aV, which can be transformed to a block-encoding of V using QSVT.[3] The OAA subroutine is used in

[2] It is shown in [687, Section 8.5] how these phase operators can be constructed using the corresponding controlled reflection operator. If only the uncontrolled reflection is available, a control can be added using, for example, [744, Fig. 5].

[3] We note that in this interpretation, one may be concerned that the phase information of the unitary V is lost by transforming the singular values. This turns out not to be problematic, as the phase information of V can be considered stored in the basis transformation matrices

the context of Hamiltonian simulation via Taylor series, where it would be problematic to have to reflect around the initial state during amplification.[4] It is also used in [297] (applied to isometries) for simulation of open quantum systems. OAA requires the block-encoded operator being amplified to preserve state norms (i.e., it must be an isometry), as this ensures that the success probability of the operation is independent of the state to which it is applied, which in turn enables amplification without reflection around the initial state.

It is also possible to amplify a block-encoding of a non-isometric operator A using QSVT; see [429, Theorem 30] and [715]. Assume $\|A\| = 1$; given the ability to implement a block-encoding U of $\sqrt{p}A$, we can use oblivious amplitude amplification to implement a block-encoding of A using $O(1/\sqrt{p})$ calls to $U, U^\dagger$. Note that for a general normalized state $|\psi\rangle$, it holds that $\||A|\psi\rangle\| \leq 1$, with equality only achieved when $|\psi\rangle$ is the singular vector corresponding to the largest singular value of A. As a result, to boost the success probability of outputting $A|\psi\rangle/\||A|\psi\rangle\|$ to unity for a general input state requires using regular amplitude amplification, involving reflections around the initial state.

- While we are unaware of a standard reference for the use of an additional ancilla qubit to account for cases where the success amplitude $a \neq \sin(\pi/(4m+2))$ for integer m, discussed above in §Caveats, it is explained more fully in [754, Appendix B].

14.2 Amplitude estimation

Rough overview (in words)

Given a quantum subroutine that succeeds with unknown success probability, amplitude estimation provides quadratic speedup over classical methods for estimating the success probability. Because many quantities of interest can be encoded in an amplitude or probability, amplitude estimation can be used as a widely applicable tool for obtaining Monte Carlo estimates with complexity $O(1/\epsilon)$, instead of the $O(1/\epsilon^2)$ achieved by classical estimation.

present in the singular value decomposition, rather than in the diagonal singular values matrix. This is taken care of automatically using QSVT. Phases are preserved when using an odd polynomial.

[4] More precisely, a robust version of OAA is used which is applicable to an operator that is ϵ close to being unitary [137, 136].

Rough overview (in math)

We are given an initial state $|\psi_0\rangle$, a target ("good") state $|\psi_g\rangle$, and a unitary U and its inverse $U^\dagger$ such that

$$U|\psi_0\rangle = |\psi\rangle = a|\psi_g\rangle + b|\psi_b\rangle\,,$$

where $|\psi_b\rangle$ is a ("bad") state orthogonal to the target state. We assume that we can mark the target state $|\psi_g\rangle$ (i.e., the ability to reflect about the state). Thus, $p = |a|^2$ is the success probability of applying U and measuring $|\psi_g\rangle$.[5] We are given the ability to implement the reflection operator around the initial state $R_{\psi_0} = I - 2|\psi_0\rangle\langle\psi_0|$ and an operation that, when restricted to the subspace spanned by $\{|\psi_g\rangle, |\psi_b\rangle\}$, acts as the reflection around the target state $R_{\psi_g} = I - 2|\psi_g\rangle\langle\psi_g|$. We can then estimate the success probability by performing quantum phase estimation on an operator $W = -UR_{\psi_0}U^\dagger R_{\psi_g}$, from the initial state $U|\psi_0\rangle = |\psi\rangle$. The standard analysis [186] proceeds by letting $|a| = \sin(\theta)$ and $|b| = \cos(\theta)$ (thus, the phases of a and b are absorbed into $|\psi_g\rangle$ and $|\psi_b\rangle$ and are not determined by the following procedure) and showing that the 2D subspace spanned by $\{|\psi_g\rangle, |\psi_b\rangle\}$ is invariant under W, where it acts as a rotation operator

$$W = \begin{pmatrix} \cos(2\theta) & \sin(2\theta) \\ -\sin(2\theta) & \cos(2\theta) \end{pmatrix}.$$

This operator has eigenvalues $e^{\pm 2i\theta}$, and we can estimate θ to additive error ϵ through quantum phase estimation. The estimate for θ can be converted into an estimate for $|a|$, or for the success probability $p = |a|^2$, which is often the quantity of interest.

Dominant resource cost (gates/qubits)

The classical approach for learning the probability p to precision ϵ has complexity scaling as $M = O(1/\epsilon^2)$, where the basic idea is to perform M incoherent repetitions of applying U and measuring in the $|\psi_g\rangle, |\psi_b\rangle$ basis, and then tally the measurement outcomes and construct the frequentist (or maximum likelihood) estimate of p. Amplitude estimation provides a quadratic speedup, learning the probability (and amplitude) with complexity scaling as $M = O(1/\epsilon)$. The textbook variant has a constant success probability, which can be boosted to $1 - \delta$ with $O(\log(1/\delta))$ overhead through standard methods (e.g., probability amplification by majority voting).

[5] Note that the original paper introducing amplitude estimation [186] uses the variable a to denote the success probability. While the algorithm is referred to as amplitude estimation, it is often the success probability that we wish to compute, and the complexity of the algorithm is often presented accordingly.

More precisely, following the analysis of [186] one can see that to learn $|a|$ to error ϵ it suffices to utilize M controlled applications of the walk operator W where M satisfies[6]

$$\epsilon \geq \frac{\pi\sqrt{1-|a|^2}}{M} + \frac{|a|\pi^2}{2M^2}\,. \tag{14.1}$$

The algorithm succeeds with probability at least $8/\pi^2$. For $|a| \approx 1 - O(\epsilon)$, a further quadratic improvement is obtained (i.e., $M = O(1/\sqrt{\epsilon})$ suffices).

To learn the success probability $p = |a|^2$ to error ϵ it suffices to utilize M controlled applications of the walk operator W where M satisfies [186]

$$\epsilon \geq \frac{2\pi\sqrt{p(1-p)}}{M} + \frac{\pi^2}{M^2}\,. \tag{14.2}$$

The algorithm once again succeeds with probability at least $8/\pi^2$. Similar to above, if $p \approx O(\epsilon)$ or $p \approx 1 - O(\epsilon)$, then it suffices to take $M = O(1/\sqrt{\epsilon})$.[7]

The overall gate complexity of an application involving amplitude estimation is given by M times the gate complexity of implementing a controlled application of W.

A common setting is the case where $|\psi_0\rangle = |0^{n+k}\rangle$, and U acts on n register qubits and k ancilla qubits such that $U|0^{n+k}\rangle = a|\psi_g\rangle|0^k\rangle_k + b|\psi_b\rangle|0^k_\perp\rangle_k$. In this case, the reflection operators are simple to implement, and W can be controlled by making these reflections controlled (adding another control qubit to a multicontrolled Z gate). We require $\log(M)$ ancilla qubits for phase estimation (which can be reduced using modern variants, see below and [855]).

Caveats

The textbook version of amplitude estimation described above produces biased estimates of $|a|$ and p. This is partly inherited from the biased nature of textbook quantum phase estimation. However, even if unbiased variants of phase estimation are used, the amplitude and probability estimates are not immediately unbiased, as they are obtained by applying nonlinear functions to the

[6] Specifically, Lemma 7 of [186] shows that if $\theta = \arcsin(|a|)$ and $\tilde{\theta} = \arcsin(|\tilde{a}|)$, then $|\theta - \tilde{\theta}| \leq \eta$ implies $|a^2 - \tilde{a}^2| \leq 2\eta\sqrt{a^2(1-a^2)} + \eta^2$. This is easily adapted to show that it also implies $|a - \tilde{a}| \leq \eta\sqrt{1-a^2} + a\eta^2/2$. They show that with probability at least $8/\pi^2$, θ is learned up to additive error at most $\eta = \pi/M$ with M calls to W, which together with the above expressions implies Eqs. (14.1) and (14.2).

[7] We can compare to the classical approach of estimating p by flipping a p-biased coin M times. Letting $\tilde{p}$ denote the estimate, which has mean p and variance $p(1-p)/M$, Chebyshev's inequality implies that $|p - \tilde{p}| \leq \epsilon$ with probability at least $8/\pi^2$ as long as $M \geq Cp(1-p)/\epsilon^2$ where $C = 1/(1 - 8/\pi^2)$. Thus, when $p \approx O(\epsilon)$ or $p \approx 1 - O(\epsilon)$, the classical approach achieves $M \sim 1/\epsilon$, and the quantum speedup is never more than quadratic.

estimate of the phase. Unbiased variants of amplitude [855][8] and probability estimation [49, 308] have been developed to address this.

The variant of amplitude estimation described above is also "destructive" in the sense that the output state is collapsed into a state $\frac{1}{\sqrt{2}}(|\psi_g\rangle \pm i|\psi_b\rangle) \neq |\psi_0\rangle, |\psi\rangle$. A nondestructive variant may be desired if the initial state is expensive to prepare and we require coherent or incoherent repetitions of amplitude estimation. Nondestructive variants have been developed in [499, 308, 855].

Example use cases

- Approximate counting of solutions marked by an oracle (e.g., topological data analysis, combinatorial optimization).
- Amplitude estimation provides a quadratic speedup for Monte Carlo estimation [773, 642] with uses in pricing financial assets. The general idea is to prepare a state $|\psi\rangle = \sum_x \sqrt{p(x)f(x)}|x\rangle|0\rangle + |\phi 0^\perp\rangle$ where $\mathbb{E}[f(x)] = \sum_x p(x)f(x)$ represents the expectation value we wish to evaluate using Monte Carlo sampling and corresponds to the probability that we measure the second register in state $|0\rangle$. Hence, amplitude estimation provides a quadratic speedup for estimating this quantity.
- A special case of amplitude estimation is overlap estimation [637], where given two states $|\psi\rangle, |\psi_0\rangle$ and a unitary such that $|\psi\rangle = U|\psi_0\rangle$, the goal is to measure $\langle\psi_0|U|\psi_0\rangle = \langle\psi_0|\psi\rangle$. This can be viewed as an application of amplitude amplification, where $|\psi_g\rangle = |\psi_0\rangle$. As a result, we only require the ability to implement $R_{\psi_0} = I - 2|\psi_0\rangle\langle\psi_0|$, $U, U^\dagger$ (or equivalently R_{ψ_0} and R_ψ). Note that in overlap estimation, one additionally wants to determine the phase of a, which can be obtained by applying amplitude estimation on a controlled variant of U, as outlined in [637]. Overlap estimation can be used for estimating observables, for example, in quantum chemistry.
- A generalization of amplitude estimation, via the quantum gradient algorithm, forms a core subroutine in some approaches for quantum state tomography [49]. Pure state tomography can be thought of as a generalization of amplitude estimation, in which we seek to learn all amplitudes individually, rather than only a single aggregate quantity. Closely related work on multivariate amplitude estimation [310] has broad applicability, including in convex optimization [51] and finance [361].

[8] In order to achieve bias $\leq \epsilon\eta$, the algorithm of [855] pays a multiplicative cost overhead $\sim \frac{1}{\eta}$ which, up to logarithmic factors, could also be achieved by merely improving the precision to $\epsilon\eta$. The additive $\sim \log(\frac{1}{\epsilon\eta})$ cost overhead of [49, 308] is much more satisfactory.

Further reading

- Variants of amplitude estimation using fewer ancilla qubits (including ancilla-free approaches), or with depth-repetition tradeoffs have been proposed [460], including work to make these methods nonadaptive [1007]. For a summary of these approaches and their unification within the QSVT framework, see [855].
- There has been some work on computing, optimizing, and comparing the constant prefactor of the $M = O(1/\epsilon)$ relation using different approaches to amplitude estimation, relevant for concrete resource estimates. For example, building on the analysis of [460], the method from [855] was estimated to scale roughly as $M \approx 4.7/\epsilon$ based on numerical experiments on a range of choices for ϵ and with fixed $a = 0.5$. This was observed to be about an order of magnitude better than the textbook method from [186] described above.[9] The method from [657] furthermore showed that a comparable total query complexity could be obtained while parallelizing across multiple processors, with maximum query depth roughly $0.4/\epsilon$.

[9] Asymptotically speaking, the complexity of the methods from [855, 460] scales suboptimally, as $O(\log\log(1/\epsilon)/\epsilon)$, but the extra $\log\log(1/\epsilon)$ factor grows sufficiently slowly that for practical values of ϵ it can be bounded by a small constant.