A Sublinear-Time Quantum Algorithm for Approximating Partition Functions

Arjan Cornelissen*† Yassine Hamoudi‡

Last update: January 7, 2024

Abstract

We present a novel quantum algorithm for estimating Gibbs partition functions in *sublinear time* with respect to the logarithm of the size of the state space. This is the first speed-up of this type to be obtained over the seminal nearly-linear time algorithm of Štefankovič, Vempala and Vigoda [ŠVV09]. Our result also preserves the quadratic speed-up in precision and spectral gap achieved in previous work by exploiting the properties of quantum Markov chains. As an application, we obtain new polynomial improvements over the best-known algorithms for computing the partition function of the Ising model, counting the number of k-colorings, matchings or independent sets of a graph, and estimating the volume of a convex body.

Our approach relies on developing new variants of the quantum phase and amplitude estimation algorithms that return nearly *unbiased* estimates with *low variance* and *without destroying* their initial quantum state. We extend these subroutines into a nearly unbiased quantum mean estimator that reduces the variance quadratically faster than the classical empirical mean. No such estimator was known to exist prior to our work. These properties, which are of general interest, lead to better convergence guarantees within the paradigm of simulated annealing for computing partition functions.

1 Introduction

The Boltzmann–Gibbs distribution is a paradigmatic tool for modeling systems that obey the principle of maximum entropy. It arises in several fields of research such as statistical mechanics [Geo11; FV17; Sin82], economic modeling [DY00], image processing [GG84; Bes86], statistical learning theory [Cat04], etc. The probability assigned by the Gibbs distribution to each possible configuration of a system is inversely proportional to the exponential of its energy multiplied by the inverse temperature. Mathematically, for a classical Hamiltonian $H : \Omega \to \{0, \dots, n\}$ of degree n specifying the energy level of each configuration x, the Gibbs distribution at inverse temperature β is given by $\pi_\beta(x) = \frac{1}{Z(\beta)} e^{-\beta H(x)}$ where the normalization factor

$$Z(\beta) = \sum_{x \in \Omega} e^{-\beta H(x)} \tag{1}$$

is called the *partition function*. While it is often straightforward to evaluate the partition function at high temperature (when $\beta = 0$, it is just the number of possible configurations), the low-temperature regime captures ground state properties that are challenging to compute. For instance, $Z(\infty)$ can represent the cardinalities of exponentially large combinatorial structures (such as the number of colorings of a graph [Jer95; ŠVV09], the volume of convex bodies [DFK91;

[*] QuSoft, University of Amsterdam. arjan.cornelissen@cwi.nl

[†] IBM Quantum, IBM T.J. Watson Research Center. arjan.cornelissen@ibm.com

[‡] Simons Institute for the Theory of Computing, University of California, Berkeley. ys.hamoudi@gmail.com

DF91] or the permanent of non-negative matrices [JSV04]) which are generally #P-hard to compute exactly [Val79; DF88; JS93].

The standard approach for evaluating partition functions at low temperature is to resort to Markov chain Monte Carlo methods [JS96]. A celebrated line of works [VC72; JVV86; DF91; BŠVV08; ŠVV09] has shown how to turn the ability to efficiently *sample* from the Gibbs distribution into that to efficiently *approximate* the partition function. At a high level, these works rely on the same two-stage simulated annealing algorithm. First, they compute a short *cooling schedule*, which is an increasing sequence of inverse temperature $0 = \beta_0 < \cdots < \beta_\ell = \infty$ with limited fluctuations in Gibbs distributions between two consecutive values. Next, the partition function at low temperature is expressed as a telescoping product

$$Z(\infty) = Z(0) \prod_{i=0}^{\ell-1} \frac{Z(\beta_{i+1})}{Z(\beta_i)} \tag{2}$$

which is approximated by using a suitable *product estimator*. As an example, the seminal algorithm of Štefankovič, Vempala and Vigoda [ŠVV09] generates a so-called "Chebyshev cooling schedule" of length[1] $\ell = \widetilde{O}(\sqrt{\log|\Omega|})$ (ignoring logarithmic dependences on the degree n, which is often on the order of $n \sim \log|\Omega|$) where each ratio $Z(\beta_{i+1})/Z(\beta_i)$ is expressed as the expectation value of a random variable X_i with bounded relative second moment $\mathbb{E}[X_i^2] \leq O(\mathbb{E}[X_i]^2)$. Such schedules are known to admit a product estimator that requires $O(\ell^2/\epsilon^2)$ classical Gibbs samples to estimate $Z(\infty) = Z(0) \cdot \mathbb{E}[X_1] \cdots \mathbb{E}[X_\ell]$ with relative error ϵ. Thus, if we let δ denote the spectral gap of a (ergodic reversible) Markov chain generating samples from the considered Gibbs distributions, the overall cost of the algorithm presented in [ŠVV09] is $\widetilde{O}(\log|\Omega|/(\epsilon^2\delta))$.

The theory of quantum algorithms provides several directions for accelerating the computation of partition functions. Quantum Markov chains [Sze04; MNRS11] can prepare coherent "qsample" encodings $|\pi_\beta\rangle = \sum_x \sqrt{\pi_\beta(x)}|x\rangle$ of the Gibbs distribution with a quadratic improvement in spectral gap for the rate of convergence (but an increase dependence on other parameters). Quantum phase estimation [Kit95] and amplitude estimation [BHMT02] lead to quadratically better convergence rates for estimating expectation values [Ter99; AW99; Hei02; BDGT11; Mon15; HM19; Ham21]. Yet, while this may hint at the existence of an $\widetilde{O}(\sqrt{\log|\Omega|}/(\epsilon\sqrt{\delta}))$ quantum algorithm for estimating partition functions, the best known algorithms [HW20; AHN+22] still require a linear scaling $\widetilde{O}(\log|\Omega|/(\epsilon\sqrt{\delta}))$ with the logarithm of the size of the state space. This bottleneck is due to additional challenges posed by current quantum algorithmic techniques. It is for instance significantly harder to prepare the qsample $|\pi_\beta\rangle$ (at low temperature) than to implement the reflection $\mathbb{I} - 2|\pi_\beta\rangle\langle\pi_\beta|$ through it. This obstacle requires using *nondestructive* procedures [MW05; WA08; WCNA09; TOV+11; ORR13; HW20] to recycle the same qsamples all along the algorithm, and to rely mostly on the reflection operator. Another fundamental limitation faced by current best quantum mean estimators [Mon15; HM19; Ham21] is the presence of *biases* in the estimates that degrade the convergence guarantee of the product estimators.

1.1 Contributions

Our main contribution is to develop the first quantum algorithm for approximating Gibbs partition functions with a complexity scaling *sublinearly* with respect to the logarithm of the size of the state space. More precisely, we prove the next theorem in Section 4.

Theorem 4.1 (Informal)**.** *There is a quantum algorithm such that, given a Gibbs distribution generated by a Markov chain with spectral gap δ, it computes an estimate $\widetilde{Z}$ of the partition function at zero temperature satisfying $|\widetilde{Z} - Z(\infty)| \leq \epsilon Z(\infty)$ by using $\widetilde{O}\big(\log^{3/4}(|\Omega|)\log^{3/2}(n)/(\epsilon\sqrt{\delta})\big)$ steps of the quantum walk operator.*

[1] We use the notation $\widetilde{O}(.)$ to hide polylogarithmic factors in the argument.

Our result reduces the polynomial dependence on $\log|\Omega|$ by a factor of $1/4$ and it achieves state-of-the-art dependence on the spectral gap δ and the accuracy ϵ up to logarithmic factors. We provide a comparison with prior work in Table 1.

	Schedule generation	Mean-value estimation	**Total cost**
[DF91; BŠVV08]	0 (non-adaptive)	$\widetilde{O}\big(\log^2\lvert\Omega\rvert/(\epsilon^2\delta)\big)$	$\widetilde{O}\big(\log^2\lvert\Omega\rvert/(\epsilon^2\delta)\big)$
[ŠVV09; Hub15; Kol18]	$\widetilde{O}\big(\log\lvert\Omega\rvert/\delta\big)$	$\widetilde{O}\big(\log\lvert\Omega\rvert/(\epsilon^2\delta)\big)$	$\widetilde{O}\big(\log\lvert\Omega\rvert/(\epsilon^2\delta)\big)$
[WCNA09]	Use [BŠVV08]	$\widetilde{O}\big(\log^2\lvert\Omega\rvert/(\epsilon\sqrt{\delta})\big)$	$\widetilde{O}\big(\log^2\lvert\Omega\rvert/(\epsilon\sqrt{\delta})\big)$
[Mon15]	Use [ŠVV09]	$\widetilde{O}\big(\log\lvert\Omega\rvert/(\epsilon\sqrt{\delta})\big)$	$\widetilde{O}\big(\log\lvert\Omega\rvert(1/\delta + 1/\epsilon\sqrt{\delta})\big)$
[HW20; AHN+22]	$\widetilde{O}\big(\sqrt{\log\lvert\Omega\rvert/\delta}\big)$	Use [Mon15]	$\widetilde{O}\big(\log\lvert\Omega\rvert/(\epsilon\sqrt{\delta})\big)$
Our work	Use [HW20; AHN+22]	$\widetilde{O}\big(\log^{3/4}\lvert\Omega\rvert/(\epsilon\sqrt{\delta})\big)$	$\widetilde{O}\big(\log^{3/4}\lvert\Omega\rvert/(\epsilon\sqrt{\delta})\big)$

Table 1: Comparison of the complexity (in terms of Markov chain steps) needed to compute partition functions over a state space Ω, where δ is the spectral gap of the Markov chain and ϵ is the accuracy parameter. Here, we omit the polylogarithmic dependencies on the degree n of the partition function since in most applications $n \sim \log|\Omega|, \mathrm{poly}(1/\delta)$. The first two rows are the classical algorithms.

The main technical ingredients of our work are new variants of the quantum phase and amplitude estimation algorithms (Theorems 2.2 and 2.4) that return (nearly) unbiased estimates and restore the qsamples used in the process with high probability. We extend these properties into new quantum mean estimators described in Section 3. In particular, the next result is crucial in the development of our improved product estimator (Theorem 3.3), and may be of independent interest. It shows that the relative error ϵ in bias can be decreased *exponentially* quickly. In comparison, the estimate $\widetilde{\mu}$ computed in previous works [Mon15; HM19; Ham21] satisfies $|\widetilde{\mu} - \mu| \leq O(\sigma/t)$ with high probability, which incurs a worst-case bias of roughly $O(\sigma/t)$.

Proposition 3.1 (Informal)**.** *For any parameters $t, \epsilon, \widetilde{\sigma}$ there is a quantum algorithm such that, given a random variable X with variance at most $\widetilde{\sigma}$, it computes a mean estimate $\widetilde{\mu}$ with bias $|\mathbb{E}[\widetilde{\mu}] - \mu| \leq \epsilon\widetilde{\sigma}$ and variance $\mathrm{Var}[\widetilde{\mu}] \leq \left(\frac{\widetilde{\sigma}}{t}\right)^2$ by using one copy of a qsample $|\pi_X\rangle$ (restored at the end of the computation) and $\widetilde{O}(t\log(1/\epsilon)^2)$ applications of the reflection $\mathbb{I} - 2|\pi_X\rangle\langle\pi_X|$.*

Finally, we provide applications of our work in Section 4.2, where we improved upon the best known algorithms for several approximate counting problems that can be phrased as partition functions. These results are described succinctly in Table 2.

	Colorings, Ising model, Independent sets	Matchings	Volume of convex body
[ŠVV09; CLV21]	$\widetilde{O}(\lvert V\rvert^2/\epsilon^2)$	$\widetilde{O}(\lvert V\rvert\lvert E\rvert/\epsilon^2)$	–
[CV18; JLLV21]	–	–	$\widetilde{O}(d^{3.5} + d^3/\epsilon^2)$
[HW20; AHN+22]	$\widetilde{O}(\lvert V\rvert^{1.5}/\epsilon)$	$\widetilde{O}(\lvert V\rvert\lvert E\rvert^{0.5}/\epsilon)$	–
[CCH+19]	–	–	$\widetilde{O}(d^3 + d^{2.5}/\epsilon)$
Our work	$\widetilde{O}(\lvert V\rvert^{1.25}/\epsilon)$	$\widetilde{O}(\lvert V\rvert^{0.75}\lvert E\rvert^{0.5}/\epsilon)$	$\widetilde{O}(d^3 + d^{2.25}/\epsilon)$

Table 2: Comparison of the best known classical (rows 1–2) and quantum (rows 3–5) algorithms for approximate counting problems. Here, ϵ is the accuracy parameter, V is the vertex set, E is the edge set and d is the dimension of the convex body. The complexities are expressed in terms of number of random or quantum walk steps (for volume estimation, each step requires one query to a membership oracle).

1.2 Proof overview

We give a high-level description of the algorithms developed in this paper. We first explain the main differences with previous works [WCNA09; Mon15; HW20; AHN+22] that allow us to obtain a sublinear algorithm for estimating partition functions. Next, we present new variants of *phase estimation*, *amplitude estimation* and *quantum mean estimation* that are needed for implementing our approach.

Estimating partition functions (Section 4). The current bottleneck in quantizations of the classical simulated annealing algorithm of Štefankovič, Vempala and Vigoda [ŠVV09] lies in estimating the expectation of a Chebyshev cooling schedule for the telescoping product displayed in Equation (2). Abstractly, this problem amounts to estimating with relative error ϵ the expectation of the product $X = X_1 \cdots X_\ell$ of ℓ independent random variables with bounded relative variance $\mathrm{Var}[X_i] \leq O(\mathbb{E}[X_i]^2)$. By Bernoulli's inequality, one can show that estimating the expectation of each inner random variable X_i with relative error $O(\epsilon/\ell)$ and taking the product of the estimates is sufficient for that purpose. Montanaro [Mon15] developed a quantum algorithm for computing each such inner estimate in time $\widetilde{O}(\ell/\epsilon)$, quadratically faster than it is possible classically, leading to an overall complexity of $\widetilde{O}(\ell^2/\epsilon)$. Nevertheless, a different *classical* estimator from Dyer and Frieze [DF91] provides the same quadratic complexity with respect to the schedule length: replace the ϵ/ℓ-error inner estimates with *unbiased* estimates of variance $O((\epsilon^2/\ell) \cdot \mathbb{E}[X_i]^2)$. The latter estimates can simply be obtained by averaging $O(\ell/\epsilon^2)$ samples due to the bounded variance property. The overall variance of this new product estimator is shown to be $O((\epsilon\mathbb{E}[X])^2)$ which, by Chebyshev's inequality and the fact that it is unbiased, implies that the overall relative error is ϵ. We follow a similar approach in our work to speed up the estimation of $\mathbb{E}[X]$. Our main contribution is a new quantum estimator that reduces the variance quadratically faster than classically, while keeping the estimates (nearly) unbiased. This allows us to compute each inner estimate in time $\widetilde{O}(\sqrt{\ell}/\epsilon)$, leading to an overall complexity of $\widetilde{O}(\ell^{3/2}/\epsilon)$. Our estimator requires new unbiased variants of the *phase* and *amplitude estimation* algorithms, which are sketched in the next paragraphs. Additionally, these algorithms are made *nondestructive* (i.e. they restore a copy of their starting state) to ensure the reusability of the *qsamples* which are an expensive resource in our applications.

Unbiased phase estimation (Section 2.1). The *phase estimation* problem asks the question of estimating the eigenphase $\theta \in [0,1)$ of a quantum state $|\psi\rangle$ satisfying $U|\psi\rangle = e^{2\pi i\theta}|\psi\rangle$ for a given unitary operator U. Kitaev's algorithm [Kit95] returns an estimate $\widetilde{\theta}$ within distance $1/t$ from θ with constant success probability by using $O(t)$ controlled-U operations. By standard success amplification techniques, the variance of this estimate can be guaranteed to be of the order of $1/t^2 + \epsilon$ at the cost of an $O(\log 1/\epsilon)$ overhead in complexity. One simple idea to make this estimator unbiased is to consider the *phase-shifted* unitary $U_\varphi|\psi\rangle = e^{2\pi i(\theta+\varphi)}|\psi\rangle$ for a random φ, run phase estimation on it and subtract the shift φ from the estimate. Although it provides an unbiased estimate of $e^{2\pi i\theta}$ (and, therefore, of $\sin^2(\theta)$), we do not know how to reduce its variance without introducing a new bias. We overcome this difficulty by exploiting more properties of the output distribution of phase estimation and looking for an unbiased estimate of θ directly. First, we show that it is possible to compute with high probability the *exact* value of the first $\log t$ bits of $\theta + \varphi = 0.b_1 b_2 \ldots$ for a well-chosen phase shift φ (steps 1–5 in Algorithm 1). This relies on the ability to pinpoint exactly the interval $\left[\frac{k}{t}, \frac{k+1}{t}\right)$ for $k \in \{0, \ldots, t-1\}$ containing the phase $\theta + \varphi$, as long as the latter is not too close to an interval endpoint (which is avoided by the choice of φ). Secondly, we look for a low-variance unbiased estimate of the remaining bits $0.0^{\log t} b_{\log t+1} b_{\log t+2} \ldots$ (steps 6–7 in Algorithm 1). For that, we consider the exponentiated unitary U_φ^t which amplifies the eigenphase to $\lambda = 0.b_{\log t+1} b_{\log t+2} \ldots$. Using *phase-to-amplitude* conversion techniques, we transform it into a new unitary V that encodes λ into the amplitude of a new auxiliary quantum register. This register can simply be measured to obtain an unbiased binary estimate $\widetilde{\lambda} \in \{-1, 1\}$ of λ. The variance of the correction term $\widetilde{\lambda}/t$ added to the previously computed bits $0.b_1 b_2 \ldots b_{\log t}$ is thus of order $O(1/t^2)$. We also have to be careful that λ is bounded away from 0 and 1 in order for the conversion techniques to succeed. This is again guaranteed by carefully choosing the phase shift. The final algorithm is summarized in Theorem 2.2.

Nondestructive amplitude estimation (Appendix B). The *amplitude estimation* algorithm [BHMT02] approximates the measurement probability $p = \|\Pi|\psi\rangle\|^2$ for a state $|\psi\rangle$ and a projector Π, given access to the reflection operators $R_\psi = \mathbb{I} - 2|\psi\rangle\langle\psi|$ and $R_\Pi = \mathbb{I} - 2\Pi$. This algorithm relies on the observation that the Grover operator $G = -R_\psi R_\Pi$ has two conjugate eigenvalues $e^{\pm 2i \arcsin(\sqrt{p})}$ with $|\psi\rangle$ being a superposition over the two corresponding eigenvectors. In its original form [BHMT02], the algorithm consists of running phase estimation on the latter state and applying the function $x \mapsto \sin^2(\pi x)$ on the measured register. Although the post-measurement state is still a superposition over the two above eigenvectors, this process can introduce a relative phase that makes it far from the initial state $|\psi\rangle$. Harrow and Wei [HW20] argued that amplitude estimation can be made *nondestructive* (i.e. it restores the state $|\psi\rangle$) by using an *uncomputation* trick inspired by the Marriott-Watrous scheme [MW05]. We discuss this result in Appendix B and we provide a new proof that avoids some complications in the analysis.

Nondestructive unbiased amplitude estimation (Section 2.2). We aim at combining the above two properties in order to obtain a variant of amplitude estimation that is both nondestructive and unbiased. Unfortunately, the properties of the Grover operator do not allow us to simply use the unbiased phase estimation on it. Indeed, it would require the ability to shift its two eigenphases by *opposite* values simultaneously. Moreover, it would return an unbiased estimate of $\arcsin(\sqrt{p})$ instead of p. We overcome these two obstacles by first applying the function $x \mapsto \sin^2(\pi x)$ on the eigenphases of the Grover operator (via amplitude-to-phase conversion techniques [GAW17]), which produces a new unitary G' with "merged" eigenvalue $e^{2\pi i p}$ and eigenvector $|\psi\rangle$. We next apply the unbiased phase estimation on G' and $|\psi\rangle$, which is inherently nondestructive since the latter state is now an eigenvector. Yet, this unbiased estimator poses a new problem: its variance is worse than that of the original amplitude estimation algorithm (roughly, $1/t^2$ vs. $p/t^2 + 1/t^4$), which is due to encoding p instead of $\arcsin(\sqrt{p})$ into the phase. Our solution to this problem is to first compute a rough *biased* amplitude estimate $q = \Theta(\max(p, 1/t^2))$ (via standard nondestructive amplitude estimation) and to apply the new function $x \mapsto \sin^2(\pi x)/q$ on the eigenphases of the Grover operator (via linear amplitude amplification [Low17]). Since the obtained operator will require time $O(1/\sqrt{q})$ to be simulated, we are left with $O(\sqrt{q}t)$ steps available to perform unbiased phase estimation on it. The variance is then shown to be of the order of p/t^2 when $p = O(1/t^2)$ and q/t^2 otherwise. A side tool of independent interest (needed when $p = O(1/t^2)$) is a *nondestructive coin flipping* algorithm (Proposition 2.3) for sampling from the Bernoulli variable of parameter $\|\Pi|\psi\rangle\|^2$ in constant expected time.

Nondestructive unbiased mean estimator (Section 3). Our final step is to generalize the amplitude estimation algorithm to estimate the expectation of an arbitrary random variable X whose distribution is encoded into a qsample $|\pi_X\rangle = \sum_x \sqrt{\Pr[X = x]}|x\rangle$ (amplitude estimation handles the case of Bernoulli distributions). The estimator shall meet the three requirements of being unbiased, nondestructive and low-variance in order to be applied efficiently to the partition function problems. We build upon the quantum estimators from [Hei02; Mon15; HM19; Ham21] that already satisfy the low-variance property. In the latter works, the estimation of $\mathbb{E}[X]$ is reduced to estimating the means of a series of Bernoulli distributions whose parameters are determined by certain *quantiles* of the input distribution. We follow the same approach, by instead using the nondestructive unbiased amplitude estimation algorithm in the reduction. This is not enough yet to make the estimator nondestructive and unbiased. First, we modify a recentering step in [Mon15; Ham21] that computed a rough mean estimate by averaging classical samples (obtained from destructive measurements of copies of $|\pi_X\rangle$). We develop a new *median estimator* for this purpose that needs not destroy any copy of $|\pi_X\rangle$ (Proposition 3.2). Secondly, the existing estimators require *truncating* the random variable X by replacing its outcomes larger than a certain threshold value with zero. We suppress most of the truncation-induced bias by adding extra Bernoulli distributions to the reduction. The sum of these added estimates has low variance and is equal (in expectation) to most of the bias introduced by previous approaches. The final estimator is summarized in Proposition 3.1.

2 Unbiased and nondestructive quantum subroutines

2.1 Phase estimation

In this section, we describe a variant of phase estimation that returns a (nearly) unbiased estimate. We use the next notation for separating the most and least significant bits of a phase $\theta \in [0,1]$.

Definition 2.1. Let $\theta \in [0,1]$ be a real number with binary expansion $\theta = 0.\theta_1\theta_2\theta_3\ldots$ where $\theta_j \in \{0,1\}$. Then, for any integer $\tau \geq 0$, we denote $\theta_{\leq\tau} = 0.\theta_1\theta_2\ldots\theta_\tau$ and $\theta_{>\tau} = 0.\theta_{\tau+1}\theta_{\tau+2}\ldots$ such that $\theta = \theta_{\leq\tau} + 2^{-\tau}\theta_{>\tau}$.

Suppose that we can call a controlled unitary U, and that we have a copy of an eigenstate $|\psi\rangle$, satisfying $U|\psi\rangle = e^{2\pi i\theta}|\psi\rangle$. Then, we can implement the *phase-shift* operation $e^{2\pi i\varphi}U|\psi\rangle = e^{2\pi i(\theta+\varphi)}|\psi\rangle$ for any φ with one call to U, and the *bit-shift* operation $U^{2^\tau}|\psi\rangle = e^{2\pi i\theta_{>\tau}}|\psi\rangle$ with 2^τ calls to U. We make use of these two operations to design the following unbiased, nondestructive, low-variance phase estimation algorithm.

Algorithm 1. Unbiased low-variance phase estimation, $\mathsf{Ub\text{-}Phase}(U, |\psi\rangle, t, \epsilon)$.

```
1: for $k = 0, \ldots, 8$ do
2:     Set $\varphi = k/(16t')$ and $U_\varphi = e^{2\pi i\varphi}U$ where $t' = 8t$.
3:     Run phase estimation $N = \lceil 200\log(72t'/\epsilon)\rceil$ times on the phase-shifted unitary $U_\varphi$
       with initial state $|\psi\rangle$ and time $t'$ to collect $N$ independent estimates $\widetilde{\varphi}_1, \ldots, \widetilde{\varphi}_N$ of $\varphi$.
       Compute the frequencies $(\widetilde{p}(i))_{i\in\{0,\ldots,t'-1\}}$ where $\widetilde{p}(i) = |\{j \in \{1,\ldots,N\} : \widetilde{\varphi}_j = i/t'\}|$.
4:     if there is an index $i$ such that $\widetilde{p}(i), \widetilde{p}(i+1) \geq 0.17$ and $(i/t')_{>\tau} \in (4/7, 5/7)$ then
5:         Stop the for loop and set $\widetilde{\theta}_\varphi = (i/t')_{\leq\tau}$.
6: Apply phase-to-amplitude conversion on the exponentiated unitary $U_\varphi^t$ (for the last value
   of $\varphi$ computed in step 2) with precision $\epsilon/4$ to obtain a unitary $V$ such that
       $$V\big(|\psi\rangle|0\rangle^{\otimes a}\big) = |\psi\rangle\big(\sqrt{1-p'}|0\rangle^{\otimes a} + \sqrt{p'}|\Phi^\perp\rangle\big).$$
7: Compute the state $V\big(|\psi\rangle|0\rangle^{\otimes a}\big)$ and measure its last $a$ qubits in the standard basis. If the
   result is the all-zero string then set $b = 0$ else set $b = 1/(2\pi)$.
8: Output $\widetilde{\theta} = \widetilde{\theta}_\varphi + b/t - \varphi$.
```

Theorem 2.2 (Unbiased low-variance phase estimation)**.** *Let U be a unitary operator with an eigenvector $|\psi\rangle$ such that $U|\psi\rangle = e^{2\pi i\theta}|\psi\rangle$ where $\theta \in [0,1/2]$. Given $t = 2^\tau \geq 1$ and $\epsilon \in (0,1)$, the* unbiased low-variance phase estimation *algorithm* $\mathsf{Ub\text{-}Phase}(U, |\psi\rangle, t, \epsilon)$ *(Algorithm 1) outputs an estimate $\widetilde{\theta} \in [-1,1]$ such that*

$$\left|\mathbb{E}[\widetilde{\theta}] - \theta\right| \leq \epsilon \quad \text{and} \quad \mathrm{Var}[\widetilde{\theta}] \leq \frac{1}{t^2} + \epsilon.$$

The algorithm needs one copy of $|\psi\rangle$, which is restored at the end of the computation with probability at least $1-\epsilon$, and $O(t\log(t/\epsilon))$ controlled-U operations.

Proof. We decompose the analysis of Algorithm 1 into three parts. First, we show that the two phases $\varphi \in (0, 1/t)$ and $\widetilde{\theta}_\varphi \in (0,1)$ obtained at the end of steps 1–5 satisfy the equations

$$\widetilde{\theta}_\varphi = (\theta+\varphi)_{\leq\tau} \qquad \text{and} \qquad (\theta+\varphi)_{>\tau} \in (1/2, 3/4) \tag{3}$$

with probability at least $1-\epsilon/8$. We refer to this part as *exact phase estimation* since $\widetilde{\theta}_\varphi$ is exactly equal to the first τ bits of the shifted phase $\theta+\varphi$. Next, conditioned on any pair $(\varphi, \widetilde{\theta}_\varphi)$ satisfying the above equation, we prove that the estimate b obtained at the end of steps 6–7 satisfies

$$|\mathbb{E}[b \mid (\varphi, \widetilde{\theta}_\varphi)] - (\theta+\varphi)_{>\tau}| \leq \epsilon/2 \qquad \text{and} \qquad \mathrm{Var}[b \mid (\varphi, \widetilde{\theta}_\varphi)] \leq 1. \tag{4}$$

We refer to this part as *unbiased phase estimation.* Finally, we analyze the expectation and variance of the final estimate constructed at step 8.

Exact phase estimation (steps 1–5). For each phase shift φ and integer $i \in \{0, \ldots, t'-1\}$, let ϵ_i^φ denote the distance between the angles $\theta+\varphi$ and i/t' along the unit circle. The two smallest distances are achieved by $\epsilon_{i^\star}^\varphi, \epsilon_{i^\star+1}^\varphi \leq 1/t'$ where $i^\star = t'(\theta+\varphi)_{\leq 3\tau}$, whereas the other indices $j \notin \{i^\star, i^\star+1\}$ are at distance at least $\epsilon_j^\varphi > 1/t'$. By Hoeffding's inequality and a union bound over the t' possible indices, at each round of step 3 the empirical frequency vector $\widetilde{p}$ is at distance $\|p-\widetilde{p}\|_\infty \leq 0.05$ from the probability vector p with probability at least $1-\epsilon/72$. Since step 3 is executed at most 9 times, all of the frequency vectors computed by the algorithm satisfy this bound with probability at least $1-\epsilon/8$. On the other hand, there is at least one value of φ (among the 9 possible ones) such that $\epsilon_{i^\star}^\varphi, \epsilon_{i^\star+1}^\varphi \leq 5/(8t')$ and $(\theta+\varphi)_{>\tau} \in (1/2, 3/4)$. The algorithm will correctly detect such a value because of the gaps proved in Corollary A.2 between the probabilities $p(i^\star), p(i^\star+1)$ and the probabilities $p(j)$ for $j \notin \{i^\star, i^\star+1\}$. This proves Equation (3).

Unbiased phase estimation (steps 6–7). The state $|\psi\rangle$ is an eigenvector of U_φ^t such that $U_\varphi^t|\psi\rangle = e^{2\pi i(\theta+\varphi)_{>\tau}}|\psi\rangle$. If the pair $(\varphi, \widetilde{\theta}_\varphi)$ obtained at the end of steps 1–6 satisfies Equation (3) then, by Lemma A.5, the amplitude $\sqrt{p'}$ obtained at step 6 is at distance at most $|\sqrt{p'} - \sqrt{2\pi(\theta+\varphi)_{>\tau}}| \leq \epsilon/4$ from the shifted phase. Thus, conditioned on any fixed correct pair $(\varphi, \widetilde{\theta}_\varphi)$, we get $|\mathbb{E}[b \,|\, (\varphi, \widetilde{\theta}_\varphi)] - (\theta+\varphi)_{>\tau}| = |p' - (\theta+\varphi)_{>\tau}| \leq \epsilon/2$ and $\mathrm{Var}[b \,|\, (\varphi, \widetilde{\theta}_\varphi)] \leq 1$ for the estimate b computed at step 7. This proves Equation (4).

Low-variance (step 8). The expectation and variance of the final estimate $\widetilde{\theta}$ are $|\mathbb{E}[\widetilde{\theta} \,|\, (\varphi, \widetilde{\theta}_\varphi)] - \theta| \leq \epsilon/2$ and $\mathrm{Var}[\widetilde{\theta} \,|\, (\varphi, \widetilde{\theta}_\varphi)] \leq 1/t^2$ conditioned on any pair $(\varphi, \widetilde{\theta}_\varphi)$ satisfying Equation (3). Since $|\theta| \leq 1$ and Equation (3) holds with probability at least $1-\epsilon/8$, we obtain by the law of total expectation that $|\mathbb{E}[\widetilde{\theta}] - \theta| \leq \epsilon/2 + \max|\widetilde{\theta} - \theta| \cdot \epsilon/8 \leq 3\epsilon/4$ and by the law of total variance that $\mathrm{Var}[\widetilde{\theta}] = \mathbb{E}[\mathrm{Var}[\widetilde{\theta} \mid (\varphi, \widetilde{\theta}_\varphi)]] + \mathrm{Var}[\mathbb{E}[\widetilde{\theta} \mid (\varphi, \widetilde{\theta}_\varphi)]] \leq 1/t^2 + \epsilon/8 + \mathbb{E}[|\mathbb{E}[\widetilde{\theta} \mid (\varphi, \widetilde{\theta}_\varphi)] - \theta|^2] \leq 1/t^2 + \epsilon/8 + (\epsilon/2)^2 + 4 \cdot \epsilon/8 \leq 1/t^2 + 7\epsilon/8$. The number of controlled-U operations used by the algorithm is $O(t\log(t/\epsilon))$ at steps 1–5 and $O(t\log(1/\epsilon))$ at steps 6–7. Moreover, the state $|\psi\rangle$ is preserved by the phase estimation and the phase-to-amplitude conversion algorithms. □

2.2 Amplitude estimation

In this section, we describe a variant of amplitude estimation that returns a (nearly) unbiased estimate and restores the initial state with high probability. We start with a simple procedure, adapted from the Marriott-Watrous scheme [MW05], to sample nondestructively from a Bernoulli variable of parameter $\|\Pi|\psi\rangle\|^2$ given a projector Π and a single copy of a state $|\psi\rangle$.

Algorithm 2. Nondestructive coin flip

```
Measure the state $|\psi\rangle$ according to the projectors $\{\Pi, \mathbb{I}-\Pi\}$. Set $b=1$ if the result is $\Pi|\psi\rangle$
  and $b=0$ if it is $(\mathbb{I}-\Pi)|\psi\rangle$.
Measure the residual state according to the projectors $\{|\psi\rangle\langle\psi|, \mathbb{I}-|\psi\rangle\langle\psi|\}$.
while the residual state is not $|\psi\rangle$ do
    Measure the residual state according to the projectors $\{\Pi, \mathbb{I}-\Pi\}$.
    Measure the residual state according to the projectors $\{|\psi\rangle\langle\psi|, \mathbb{I}-|\psi\rangle\langle\psi|\}$.
Output $b$ and $|\psi\rangle$.
```

Proposition 2.3 (Nondestructive coin flip). *Let $|\psi\rangle$ be a quantum state and Π be a projection operator. The* nondestructive coin flip *algorithm (Algorithm 2) outputs a Boolean value $b \in \{0,1\}$ such that $\mathbb{E}[b] = \|\Pi|\psi\rangle\|^2$. The algorithm needs one copy of $|\psi\rangle$, which is restored at the end of the computation, and $O(1)$ applications in expectation of the reflection operators $\mathbb{I} - 2|\psi\rangle\langle\psi|$ and $\mathbb{I} - 2\Pi$.*

Proof. Let $p = \|\Pi|\psi\rangle\|^2$ denote the probability to be estimated and define $r = 2p(1-p)$. By the standard Marriott-Watrous analysis [MW05], the residual state at the end of step 2 is $|\psi\rangle$ with probability $1-r$ (in which case steps 3–5 need not be executed) and the probability that one iteration of steps 4–5 succeeds in restoring $|\psi\rangle$ is r. Thus, the expected number of iterations needed to restore $|\psi\rangle$ is $r^2 \sum_{i=0}^{\infty}(i+1)(1-r)^i = 1$. □

We now extend the above algorithm to a low-variance unbiased estimator.

Algorithm 3. Nondestructive unbiased amplitude estimation, $\mathsf{NdUb\text{-}Amplitude}(|\psi\rangle, \Pi, t, \epsilon)$.

```
1: Run nondestructive amplitude estimation $\mathsf{Nd\text{-}Amplitude}(|\psi\rangle, \Pi, 2t, \epsilon/8)$ on $|\psi\rangle$ with projector $\Pi$ to obtain a (biased) amplitude estimate $q$ of $p = \|\Pi|\psi\rangle\|^2$.
2: Let $\tau = \max\{1, \frac{1}{4}\min\{t, 1/\sqrt{q}\}\}$. Run linear amplitude amplification on $|\psi\rangle$ with projector $\Pi$, time $\tau$ and precision $\epsilon/40$ to obtain a unitary $V_{\tau,\epsilon/40}$ preparing a state $|\psi'\rangle = V_{\tau,\epsilon/40}|\psi\rangle$ such that
       $\big|\|\Pi|\psi'\rangle\| - \tau\|\Pi|\psi\rangle\|\big| \leq \epsilon/40.$
3: if $q \leq 1/t^2$ then
4:     Measure the state $|\psi'\rangle$ according to the projectors $\{\Pi, \mathbb{I}-\Pi\}$ by using the nondestructive coin flip algorithm. Let $b \in \{0,1\}$ denote the Boolean value it returns.
5:     Output $\widetilde{p} = b/\tau^2$.
6: else
7:     Run amplitude-to-phase conversion on the unitary $V_{\tau,\epsilon/40}$ with precision $\epsilon' = \Omega\big(\frac{\tau^2}{t^2\log(t/(\epsilon\tau))}\big)$ to obtain a unitary $U_{\epsilon'}$ such that
           $\|U_{\epsilon'}\big(|\psi'\rangle|0\rangle^{\otimes a}\big) - e^{i\|\Pi|\psi'\rangle\|^2}|\psi'\rangle|0\rangle^{\otimes a}\| \leq \epsilon'.$
8:     Run the unbiased phase estimation algorithm $\mathsf{Ub\text{-}Phase}(U_{\epsilon'}, |\psi'\rangle|0\rangle^{\otimes a}, 7t/\tau, \epsilon/90)$ on $U_{\epsilon'}$ with (approximate) eigenstate $|\psi'\rangle|0\rangle^{\otimes a}$ to compute an eigenphase estimate $\widetilde{p}'$.
9:     Output $\widetilde{p} = 2\pi\widetilde{p}'/\tau^2$.
```

Theorem 2.4 (Nondestructive unbiased amplitude estimation). *Let $|\psi\rangle$ be a quantum state and Π be a projection operator with $p = \|\Pi|\psi\rangle\|^2$. Given $t \geq 4$ and $\epsilon \in (0,1)$, the* nondestructive unbiased amplitude estimation *algorithm* $\mathsf{NdUb\text{-}Amplitude}(|\psi\rangle, \Pi, t, \epsilon)$ *(Algorithm 3) outputs an estimate $\widetilde{p} \in [-2\pi, 2\pi]$ such that*

$$|\mathbb{E}[\widetilde{p}] - p| \leq \epsilon \quad \textit{and} \quad \mathrm{Var}[\widetilde{p}] \leq \frac{91p}{t^2} + \epsilon.$$

The algorithm needs one copy of $|\psi\rangle$, which is restored at the end of the computation with probability at least $1-\epsilon$, and $O(t\log\log(t)\log(t/\epsilon))$ applications in expectation of the reflection operators $\mathbb{I} - 2|\psi\rangle\langle\psi|$ and $\mathbb{I} - 2\Pi$.

Proof. By a simple application of Proposition B.2, with probability at least $1-\epsilon/8$, the amplitude estimate q computed at step 1 satisfies

$$\widetilde{q} = 0 \text{ if } p \leq 1/(4t^2), \qquad q \leq 7p \text{ if } p \geq 1/(4t^2), \qquad q \geq p/4 \text{ if } q > 1/t^2.$$

We analyze the rest of the algorithm conditioned on the above inequalities being true. Similarly to the proof of Theorem 2.2, by the laws of total expectation and total variance, when removing the condition the expectation and variance of the estimate $\widetilde{p}$ will differ respectively by at most $\epsilon/4$ and $7\epsilon/8$ from the values computed below.

The amplification parameter τ at step 2 satisfies $\tau\|\Pi|\psi\rangle\| = \tau\sqrt{p} \leq 1/2$ when $\tau > 1$, as required by Lemma A.3. Thus, the norm of the post-amplification state $|\psi'\rangle$ projected along Π is

$$\big|\|\Pi|\psi'\rangle\|^2 - \tau^2 p\big| = \big|\|\Pi|\psi'\rangle\| + \tau\|\Pi|\psi\rangle\|\big| \cdot \big|\|\Pi|\psi'\rangle\| - \tau\|\Pi|\psi\rangle\|\big| \leq \epsilon/20. \tag{5}$$

We split the analysis of the subsequent steps of Algorithm 3 into two cases depending on whether the biased estimate q is smaller or larger than $1/t^2$.

If $q \leq 1/t^2$ (steps 3–5). We have $\tau = t/4$ and, by Proposition 2.3, the Boolean value b at step 4 is sampled from a Bernoulli distribution of parameter $\mathbb{E}[b] = \|\Pi|\psi'\rangle\|^2$. Thus, the expectation of the output estimate satisfies $|\mathbb{E}[\widetilde{p}] - p| = \big|(\|\Pi|\psi'\rangle\|/\tau)^2 - p\big| \leq \epsilon/(20\tau^2) \leq \epsilon/2$ and its variance is $\mathrm{Var}[\widetilde{p}] \leq \mathbb{E}[\widetilde{p}^2] = \|\Pi|\psi'\rangle\|^2/\tau^4 \leq (\tau^2 p + \epsilon/20)/\tau^4 \leq p/\tau^2 + \epsilon/(20\tau^4) \leq 16p/t^2 + \epsilon/8$.

If $q > 1/t^2$ (steps 6–9). We have $\tau = \max\{1, 1/(4\sqrt{q})\}$. We start by analysing the estimate obtained when the unitary $U_{\epsilon'}$ used in phase estimation (step 8) is replaced with a unitary U that perfectly satisfies

$$U\big(|\psi'\rangle|0\rangle^{\otimes a}\big) = e^{i\|\Pi|\psi'\rangle\|^2}|\psi'\rangle|0\rangle^{\otimes a}$$

and is at distance at most $\|U - U_{\epsilon'}\| \leq \epsilon'$ from $U_{\epsilon'}$ (such a unitary exists since $\|U_{\epsilon'}\big(|\psi'\rangle|0\rangle^{\otimes a}\big) - e^{i\|\Pi|\psi'\rangle\|^2}|\psi'\rangle|0\rangle^{\otimes a}\| \leq \epsilon'$ by Lemma A.4). We rename the phase and output estimates computed at steps 8 and 9 to $\widehat{p}'$ and $\widehat{p}$ respectively in this case. Observe that the eigenphase $\|\Pi|\psi'\rangle\|^2/(2\pi)$ satisfies the boundedness requirement of Theorem 2.2 since it is at most $(\tau^2 p + \epsilon/20)/(2\pi) \leq 1/2$ by Equation (5) and $\tau = \max\{1, 1/(4\sqrt{q})\} \leq \max\{1, 1/(2\sqrt{p})\}$. Thus, the expectation of the estimate $\widehat{p}'$ computed at step 8 satisfies $\big|\mathbb{E}[2\pi\widehat{p}'] - \tau^2 p\big| \leq \big|\mathbb{E}[2\pi\widehat{p}'] - \|\Pi|\psi'\rangle\|^2\big| + \big|\|\Pi|\psi'\rangle\|^2 - \tau^2 p\big| \leq 2\pi\epsilon/90 + \epsilon/20$, where the first step is by the triangle inequality and the second step is by Theorem 2.2 and Equation (5). The variance is at most $\mathrm{Var}[\widehat{p}'] \leq \tau^2/(7t)^2 + \epsilon/90$ by Theorem 2.2. Since the final estimate is the scalar multiple $\widehat{p} = 2\pi\widehat{p}'/\tau^2$ of $\widehat{p}'$, its expectation and variance are $|\mathbb{E}[\widehat{p}] - p| \leq 2\pi(2\pi\epsilon/90 + \epsilon/20)/\tau^2 \leq \epsilon/4$ and $\mathrm{Var}[\widehat{p}] \leq 4\pi^2/(7t\tau)^2 + 4\pi^2\epsilon/(90\tau^4) \leq 91p/t^2 + \epsilon/9$.

Consider now the case where U is replaced with the actual unitary U_ϵ. Let T denote the number of times the operator U_ϵ is used by the unbiased phase estimation procedure at step 8. By Theorem 2.2, we have $T = O((t/\tau)\log(t/(\epsilon\tau)))$. By standard approximation arguments (see [NC11, p.194] for instance), the distribution of the actual estimate $\widetilde{p}$ is at distance at most $2T\|U - U_{\epsilon'}\|$ (in ℓ_∞-norm) from that of $\widehat{p}$. Moreover, by a simple analysis of Algorithm 1, both estimates are distributed in $[-1, 1]$ and can take at most $36\pi t/\tau$ different values. We conclude that $|\mathbb{E}[\widetilde{p}] - p| \leq |\mathbb{E}[\widehat{p}] - p| + 2T\|U - U_{\epsilon'}\| \cdot 36\pi t/\tau \leq \epsilon/4 + 2T\epsilon' \cdot 36\pi t/\tau \leq \epsilon/2$ and $\mathrm{Var}[\widetilde{p}] \leq \mathrm{Var}[\widehat{p}] + 8T\|U - U_{\epsilon'}\| \cdot 36\pi t/\tau \leq 91p/t^2 + \epsilon/9 + 8T\epsilon' \cdot 36\pi t/\tau \leq 91p/t^2 + \epsilon/8$ by choosing a value $\epsilon' = \Theta\big(\frac{\tau^2}{t^2\log(t/(\epsilon\tau))}\big)$. □

3 Mean estimation

We construct first a quantum estimator that decreases the variance quadratically faster than classically while keeping the estimate (nearly) unbiased. Moreover, it requires a single qsample, which is not destroyed with high probability. The main ingredient is the nondestructive amplitude estimation algorithm constructed in the previous section. The estimator is based on expressing the mean as a linear combination of carefully chosen Bernoulli random variables (with parameters $\widetilde{\mu}_j^+$ and $\widetilde{\mu}_j^-$ defined at steps 4 and 7 of Algorithm 4) which are related to different truncations of the input random variable. We use a more sophisticated truncation pattern than in previous work [Hei02; Mon15; Ham21; CHJ22] to ensure that the estimate is nearly unbiased. Our estimator requires knowing a rough estimate $\widetilde{m}$ of the mean and an upper-bound $\widetilde{\sigma}$ on the variance of the random variable. We explain how to find $\widetilde{m}$ nondestructively in Proposition 3.2, whereas $\widetilde{\sigma}$ will be provided directly by our applications.

Algorithm 4. Unbiased mean estimator, $\mathsf{Unbiased}(X, t, \widetilde{m}, \widetilde{\sigma}, \epsilon)$.

```
1: Define the shifted random variable $\bar{X} = X - \widetilde{m}$.
2: Set $a_{-1} = 0$ and $a_j = 2^j \widetilde{\sigma}$ for $j \in \{0, \dots, k\}$, where $k = \lceil \log(580/\epsilon) \rceil$.
3: for $j = 0$ to $k$ do
4:     Use a controlled rotation to transform $|\pi_X\rangle$ into the state
           $|\pi_j^+\rangle = \sum_{x \notin (a_{j-1}, a_j]} \sqrt{\pi_{\bar{X}}(x)}|x\rangle|0\rangle + \sum_{x \in (a_{j-1}, a_j]} \sqrt{\pi_{\bar{X}}(x)}|x\rangle \left( \sqrt{1 - \frac{x}{a_j}}|0\rangle + \sqrt{\frac{x}{a_j}}|1\rangle \right)$
       and denote $\mu_j^+ = \mathbb{E}\left[\frac{\bar{X}}{a_j} \mathbb{1}_{\bar{X} \in (a_{j-1}, a_j]}\right] = \|(\mathbb{I} \otimes |1\rangle\langle 1|)|\pi_j^+\rangle\|^2$.
5:     Compute an estimate $\widetilde{\mu}_j^+$ of $\mu_j^+$ by using the nondestructive unbiased amplitude
       estimation algorithm $\mathsf{NdUb\text{-}Amplitude}(|\pi_j^+\rangle, \mathbb{I} \otimes |1\rangle\langle 1|, 300t, \epsilon')$ where $\epsilon' = \left(\frac{\epsilon}{2320t}\right)^2$.
6:     Restore $|\pi_j^+\rangle$ to $|\pi_X\rangle$ by undoing the controlled rotation.
7:     Repeat steps 4–6 with the state
           $|\pi_j^-\rangle = \sum_{-x \notin (a_{j-1}, a_j]} \sqrt{\pi_{\bar{X}}(x)}|x\rangle|0\rangle + \sum_{-x \in (a_{j-1}, a_j]} \sqrt{\pi_{\bar{X}}(x)}|x\rangle \left( \sqrt{1 - \frac{|x|}{a_j}}|0\rangle + \sqrt{\frac{|x|}{a_j}}|1\rangle \right)$
       to obtain an estimate $\widetilde{\mu}_j^-$ of $\mu_j^- = \mathbb{E}\left[\frac{|\bar{X}|}{a_j} \mathbb{1}_{\bar{X} \in [-a_j, -a_{j-1})}\right]$.
8: Output $\widetilde{\mu} = \widetilde{m} + \sum_{j=0}^k a_j (\widetilde{\mu}_j^+ - \widetilde{\mu}_j^-)$.
```

Proposition 3.1 (UNBIASED MEAN ESTIMATOR). *Let X be a finite random variable with mean μ and variance σ^2 encoded as a qsample $|\pi_X\rangle$. Given two reals $\epsilon \in (0, 1)$ and $t \geq 1$ and two estimates $\widetilde{m}, \widetilde{\sigma}$ such that $|\mu - \widetilde{m}| \leq 17\sigma$ and $\widetilde{\sigma} \geq \sigma$, the* unbiased mean estimator $\mathsf{Unbiased}(X, t, \widetilde{m}, \widetilde{\sigma}, \epsilon)$ *(Algorithm 4) outputs a mean estimate $\widetilde{\mu}$ such that*

$$|\mathbb{E}[\widetilde{\mu}] - \mu| \leq \epsilon\widetilde{\sigma} \quad \textit{and} \quad \mathrm{Var}[\widetilde{\mu}] \leq \left(\frac{\widetilde{\sigma}}{t}\right)^2.$$

The algorithm needs one copy of $|\pi_X\rangle$, which is restored at the end of the computation with probability at least $1 - \epsilon$, and $\widetilde{O}(t \log(1/\epsilon)^2)$ applications of the reflection $\mathbb{I} - 2|\pi_X\rangle\langle\pi_X|$ in expectation.

Proof. We start by giving several properties of the shifted random variable $\bar{X} = X - \widetilde{m}$. First, its second moment is at most

$$\mathbb{E}[\bar{X}^2] = \mathbb{E}[(X - \mu)^2] + (\mu - \widetilde{m})^2 \leq 290\sigma^2 \tag{6}$$

since, by assumption, the input estimate $\widetilde{m}$ satisfies $|\mu - \widetilde{m}| \leq 17\sigma$. Next, the expectation of $|\bar{X}|$ over any interval $(a_{j-1}, a_j]$ defined in Algorithm 4 when $j > 0$ is at most $\mathbb{E}\big[|\bar{X}|\mathbb{1}_{|\bar{X}| \in (a_{j-1}, a_j]}\big] \leq \mathbb{E}\big[\bar{X}^2 \mathbb{1}_{|\bar{X}| \in (a_{j-1}, a_j]}\big]/a_{j-1}$ because $a_{j-1} > 0$. Hence, the normalized means μ_j^+ and μ_j^- satisfy

$$\mu_j^+ + \mu_j^- \leq \frac{2}{a_j^2}\mathbb{E}\big[\bar{X}^2 \mathbb{1}_{|\bar{X}| \in (a_{j-1}, a_j]}\big] \qquad \text{when } j > 0. \tag{7}$$

Lastly, the expectation of $|\bar{X}|$ over the interval $[a_k, +\infty)$ is at most

$$\mathbb{E}[|\bar{X}|\mathbb{1}_{|\bar{X}|} \geq a_k] \leq \frac{\mathbb{E}[\bar{X}^2]}{a_k} \leq \frac{290\sigma}{2^k} \tag{8}$$

where we used Equation (6) and the assumption that $\widetilde{\sigma} \geq \sigma$.

We now analyze the expectation and variance of the final estimate $\widetilde{\mu}$. Each of the $\widetilde{\mu}_j^+$ and $\widetilde{\mu}_j^-$ estimates introduces a bias of at most ϵ' by Theorem 2.4 and there is another bias due to the truncated part $\mathbb{E}[\bar{X}\mathbb{1}_{|\bar{X}|} \geq a_k]$ analyzed in Equation (8). Overall, we have $|\mathbb{E}[\widetilde{\mu}] - \mu| \leq \sum_{j=0}^k a_j \epsilon' + \frac{290\sigma}{2^k} \leq (2^{k+1}\epsilon' + \frac{290}{2^k})\widetilde{\sigma} \leq \epsilon\widetilde{\sigma}$. Since the estimates $\widetilde{\mu}_j$ are independent random variables, the variance is $\mathrm{Var}[\widetilde{\mu}] = \sum_{j=0}^k a_j^2 \mathrm{Var}[\widetilde{\mu}_j]$. Thus, by Theorem 2.4 and Equation (7), $\mathrm{Var}[\widetilde{\mu}] \leq \sum_{j=0}^k a_j^2 \left(\frac{\mu_j^+ + \mu_j^-}{(30t)^2} + \epsilon'\right) \leq a_0^2\left(\frac{2}{(30t)^2} + \epsilon'\right) + \sum_{j=1}^k \left(\frac{2\mathbb{E}\left[\bar{X}^2 \mathbb{1}_{|\bar{X}| \in (a_{j-1}, a_j]}\right]}{(30t)^2} + a_j^2 \epsilon'\right) \leq \frac{2\widetilde{\sigma}^2 + 2\mathbb{E}\left[\bar{X}^2\right]}{(30t)^2} + 2^{2k+1}\widetilde{\sigma}^2\epsilon' \leq \left(\frac{582}{(30t)^2} + 2^{2k+1}\epsilon'\right)\widetilde{\sigma}^2 \leq \left(\frac{\widetilde{\sigma}}{t}\right)^2$.

By Theorem 2.4, the algorithm uses $\widetilde{O}\big(kT\log(1/\epsilon')\big) = \widetilde{O}(t\log(1/\epsilon)^2)$ reflections through $|\pi_X\rangle$ and it preserves the copy of $|\pi_X\rangle$ with probability at least $1 - 2(k+1)\epsilon' \geq 1 - \epsilon$. □

We now explain how to find an estimate $\widetilde{m}$ that satisfies the first requirement $|\mu - \widetilde{m}| \leq 17\sigma$ needed to apply the above estimator. If we had access to classical samples then it would suffice to compute the average of $O(\log 1/\eta)$ samples to get the above inequality with probability $1 - \eta$. Nevertheless, we do not know how to extract a classical sample from a qsample $|\pi_X\rangle$ without destroying it. Since the mean is always within one standard deviation of the median, we can instead attempt to estimate the median. There exist quantum algorithms [NW99; Nay99; Ham21] for computing the quantiles of a distribution, however they also require measuring qsamples. We follow a new approach that incurs a logarithmic overhead in the size of the support but is nondestructive. Our algorithm consists of performing a binary search over the support of the random variable, where we test if a value x is close to the median by estimating the tail probability $\Pr[X \geq x]$ with the nondestructive amplitude estimation algorithm.

Algorithm 5. Nondestructive median estimator, $\mathsf{Median}(X, \eta)$.

```
1: Let $x_1 < \cdots < x_n$ denote the values in the support of $X$.
2: Set $a = 1$ and $b = n + 1$. Run the following binary search until $a = b$:
3:     Set $k = \lfloor \frac{a+b}{2} \rfloor$.
4:     Compute an estimate $\widetilde{p}_k$ of $p_k = \Pr[X \geq x_k]$ by running the nondestructive amplitude
       estimation algorithm $\mathsf{Nd\text{-}Amplitude}(|\pi_k\rangle, \mathbb{I} \otimes |1\rangle\langle 1|, 3\sqrt{2}, \eta/\log(m))$ where

       $$|\pi_k\rangle = \left(\sum_{x < x_k} \sqrt{\pi_X(x)}|x\rangle\right)|0\rangle + \left(\sum_{x \geq x_k} \sqrt{\pi_X(x)}|x\rangle\right)|1\rangle.$$

5:     If $\widetilde{p}_k \leq 1/6$ then set $b = k$ else set $a = k + 1$.
6: Output $\widetilde{m} = x_{a-1}$.
```

Proposition 3.2 (Nondestructive median estimator). *Let X be a finite random variable with support size n, mean μ and variance σ^2. Given a real $\eta \in (0,1)$, the* nondestructive median estimator $\mathsf{Median}(X, \eta)$ *(Algorithm 5) outputs an approximate median $\widetilde{m}$ such that*

$$|\mu - \widetilde{m}| \leq 17\sigma$$

with probability at least $1 - \eta$. The algorithm needs one copy of $|\pi_X\rangle$, which is restored at the end of the computation with probability at least $1 - \eta$, and $O(\log(n)\log(\log(n)/\eta))$ applications of the reflection $\mathbb{I} - 2|\pi_X\rangle\langle\pi_X|$.

Proof. By Theorem B.2 and a union-bound argument, all the estimates $\widetilde{p}_k$ computed by Algorithm 5 satisfy $|\widetilde{p}_k - p_k| \leq \sqrt{p_k}/(3\sqrt{2}) + 1/18$ with probability at least $1 - \eta$. In particular, if $p_k \geq 1/2$ then $\widetilde{p}_k \geq p_k - \sqrt{p_k}/(3\sqrt{2}) - 1/18 \geq 5/18$ and if $p_k \leq 1/18$ then $\widetilde{p}_k \leq p_k + \sqrt{p_k}/(3\sqrt{2}) + 1/18 \leq 1/6$. Thus, the binary search stops with $a = b$ satisfying

$p_a < 1/2 \leq 9p_{a-1}$ with probability at least $1-\eta$ (note that $a \geq 2$ since $p_1 = 1$). Consequently, if we define the quantile function $Q_X(p) = \sup\{x \in \mathbb{R} : \Pr[X \geq x] \geq p\}$ for all $p \in [0,1]$, we have $Q_X(1/2) \leq \widetilde{m} \leq Q_X(1/18)$. Finally, the result $|\mu - \widetilde{m}| \leq 17\sigma$ follows from the quantile inequality $\mu - \sigma\sqrt{\frac{p}{1-p}} \leq Q_X(p) \leq \mu + \sigma\sqrt{\frac{1-p}{p}}$ proved in [BB04, Section 3.1]. □

We describe our final estimator for estimating the mean of a product of ℓ independent random variables $X_1, \ldots, X_\ell$ with relative error ϵ. We make the standard assumptions that the relative second moment $\frac{\mathbb{E}[X_i^2]}{\mathbb{E}[X_i]^2}$ and the inverse fidelity $1/|\langle \pi_{X_i} | \pi_{X_{i+1}} \rangle|^2$ are upper-bounded by some known value B, which is a constant in our applications (Section 4). Our estimator uses a similar approach to that of Dyer and Frieze [DF91] by taking the product of ℓ (nearly) unbiased estimates $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_\ell$ of $\mathbb{E}[X_1], \ldots, \mathbb{E}[X_\ell]$ respectively, each with variance $O(\epsilon^2/\ell)$. We obtain a quantum speed-up by using the quantum unbiased estimator of Proposition 3.1 to reduce the variances quadratically faster than classically.

Algorithm 6. Product estimator, $\mathsf{Product}(X_1, \ldots, X_\ell, B, \epsilon)$.

```
1: for $i = 1$ to $\ell$ do
2:     Define $\widehat{X}_i$ to be the average of $K = \max\{1, 1156(B-1)\}$ independent samples from $X_i$.
3:     Compute an estimate $\widetilde{m}_i$ of the median of $\widehat{X}_i$ by using the nondestructive median
       estimator $\mathsf{Median}(\widehat{X}_i, 1/(11\ell))$. Set $\widetilde{\sigma}_i = \widetilde{m}_i B$.
4:     Compute an estimate $\widetilde{\mu}_i$ of the expectation $\mathbb{E}[\widehat{X}_i]$ by using the unbiased mean estimator
       $\mathsf{Unbiased}(\widehat{X}_i, \frac{96B\sqrt{\ell}}{\epsilon}, \widetilde{m}_i, \widetilde{\sigma}_i, \frac{\epsilon}{6\ell B})$.
5:     Anneal the qsample $|\pi_{\widehat{X}_i}\rangle$ to $|\pi_{\widehat{X}_{i+1}}\rangle$ with probabilistic annealing.
6: Output $\widetilde{\mu} = \widetilde{\mu}_1 \cdots \widetilde{\mu}_\ell$.
```

Theorem 3.3 (Product estimator). *Let $B > 1$ and $\epsilon \in (0,1)$. Consider a sequence $X_1, \ldots, X_\ell$ of ℓ independent random variables with support size n, bounded relative second moment $\frac{\mathbb{E}[X_i^2]}{\mathbb{E}[X_i]^2} \leq B$ and bounded fidelity $|\langle \pi_{X_i} | \pi_{X_{i+1}} \rangle|^2 \geq 1/B$ for all i. Denote their product as $X = X_1 \cdots X_\ell$. Then, the* product estimator $\mathsf{Product}(X_1, \ldots, X_\ell, B, \epsilon)$ *(Algorithm 6) outputs a multiplicative-error estimate $\widetilde{\mu}$ such that*

$$\left|\widetilde{\mu} - \mathbb{E}\Big[\prod_{i=1}^{\ell} X_i\Big]\right| \leq \epsilon \mathbb{E}\Big[\prod_{i=1}^{\ell} X_i\Big]$$

with probability at least $2/3$. It uses $O(B)$ copies of $|\pi_{X_1}\rangle$ and $\widetilde{O}\big(B^2\ell^{3/2}/\epsilon + B\ell\log(n)\big)$ reflections through the states $|\pi_{X_1}\rangle, \ldots, |\pi_{X_\ell}\rangle$ in expectation.

Proof. By a union-bound argument, with probability at least $10/11$, all the median estimates $\widetilde{m}_i$ computed at step 3 are within distance $17\sqrt{\mathrm{Var}[\widehat{X}_i]}$ of the means $\mathbb{E}[\widehat{X}_i]$. We condition on this event happening for the rest of the proof.

We start by analysing steps 2–5 for a given random variable X_i. First, notice that we can construct a qsample for $\widehat{X}_i$ by taking $|\pi_{\widehat{X}_i}\rangle = U_{\mathrm{average}}(|\pi_{X_i}\rangle^{\otimes K} \otimes |0\rangle)$ where U_{average} is a unitary computing the average of the first K registers in the last register. The expectation is unchanged $\mathbb{E}[X_i] = \mathbb{E}[\widehat{X}_i]$, whereas the relative variance is at most

$$\frac{\mathrm{Var}[\widehat{X}_i]}{\mathbb{E}[X_i]^2} = \frac{\mathrm{Var}[X_i]}{K\mathbb{E}[X_i]^2} \leq \frac{B-1}{K}. \tag{9}$$

The next equation shows that the distance between the rough estimate $\widetilde{m}_i$ computed at step 3 and the expectation is bounded by a multiple of the standard deviation (by Proposition 3.2), which in turn is a small multiple of the expectation (since the relative variance is small by Equation (9)).

$$|\widetilde{m}_i - \mathbb{E}[X_i]| \leq 17\sqrt{\mathrm{Var}[\widehat{X}_i]} \leq 17\sqrt{\frac{B-1}{K}}\mathbb{E}[X_i] = \frac{\mathbb{E}[X_i]}{2} \tag{10}$$

Finally, the expectation of each estimate $\widetilde{\mu}_i$ computed at step 4 is bounded as

$$|\mathbb{E}[\widetilde{\mu}_i] - \mathbb{E}[X_i]| \leq \frac{\epsilon}{6\ell B}\widetilde{\sigma}_i \leq \frac{\epsilon}{4\ell}\mathbb{E}[X_i] \tag{11}$$

where the first inequality is by Theorem 2.4 and the second is by $\widetilde{\sigma}_i = B\widetilde{m}_i$ and Equation (10).

Consider now the product random variable $X = X_1 \cdots X_\ell$. The expectation of the final estimate $\widetilde{\mu}$ is at most $\mathbb{E}[\widetilde{\mu}] \leq (1 + \frac{\epsilon}{4\ell})^\ell \prod_{i=1}^{\ell} \mathbb{E}[\mu_i] \leq (\epsilon/2)\mathbb{E}[X]$ and at least $\mathbb{E}[\widetilde{\mu}] \geq (1 - \frac{\epsilon}{4\ell})^\ell \prod_{i=1}^{\ell} \mathbb{E}[\mu_i] \geq (\epsilon/4)\mathbb{E}[X]$ by Equation (11). Thus, $|\mathbb{E}[\widetilde{\mu}] - \mathbb{E}[X]| \leq (\epsilon/2)\mathbb{E}[X]$. The relative variance of $\widetilde{\mu}$ is at most $\frac{\mathrm{Var}[\widetilde{\mu}]}{\mathbb{E}[\widetilde{\mu}]^2} = \prod_{i=1}^{\ell}\Big(1 + \frac{\mathrm{Var}[\widetilde{\mu}_i]}{\mathbb{E}[\widetilde{\mu}_i]^2}\Big) - 1 \leq \prod_{i=1}^{\ell}\Big(1 + \frac{\widetilde{\sigma}_i^2}{(t\mathbb{E}[\widetilde{\mu}_i])^2}\Big) - 1 = \prod_{i=1}^{\ell}\Big(1 + \Big(\frac{\epsilon B\widetilde{m}_i}{96\sqrt{\ell}B\mathbb{E}[\widetilde{\mu}_i]}\Big)^2\Big) - 1 \leq \Big(1 + \frac{\epsilon^2}{32\ell}\Big)^\ell - 1 \leq \epsilon^2/16$ where the second step is by Proposition 3.1 and the fourth step uses that $\widetilde{m}_i/\mathbb{E}[\widetilde{\mu}_i] \leq 3$ by Equations (10) and (11). We conclude by using the triangle and Chebyshev inequalities that $|\widetilde{\mu} - \mathbb{E}[X]| \leq |\widetilde{\mu} - \mathbb{E}[\widetilde{\mu}]| + |\mathbb{E}[\widetilde{\mu}] - \mathbb{E}[X]| \leq (\epsilon/2 + \epsilon/2)\mathbb{E}[X]$ with probability at least $1 - 4\frac{\mathrm{Var}[\widetilde{\mu}]}{(\epsilon\mathbb{E}[\widetilde{\mu}])^2} \geq 3/4$.

The algorithm needs K copies of the initial qsample $|\pi_{X_1}\rangle$ to anneal successively into K copies of $|\pi_{X_i}\rangle$ at each stage of steps 2–5. The transition from $|\pi_{X_i}\rangle^{\otimes K}$ to $|\pi_{X_{i+1}}\rangle^{\otimes K}$ requires $O(KB)$ reflections through $|\pi_{X_i}\rangle$ and $|\pi_{X_{i+1}}\rangle$, by using Lemma A.6 on each of the K subsystems. Furthermore, at each stage, step 3 uses $\widetilde{O}(\log(n)\log(\ell))$ reflections through $|\pi_{\widehat{X}_i}\rangle$ by Proposition 3.2 and step 4 uses $\widetilde{O}(B\sqrt{\ell}/\epsilon)$ reflections through $|\pi_{\widehat{X}_i}\rangle$ by Proposition 3.1. Overall, the number of reflections through the original states $|\pi_{X_1}\rangle, \ldots, |\pi_{X_\ell}\rangle$ is thus $\widetilde{O}\big(\ell K\big(B + \log(n) + B\sqrt{\ell}/\epsilon\big)\big)$. □

4 Partition function estimation

In this section, we showcase an application of our newly-constructed quantum mean estimation algorithm. Specifically, we show how it can be used to speed up existing quantum algorithms for estimating partition functions. In Section 4.1, we elaborate on the generic algorithm for partition function estimation and how our results provide a speed up and, in Section 4.2, we discuss how this gives rise to more efficient quantum algorithms for several applications.

4.1 General algorithm for partition function estimation

In this subsection, we first describe our partition function estimation algorithm in high level, and define the required notation along the way. We follow the exposition in [HW20], but use slightly different notations to match the rest of this document more closely.

Let Ω be a state space, and let $H : \Omega \to \{0, \ldots, n\}$ be a Hamiltonian, i.e., a function associating an energy to each of the states. Our goal will be to compute the number of states whose energy is 0, i.e., $|H^{-1}(0)|$. To that end, we define the partition function $Z : [0, \infty] \to \mathbb{R}$ as

$$Z(\beta) = \sum_{x \in \Omega} e^{-\beta H(x)},$$

and we observe that $Z(0) = |\Omega|$, and $Z(\infty) = |H^{(-1)}(0)|$. The partition function arises frequently in statistical physics, where β is referred to as the inverse temperature.

Next, we define a sequence of inverse temperatures $0 = \beta_0 < \beta_1 < \cdots < \beta_\ell = \infty$, and express $Z(\infty)$ as the telescoping product

$$Z(\infty) = Z(0) \cdot \frac{Z(\beta_1)}{Z(0)} \cdot \frac{Z(\beta_2)}{Z(\beta_1)} \cdot \ldots \cdot \frac{Z(\infty)}{Z(\beta_\ell)} = Z(0) \prod_{i=0}^{\ell-1} \frac{Z(\beta_{i+1})}{Z(\beta_i)}. \tag{12}$$

Since throughout the sequence of β_i's, we are increasing the inverse temperature, and hence decreasing the temperature, this sequence is referred to as the *cooling schedule*. Its length ℓ is called the *schedule length*. The core idea for estimating $Z(\infty)$ is to evaluate each of the factors in the product on the right-hand side of Equation (12) individually.

Thus, let $i \in [\ell]$. We endow Ω with a probability distribution π_{β_i}, called the Gibbs distribution in statistical physics, and define a random variable $X_i : \Omega \to \mathbb{R}$ on it, with

$$\pi_{\beta_i}(x) = \frac{e^{-\beta_i H(x)}}{Z(\beta_i)}, \qquad \text{and} \qquad X_i(x) = e^{-(\beta_{i+1}-\beta_i)H(x)}.$$

It follows immediately that

$$\mathbb{E}[X_i] = \sum_{x\in\Omega} \frac{e^{-\beta_i H(x)}}{Z(\beta_i)} \cdot e^{-(\beta_{i+1}-\beta_i)H(x)} = \sum_{x\in\Omega} \frac{e^{-\beta_{i+1} H(x)}}{Z(\beta_i)} = \frac{Z(\beta_{i+1})}{Z(\beta_i)}.$$

The idea is now to devise a procedure that samples from the Gibbs distribution and obtains an estimator $\widetilde{\mu}_i$ of $\mathbb{E}[X_i]$.

In order to be able to sample from the Gibbs distribution, we encode it in a quantum state. To that end, we define the Gibbs state $|\pi_{\beta_i}\rangle \in \mathbb{C}^\Omega$ at inverse temperature β_i, defined as

$$|\pi_{\beta_i}\rangle = \frac{1}{\sqrt{Z(\beta_i)}} \sum_{x\in\Omega} \sqrt{e^{-\beta_i H(x)}}|x\rangle.$$

In typical applications (see the next subsection), it can be much easier to reflect through the Gibbs state than to prepare it. Therefore, we estimate $\mathbb{E}[X_i]$ using the nondestructive mean estimation algorithm constructed in the previous sections. This has the benefit of restoring all the Gibbs state needed in the computation of $\widetilde{\mu}_i$, which makes it possible to reuse them for computing the next factor in the products on the right-hand side of Equation (12).

It remains to choose the cooling schedule, i.e., the inverse temperatures $\beta_1, \dots, \beta_{\ell-1}$. For reasons that are sketched on high level, we want our cooling schedule to have the following two properties.

1. *B-Chebyshev.* For any $B \geq 1$, a cooling schedule is called B-Chebyshev if for any two subsequent β_i and β_{i+1}, we have

   $$\frac{\mathbb{E}[X_i^2]}{\mathbb{E}[X_i]^2} = \frac{Z(2\beta_{i+1}-\beta_i)Z(\beta_i)}{Z(\beta_{i+1})^2} \leq B.$$

   This requirement can be viewed as a tail bound – it tells us that X_i concentrates well around its mean $\mathbb{E}[X_i]$.

2. *B-slowly varying.* For any $B \geq 1$, a cooling schedule is called B-slowly varying if

   $$\left|\langle \pi_{\beta_i}|\pi_{\beta_{i+1}}\rangle\right|^2 = \frac{Z\left(\frac{\beta_i+\beta_{i+1}}{2}\right)^2}{Z(\beta_i)Z(\beta_{i+1})} \geq 1/B.$$

   This requirement can be seen as a proximity (or *warm start*) requirement – it tells us that the Gibbs states $|\pi_{\beta_i}\rangle$ and $|\pi_{\beta_{i+1}}\rangle$ have some non-negligible overlap. It ensures that we can *anneal*, i.e., transform, $|\pi_{\beta_i}\rangle$ into $|\pi_{\beta_{i+1}}\rangle$ without having to do too much work, and hence circumvents having to prepare $|\pi_{\beta_{i+1}}\rangle$ from scratch for estimating the next mean $\mathbb{E}[X_{i+1}]$.

When we set $B = e^2$, it is shown in [ŠVV09] that there exists a cooling schedule of length $\ell = \widetilde{O}(\sqrt{\log|\Omega|\log(n)})$, which is B-Chebyshev. Subsequently, it was shown in [HW20] that one can modify the construction so that the resulting cooling schedule is not only B-Chebyshev, but also slowly varying. Moreover, computing this cooling schedule can be done on the fly, with associated cost scaling approximately linearly in the schedule length. Combining this work with our product estimator, Theorem 3.3, gives rise to the following result.

Theorem 4.1 (Partition function estimator)**.** *Let $H : \Omega \to \{0, \ldots, n\}$ be a Hamiltonian, and let Z be its partition function. Suppose that, for every inverse temperature β, the associated Gibbs distribution π_β is the stationary distribution of an ergodic reversible Markov chain with spectral gap at least δ. Then, we can estimate $Z(\infty)$ up to multiplicative error ϵ, with probability at least $2/3$, by using*

$$\widetilde{O}\Big(\big(\log^{3/4}|\Omega| \log^{3/4}(n)/\epsilon + \sqrt{\log|\Omega|} \log^{3/2} n\big)/\sqrt{\delta}\Big)$$

steps of the quantum walk operator in expectation.

Proof. We run the algorithm from [HW20] to compute the cooling schedule on the fly. The total number of Gibbs state reflections used in this step is $\widetilde{O}(\sqrt{\log|\Omega|} \log^{3/2} n)$ [HW20, Theorem 13], and the resulting cooling schedule is both e^2-slowly varying and e^2-Chebyshev, and of length $\ell = \widetilde{O}(\sqrt{\log|\Omega| \log n})$. For computing the product in Equation (12), we use the product estimator from Theorem 3.3. The total number of Gibbs state reflections used in this algorithm is $\widetilde{O}(\log^{3/4}|\Omega| \log^{3/4}(n)/\epsilon + \sqrt{\log|\Omega|} \log^{3/2} n)$. With standard failure probability reduction techniques, we obtain that the number of Gibbs state reflections performed by the resulting algorithm scales as the sum of the two complexities. Finally, by a well-known result [Sze04; MNRS11], each reflection through a Gibbs state $|\pi_\beta\rangle$ can be implemented with $O(1/\sqrt{\delta})$ steps of the quantum walk operator corresponding to a Markov chain with spectral gap at least δ generating π_β. □

The core of our improvement lies in our product estimator – the one used in [HW20] is quadratic in the schedule length, whereas ours is subquadratic. Thus, more generally, if we are in a setting where a cooling schedule that is shorter than $\widetilde{O}(\sqrt{\log|\Omega| \log(n)})$ suffices, then our resulting algorithm scales with the schedule length to the power of $3/2$. A typical setting where this is the case is when we know a priori a lower bound on $Z(\infty)$ – then we can terminate the annealing at a lower inverse temperature, and hence obtain a shorter cooling schedule. We leave working out the details in this setting for future work.

4.2 Applications

The quantum partition function estimator can be applied to several combinatorial counting and statistical physics problems. We describe some examples that are representative of these applications and for which we obtain faster algorithms than in previous work [Mon15; CCH+19; HW20; AHN+22]. In all results that we list below, the range n of the Hamiltonian scales at most polylogarithmically in $|\Omega|$, so we can hide any dependence on $\log(n)$ in the tilde of the big-O notation.

Counting k-colorings. Let $G = (V, E)$ be a graph of maximum degree $\Delta = O(1)$, and a number $k = O(1)$ of colors such that $k > 2\Delta$. We want to count the number of ways to color the vertices of G such that no edge is monochromatic. We let Ω be the set of all colorings $x \in [k]^V$, which implies that $Z(0) = |\Omega| = k^{|V|}$, and $H(x)$ is the number of monochromatic edges in G when each vertex $v \in V$ is colored with x_v. This is also called the *Potts model* and $Z(\infty) = |H^{(-1)}(0)|$ is equal to the number of valid k-colorings of G. Jerrum [Jer95] showed that the Glauber dynamics for the Potts model mixes in time $O(|V| \log|V|)$. Thus, by Theorem 4.1, we can estimate $Z(\infty)$ up to relative error ϵ in time $\widetilde{O}\big(\log^{3/4}(|\Omega|)\sqrt{|V|}/\epsilon\big) = \widetilde{O}(|V|^{5/4}/\epsilon)$.

Ferromagnetic Ising model. Let $G = (V, E)$ be a graph of maximum degree $\Delta \geq 3$. We take Ω to be the set of all assignments $x \in \{-1, 1\}^V$ of signs ±1 to the vertices of G, which readily implies that $|\Omega| = 2^{|V|}$. The energy $H(x)$ is the number of edges in E whose two endpoints have the same sign (when the sign of $v \in V$ is x_v). Here, since the model is ferromagnetic, the Gibbs distribution is chosen to be proportional to $e^{\beta H(x)}$ and the partition function is $Z(\beta) = \sum_x e^{\beta H(x)}$ (the framework given in the previous section can easily be adapted to this setting). Mossel and Sly [MS13] showed that the Glauber dynamics for the ferromagnetic Ising model mixes in time $O(|V| \log |V|)$ in the tree uniqueness region $\beta < \ln(\Delta/(\Delta-2))$. Thus, for inverse temperatures β satisfying that condition, we can estimate $Z(\beta)$ in time $\widetilde{O}(|V|^{5/4}/\epsilon)$.

Counting matchings. Let $G = (V, E)$ be a graph of maximum degree $\Delta = O(1)$. A matching $x \subseteq E$ is a subset of disjoint edges. We let Ω be the set of all matchings, and we set $H(x) = |x|$ to be the number of edges in the matching x. This is called the *monomer-dimer model*. The state space is of size $Z(0) = |\Omega| = O(|V|!2^{|V|})$ and we have $Z(\infty) = 1$. The setting is again slightly different here since the partition function is easy to calculate at zero temperature ($\beta = \infty$) rather than at infinite temperature ($\beta = 0$). We can still run a similar algorithm as in Theorem 4.1 by annealing in the opposite direction (see [HW20; Mon15] for a more detailed explanation). Chen, Liu and Vigoda [CLV21, Theorem 1.5] showed that the Glauber dynamics for the monomer-dimer model mixes in time $O(|E| \log |V|)$. Thus, we can estimate the number $Z(0)$ of matchings in G in time $\widetilde{O}(|V|^{3/4}|E|^{1/2}/\epsilon)$.

Counting independent sets. Let $G = (V, E)$ be a graph of maximum degree $\Delta = O(1)$. An independent set $x \subseteq V$ is a subset of vertices such that no edge has its two endpoints in x. Similarly as before, we let Ω be the set of all independent sets and we define $H(x) = |x|$ to be the size (number of vertices) of the independent set x. This is also called the *hard-core model*. Again, we have $Z(\infty) = 1$ and we anneal backwards to estimate $Z(0) = |\Omega|$. Chen, Liu and Vigoda [CLV21, Theorem 1.2] showed that the Glauber dynamics for the hard-core model mixes in time $O(|V| \log |V|)$ in the tree uniqueness region $\beta > \ln\left(\frac{(\Delta-2)^\Delta}{(\Delta-1)^{\Delta-1}}\right)$. Thus, for inverse temperatures β satisfying that condition, we can estimate $Z(\beta)$ in time $\widetilde{O}(|V|^{5/4}/\epsilon)$. In particular, we can estimate the number of independent sets when the maximum degree is $\Delta \leq 5$ (since the tree uniqueness region contains the value $\beta = 0$ in that case).

Computing the volume of a convex body. We can also use our techniques to speed up the volume estimation algorithms of convex bodies from [LV06; CCH+19]. In this problem, we assume that we are given a radius $R > 0$ and a convex set $K \subseteq \mathbb{R}^d$ such that $B(1) \subseteq K \subseteq B(R)$ where $B(r)$ is the ball of radius r centered at $0 \in \mathbb{R}^d$. The goal is to compute an estimate $\widetilde{V}$ such that $(1-\epsilon)\operatorname{Vol}(K) \leq \widetilde{V} \leq (1+\epsilon)\operatorname{Vol}(K)$. The algorithm from [CCH+19] achieves this with $\widetilde{O}(d^3 + d^{2.5}/\epsilon)$ quantum queries to a membership oracle O_K to K, and we reduce the number of queries required to $\widetilde{O}(d^3 + d^{2.25}/\epsilon)$. In contrast to the other applications mentioned in this section, here we require a more significant change to the algorithm, and hence we supply the details in Appendix C.

Acknowledgements

The authors thank András Gilyén for discussions on unbiased phase estimation. Y.H. was supported by the Simons Institute and DOE NQISRC QSA grant #FP00010905.

References

[AHN+22] S. Arunachalam, V. Havlicek, G. Nannicini, K. Temme, and P. Wocjan. “Simpler (Classical) and Faster (Quantum) Algorithms for Gibbs Partition Functions”. In: *Quantum* 6 (2022), p. 789 (cit. on pp. 2, 3, 15).

[AW99] D. S. Abrams and C. P. Williams. *Fast Quantum Algorithms for Numerical Integrals and Stochastic Processes*. arXiv:quant-ph/9908083. 1999 (cit. on p. 2).

[BB04] S. C. Bagui and D. K. Bhaumik. “Glimpses of Inequalities in Probability and Statistics”. In: *International Journal of Statistical Sciences* 3 (2004), pp. 9–15 (cit. on p. 12).

[BDGT11] G. Brassard, F. Dupuis, S. Gambs, and A. Tapp. *An Optimal Quantum Algorithm to Approximate the Mean and its Application for Approximating the Median of a Set of Points over an Arbitrary Distance.* `arXiv:1106.4267 [quant-ph]`. 2011 (cit. on p. 2).

[Bes86] J. Besag. “On the Statistical Analysis of Dirty Pictures”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 48.3 (1986), pp. 259–302 (cit. on p. 1).

[BHMT02] G. Brassard, P. Høyer, M. Mosca, and A. Tapp. “Quantum Amplitude Amplification and Estimation”. In: *Contemporary Mathematics* 305 (2002), pp. 53–74 (cit. on pp. 2, 5, 20–22).

[BŠVV08] I. Bezáková, D. Štefankovič, V. V. Vazirani, and E. Vigoda. “Accelerating Simulated Annealing for the Permanent and Combinatorial Counting Problems”. In: *SIAM Journal on Computing* 37.5 (2008), pp. 1429–1454 (cit. on pp. 2, 3).

[Cat04] O. Catoni. *Statistical Learning Theory and Stochastic Optimization.* Springer Berlin Heidelberg, 2004 (cit. on p. 1).

[CCH+19] S. Chakrabarti, A. M. Childs, S.-H. Hung, T. Li, C. Wang, and X. Wu. *Quantum Algorithm for Estimating Volumes of Convex Bodies.* `arXiv:1908.03903 [quant-ph]`. 2019 (cit. on pp. 3, 15, 16, 22–24).

[CHJ22] A. Cornelissen, Y. Hamoudi, and S. Jerbi. “Near-Optimal Quantum Algorithms for Multivariate Mean Estimation”. In: *Proceedings of the 54th Symposium on Theory of Computing (STOC).* 2022, pp. 33–43 (cit. on p. 9).

[CLV21] Z. Chen, K. Liu, and E. Vigoda. “Optimal Mixing of Glauber Dynamics: Entropy Factorization via High-Dimensional Expansion”. In: *Proceedings of the 53rd Symposium on Theory of Computing (STOC).* 2021, pp. 1537–1550 (cit. on pp. 3, 16).

[CV18] B. Cousins and S. Vempala. “Gaussian Cooling and $O^*(n^3)$ Algorithms for Volume and Gaussian Volume”. In: *SIAM Journal on Computing* 47.3 (2018), pp. 1237–1273 (cit. on p. 3).

[DF88] M. E. Dyer and A. Frieze. “On the Complexity of Computing the Volume of a Polyhedron”. In: *SIAM Journal on Computing* 17.5 (1988), pp. 967–974 (cit. on p. 2).

[DF91] M. E. Dyer and A. Frieze. “Computing the Volume of Convex Bodies: A Case where Randomness Provably Helps”. In: *Proceedings of the Symposium on Probabilistic Combinatorics and Its Applications.* 1991, pp. 123–170 (cit. on pp. 2–4, 12).

[DFK91] M. Dyer, A. Frieze, and R. Kannan. “A Random Polynomial-Time Algorithm for Approximating the Volume of Convex Bodies”. In: *Journal of the ACM* 38.1 (1991), pp. 1–17 (cit. on p. 1).

[DY00] A. Dragulescu and V. M. Yakovenko. “Statistical Mechanics of Money”. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 17.4 (2000), pp. 723–729 (cit. on p. 1).

[FV17] S. Friedli and Y. Velenik. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction.* Cambridge University Press, 2017 (cit. on p. 1).

[GAW17] A. Gilyén, S. Arunachalam, and N. Wiebe. *Optimizing Quantum Optimization Algorithms via Faster Quantum Gradient Computation.* `arXiv:1711.00465 [quant-ph]`. 2017 (cit. on pp. 5, 20).

[Geo11] H.-O. Georgii. *Gibbs Measures and Phase Transitions*. De Gruyter, 2011 (cit. on p. 1).

[GG84] S. Geman and D. Geman. “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6.6 (1984), pp. 721–741 (cit. on p. 1).

[Ham21] Y. Hamoudi. “Quantum Sub-Gaussian Mean Estimator”. In: *Proceedings of the 29th European Symposium on Algorithms (ESA)*. 2021, 50:1–50:17 (cit. on pp. 2, 3, 5, 9, 11).

[Hei02] S. Heinrich. “Quantum Summation with an Application to Integration”. In: *Journal of Complexity* 18.1 (2002), pp. 1–50 (cit. on pp. 2, 5, 9).

[HM19] Y. Hamoudi and F. Magniez. “Quantum Chebyshev’s Inequality and Applications”. In: *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming (ICALP)*. 2019, 69:1–69:16 (cit. on pp. 2, 3, 5).

[Hub15] M. Huber. “Approximation Algorithms for the Normalizing Constant of Gibbs Distributions”. In: *The Annals of Applied Probability* 25.2 (2015), pp. 974–985 (cit. on p. 3).

[HW20] A. W. Harrow and A. Y. Wei. “Adaptive Quantum Simulated Annealing for Bayesian Inference and Estimating Partition Functions”. In: *Proceedings of the 31st Symposium on Discrete Algorithms (SODA)*. 2020, pp. 193–212 (cit. on pp. 2, 3, 5, 13–16, 21, 22).

[Jer95] M. Jerrum. “A Very Simple Algorithm for Estimating the Number of k-Colorings of a Low-Degree Graph”. In: *Random Structures & Algorithms* 7.2 (1995), pp. 157–165 (cit. on pp. 1, 15).

[JLLV21] H. Jia, A. Laddha, Y. T. Lee, and S. Vempala. “Reducing Isotropy and Volume to KLS: An $O^*(n^3\Psi^2)$ Volume Algorithm”. In: *Proceedings of the 53rd Symposium on Theory of Computing (STOC)*. 2021, pp. 961–974 (cit. on p. 3).

[JS93] M. Jerrum and A. Sinclair. “Polynomial-Time Approximation Algorithms for the Ising Model”. In: *SIAM Journal on Computing* 22.5 (1993), pp. 1087–1116 (cit. on p. 2).

[JS96] M. Jerrum and A. Sinclair. “The Markov Chain Monte Carlo Method: An Approach to Approximate Counting and Integration”. In: *Approximation Algorithms for NP-Hard Problems*. PWS Publishing, 1996, pp. 482–520 (cit. on p. 2).

[JSV04] M. Jerrum, A. Sinclair, and E. Vigoda. “A Polynomial-Time Approximation Algorithm for the Permanent of a Matrix with Nonnegative Entries”. In: *Journal of the ACM* 51.4 (2004), pp. 671–697 (cit. on p. 2).

[JVV86] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. “Random Generation of Combinatorial Structures from a Uniform Distribution”. In: *Theoretical Computer Science* 43 (1986), pp. 169–188 (cit. on p. 2).

[Kit95] A. Kitaev. *Quantum Measurements and the Abelian Stabilizer Problem*. arXiv:quant-ph/9511026. 1995 (cit. on pp. 2, 4, 20).

[Kol18] V. Kolmogorov. “A Faster Approximation Algorithm for the Gibbs Partition Function”. In: *Proceedings of the 31st Conference On Learning Theory (COLT)*. 2018, pp. 228–249 (cit. on p. 3).

[Low17] G. H. Low. “Quantum Signal Processing by Single-Qubit Dynamics”. PhD thesis. Massachusetts Institute of Technology, 2017 (cit. on pp. 5, 20).

[LV06] L. Lovász and S. Vempala. “Simulated annealing in convex bodies and an O*(n4) volume algorithm”. In: *Journal of Computer and System Sciences* 72.2 (2006), pp. 392–417 (cit. on pp. 16, 23).

[MNRS11] F. Magniez, A. Nayak, J. Roland, and M. Santha. “Search via Quantum Walk”. In: *SIAM Journal on Computing* 40.1 (2011), pp. 142–164 (cit. on pp. 2, 15).

[Mon15] A. Montanaro. “Quantum Speedup of Monte Carlo Methods”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 471.2181 (2015), p. 20150301 (cit. on pp. 2–5, 9, 15, 16).

[MS13] E. Mossel and A. Sly. “Exact thresholds for Ising–Gibbs Samplers on General Graphs”. In: *The Annals of Probability* 41.1 (2013), pp. 294–328 (cit. on p. 16).

[MW05] C. Marriott and J. Watrous. “Quantum Arthur–Merlin games”. In: *computational complexity* 14.2 (2005), pp. 122–152 (cit. on pp. 2, 5, 7, 8, 21).

[Nay99] A. Nayak. “Lower Bounds for Quantum Computation and Communication”. PhD thesis. University of California, Berkeley, 1999 (cit. on p. 11).

[NC11] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. 10th ed. Cambridge University Press, 2011 (cit. on p. 9).

[NW99] A. Nayak and F. Wu. “The Quantum Query Complexity of Approximating the Median and Related Statistics”. In: *Proceedings of the 31st Symposium on Theory of Computing (STOC)*. 1999, pp. 384–393 (cit. on p. 11).

[ORR13] M. Ozols, M. Roetteler, and J. Roland. “Quantum Rejection Sampling”. In: *ACM Transactions on Computation Theory* 5.3 (2013), 11:1–11:33 (cit. on p. 2).

[Sin82] Y. G. Sinai. *Theory of Phase Transitions: Rigorous Results*. Pergamon, 1982 (cit. on p. 1).

[ŠVV09] D. Štefankovič, S. Vempala, and E. Vigoda. “Adaptive Simulated Annealing: A Near-Optimal Connection between Sampling and Counting”. In: *Journal of the ACM* 56.3 (2009), pp. 1–36 (cit. on pp. 1–4, 14).

[Sze04] M. Szegedy. “Quantum Speed-Up of Markov Chain Based Algorithms”. In: *Proceedings of the 45th Symposium on Foundations of Computer Science (FOCS)*. 2004, pp. 32–41 (cit. on pp. 2, 15).

[Ter99] B. M. Terhal. “Quantum Algorithms and Quantum Entanglement”. PhD thesis. University of Amsterdam, 1999 (cit. on p. 2).

[TOV+11] K. Temme, T. J. Osborne, K. G. H. Vollbrecht, D. Poulin, and F. Verstraete. “Quantum Metropolis sampling”. In: *Nature* 471 (2011), pp. 87–90 (cit. on p. 2).

[Val79] L. G. Valiant. “The Complexity of Computing the Permanent”. In: *Theoretical Computer Science* 8.2 (1979), pp. 189–201 (cit. on p. 2).

[VC72] J. P. Valleau and D. N. Card. “Monte Carlo Estimation of the Free Energy by Multistage Sampling”. In: *The Journal of Chemical Physics* 57.12 (1972), pp. 5457–5462 (cit. on p. 2).

[WA08] P. Wocjan and A. Abeyesinghe. “Speedup via Quantum Sampling”. In: *Physical Review A* 78.4 (2008), p. 042336 (cit. on p. 2).

[WCNA09] P. Wocjan, C.-F. Chiang, D. Nagaj, and A. Abeyesinghe. “Quantum Algorithm for Approximating Partition Functions”. In: *Physical Review A* 80.2 (2009), p. 022340 (cit. on pp. 2, 3).

[YLC14] T. J. Yoder, G. H. Low, and I. L. Chuang. “Fixed-Point Quantum Search with an Optimal Number of Queries”. In: *Physical Review Letters* 113 (2014), p. 210501 (cit. on p. 22).

Appendix A Algorithmic primitives

We use the next properties that characterize the output distribution of the quantum phase estimation algorithm.

Lemma A.1 (Phase estimation [Kit95; BHMT02]). *Let U be a unitary operator with an eigenvector $|\psi\rangle$ such that $U|\psi\rangle = e^{2\pi i\theta}|\psi\rangle$ where $\theta \in [0,1)$. Given a power-of-two integer t, the* phase estimation *algorithm outputs an estimate $\widetilde{\theta} = i/t$ where the number $i \in \{0,\dots,t-1\}$ is chosen with probability*

$$p(i) = \frac{\sin^2(t\Delta_i\pi)}{t^2\sin^2(\Delta_i\pi)} \tag{13}$$

and $\Delta_i = \min\{|\theta - i/t|, |1+\theta-i/t|, |1-\theta+i/t|\}$ is the distance between θ and i/t along the unit circle. The algorithm needs one copy of $|\psi\rangle$, which is restored at the end of the computation, and $O(t)$ controlled-U operations.

Corollary A.2. *If $t \geq 8$ then the probability distribution $(p(i))_i$ defined in Equation (13) satisfies $p(i) \geq 0.22$ when $\Delta_i \leq 5/(8t)$ and $p(i) \leq 0.11$ when $\Delta_i \geq 1/t$.*

Proof. Suppose first that $\Delta_i \leq 5/(8t)$. Using the inequality $\sin(x) \leq x$ when $0 \leq x \leq \pi/2$ and the fact that $y \mapsto (\sin(y)/y)^2$ is non-increasing on $[0, 5\pi/8]$, we have $p(i) \geq (\sin(t\Delta_i\pi)/(t\Delta_i\pi))^2 \geq (\sin(5\pi/8)/(5\pi/8))^2 \geq 0.22$. Suppose now that $\Delta_i \geq 1/t$. Using that $\Delta_i \leq 1/2$ and $t \geq 8$, we have $p(i) \leq 1/(t\sin(\pi/t))^2 \leq 1/(8\sin(\pi/8))^2 \leq 0.11$. □

We consider a variant of quantum amplitude amplification [BHMT02] that provides a precise linear amplification of the amplitude.

Lemma A.3 (Linear amplitude amplification, Theorem 6.10 in [Low17]). *Consider two reals $t \geq 1$ and $\epsilon \in (0,1)$. Let $|\psi\rangle$ be a quantum state and Π be a projection operator such that $\|\Pi|\psi\rangle\| \leq 1/(2t)$ if $t > 1$. Then, there is a unitary operator $V_{t,\epsilon}$ such that*

$$\big|\|\Pi V_{t,\epsilon}|\psi\rangle\| - t\|\Pi|\psi\rangle\|\big| \leq \epsilon$$

which can be implemented with $O(t\log(1/\epsilon))$ applications of the reflection operators $\mathbb{I} - 2|\psi\rangle\langle\psi|$ and $\mathbb{I} - 2\Pi$.

The next two results provide efficient quantum algorithms for converting between phase and amplitude encodings of a real parameter.

Lemma A.4 (Amplitude-to-Phase conversion, Theorem 14 in [GAW17]). *Let $|\psi\rangle$ be a quantum state and Π be a projection operator with $p = \|\Pi|\psi\rangle\|^2$. Then, given a real $\epsilon \in (0,1)$, there is a unitary operator U_ϵ and an integer $a = O(\log\log 1/\epsilon)$ such that*

$$U_\epsilon\big(|\psi\rangle|0\rangle^{\otimes a}\big) = |\psi\rangle|\varphi_p\rangle \quad \textit{where} \quad \||\varphi_p\rangle - e^{ip}|0\rangle^{\otimes a}\| \leq \epsilon$$

and U_ϵ can be implemented with $O(\log(1/\epsilon))$ applications of the reflection operators $\mathbb{I} - 2|\psi\rangle\langle\psi|$ and $\mathbb{I} - 2\Pi$.

Lemma A.5 (Phase-to-Amplitude conversion, Lemma 16 in [GAW17]). *Let U be a unitary operator with an eigenvector $|\psi\rangle$ such that $U|\psi\rangle = e^{ip}|\psi\rangle$ where $p \in [1/4, 3/4]$. Then, given a real $\epsilon \in (0,1)$, there is a unitary operator V_ϵ and an integer $a = O(\log 1/\epsilon)$ such that*

$$V_\epsilon\big(|\psi\rangle|0\rangle^{\otimes a}\big) = |\psi\rangle\big(\sqrt{1-p'}|0\rangle^{\otimes a} + \sqrt{p'}|\Phi^\perp\rangle\big) \quad \textit{where} \quad |\sqrt{p'} - \sqrt{p}| \leq \epsilon \textit{ and } \langle 0|\Phi^\perp\rangle = 0$$

and V_ϵ can be implemented with $O(\log(1/\epsilon))$ applications of the (controlled) U and $U^\dagger$ operators.

Finally, we explain how to anneal from one state to the next, given that there is a non-negligible overlap between the two. The idea is due to Marriott and Watrous [MW05] and is similar to the coin flipping algorithm (Proposition 2.3).

Algorithm 7. Annealing from $|\psi\rangle$ to $|\phi\rangle$, $\mathsf{Anneal}(|\psi\rangle, |\phi\rangle)$

```
Start with the state $|\psi\rangle$.
repeat
    Measure in the $\{|\psi\rangle\langle\psi|, \mathbb{I} - |\psi\rangle\langle\psi|\}$ basis.
    Measure in the $\{|\phi\rangle\langle\phi|, \mathbb{I} - |\phi\rangle\langle\phi|\}$ basis.
until the final measurement outcome is $|\phi\rangle$.
```

Lemma A.6 (Probabilistic annealing)**.** *Let $|\psi\rangle$ and $|\phi\rangle$ be two linearly independent quantum states of equal dimension. Then the expected number of reflections through $|\psi\rangle$ and $|\phi\rangle$ performed by Algorithm 7 is $1 + 1/(2|\langle\psi|\phi\rangle|^2)$.*

Proof. Let $p = |\langle\psi|\phi\rangle|^2$. Since the first iteration starts in state $|\psi\rangle$, the outcome of the first measurement is deterministic, and hence we will terminate with probability p. Afterward, every iteration starts in a state $|\phi^\perp\rangle$, which is orthogonal to $|\phi\rangle$ and in the 2D-subspace spanned by $|\psi\rangle$ and $|\phi\rangle$. Thus, the algorithm terminates with probability $f = 2p(1-p)$, in this iteration. Putting everything together, let T be the number of iterations the algorithm uses. Then, $\mathbb{E}[T-1|T \geq 2] = \sum_{t=2}^{\infty}(t-1)f(1-f)^{t-2} = 1/f$, and so $\mathbb{E}[T] = 1 + (1-p)/f = 1 + 1/(2p)$. □

Appendix B Nondestructive amplitude estimation

The *nondestructive amplitude estimation* algorithm is a variant of *amplitude estimation* introduced by Harrow and Wei [HW20] for restoring the initial quantum state with high probability. This is a crucial ingredient in their work for computing a cooling schedule faster than classically. In Section 2.2, we further extend the properties of amplitude estimation by making the output nearly unbiased. Our algorithm requires a rough amplitude estimate to start with, which can be obtained from the (biased) procedure of [HW20].

Below, we recall the statement of the nondestructive amplitude estimation theorem of Harrow and Wei [HW20] and we suggest a new (still biased) construction achieving this result. While the original algorithm of Harrow and Wei may be correct, its analysis seems to use two properties of amplitude estimation [BHMT02] that do not hold in general: the collapsing of the post-measurement state to an eigenvector of the Grover operator (it can be a superposition of two eigenvectors when the amplitude p is close to 0), and the output distribution giving overwhelming probability to a single amplitude estimate (there are two high-probability estimates when the underlying phase $\frac{1}{\pi}\arcsin(\sqrt{p})$ is at equal distance from the two nearest phase estimates on the unit circle). We avoid such complications by describing a new nondestructive amplitude estimation algorithm based on the next "uncomputation trick" with boosted success probability.

Lemma B.1 (Amplified uncomputation)**.** *Consider a quantum algorithm that consists of applying a unitary U on a state $|\psi\rangle|0\rangle^{\otimes k}$ and measuring the last k qubits in the standard basis. Let $\pi(1), \ldots, \pi(2^k)$ denotes the probabilities of measuring the outcomes $1, \ldots, 2^k$ respectively. Then, given a lower bound $\lambda \leq \sum_{i=1}^{2^k} \pi(i)^2$ and a success parameter $\eta \in (0,1)$, one can sample $i \in \{1, \ldots, 2^k\}$ from the distribution*

$$i \sim \frac{\pi(i)^2}{\sum_{j=1}^{2^k} \pi(j)^2}$$

and restore the state $|\psi\rangle$ with probability at least $1-\eta$ by using $O\big(\log(1/\eta)/\sqrt{\lambda}\big)$ applications of U, $U^\dagger$ and $\mathbb{I} - 2|\psi\rangle\langle\psi|$.

Proof. Let $U(|\psi\rangle|0\rangle^{\otimes k}) = \sum_i \alpha_i|\psi_i\rangle|i\rangle$ denote the state computed by the unitary U. Define the "uncomputation" unitary $U_{\mathrm{uncomp}} = (U^\dagger \otimes \mathbb{I}_{\mathbb{C}^{2^k}})(\mathbb{I}_{\mathcal{H}} \otimes \mathrm{CNOT})(U \otimes \mathbb{I}_{\mathbb{C}^{2^k}})$ that consists of running U, copying the output register into a new register and running the inverse $U^\dagger$. Then, the amplitude of $|\psi\rangle|0\rangle|i\rangle$ in the state obtained by applying U_{uncomp} on $|\psi\rangle|0\rangle^{\otimes k}|0\rangle^{\otimes k}$ is

$$(\langle\psi|\langle 0|\langle i|)U_{\mathrm{uncomp}}(|\psi\rangle|0\rangle|0\rangle) = \Big(\sum_j \alpha_j^*\langle\psi_j|\langle j|\langle i|\Big)\Big(\sum_j \alpha_j|\psi_j\rangle|j\rangle|j\rangle\Big) = |\alpha_i|^2 = \pi(i).$$

We run fixed-point amplitude amplification [YLC14] on the projector $\Pi = |\psi\rangle\langle\psi| \otimes |0\rangle\langle 0| \otimes \mathbb{I}$ and the state $U_{\mathrm{uncomp}}(|\psi\rangle|0\rangle|0\rangle)$, using the amplitude lower bound $\|\Pi U_{\mathrm{uncomp}}(|\psi\rangle|0\rangle|0\rangle)\|^2 \geq \lambda$ and the amplification parameter η. The probability of measuring $|\psi\rangle|0\rangle$ in the first two registers of the resulting state is at least $1-\eta$, in which case the last register collapses to $\frac{1}{\sqrt{\sum_i \pi(i)^2}}\sum_i \pi(i)|i\rangle$. Thus, we can restore $|\psi\rangle$ and sample $i \sim \frac{\pi(i)^2}{\sum_j \pi(i)^2}$ (by measuring the last register) with success probability at least $1-\eta$. The algorithm requires $O(\log(1/\eta)/\sqrt{\lambda})$ applications of (controlled) U_{uncomp}, $U_{\mathrm{uncomp}}^\dagger$, $\mathbb{I} - 2\Pi$ and $\mathbb{I} - 2|\psi\rangle\langle\psi|$. □

We obtain below the same statement as [HW20] for performing nondestructive amplitude estimation. Furthermore, we show that the amplitude estimate is zero with high probability when the amplitude is sufficiently small (as is the case for the regular amplitude estimation [BHMT02]).

Proposition B.2 (Nondestructive amplitude estimation)**.** *Let $|\psi\rangle$ be a quantum state and Π be a projection operator with $p = \|\Pi|\psi\rangle\|^2$. Given $t \geq 1$ and $\eta \in (0, 1/2)$, the* nondestructive amplitude estimation *algorithm* Nd-Amplitude$(|\psi\rangle, \Pi, t, \eta)$ *outputs an estimate $\widetilde{p} \in (0,1)$ such that*

$$|\widetilde{p} - p| < \frac{\sqrt{p(1-p)}}{t} + \frac{1}{t^2}$$

with probability at least $1-\eta$. Moreover, if $p \leq 1/(4t^2)$ then it outputs $\widetilde{p} = 0$ with probability at least $1-\eta$. The algorithm needs one copy of $|\psi\rangle$, which is restored at the end of the computation with probability at least $1-\eta$, and $O(t\log(1/\eta))$ applications of the reflection operators $\mathbb{I} - 2|\psi\rangle\langle\psi|$ and $\mathbb{I} - 2\Pi$.

Proof. The standard analysis of the amplitude estimation algorithm (Theorems 11 and 12 in [BHMT02]) shows that, after t steps, it outputs with probability at least $8/\pi^2$ one of two values $\widetilde{p} \in \{\widetilde{p}_1, \widetilde{p}_2\}$ that both satisfy $|\widetilde{p}_i - p| < 2\pi\frac{\sqrt{p(1-p)}}{t} + \frac{\pi^2}{t^2}$. Moreover, if $p \leq 1/(4t^2)$ then the probability to output $\widetilde{p} = 0$ is at least $8/\pi^2$. Thus, by taking the median of $O(\log(1/\eta))$ independent runs of amplitude estimation, we obtain $\widetilde{p} \in \{\widetilde{p}_1, \widetilde{p}_2\}$, or $\widetilde{p} = 0$ when $p \leq 1/(4t^2)$, with probability at least $1-\eta/2$. We apply the amplified uncomputation result (Lemma B.1) on this algorithm with amplitude lower bound $\lambda = 1/8$ and success parameter $\eta/2$. If the algorithm succeeds, the probability to sample $\widetilde{p} \in \{\widetilde{p}_1, \widetilde{p}_2\}$ is at least $\frac{2(1/2-\eta/4)^2}{2(1/2-\eta/4)^2+(\eta/2)^2} \geq 1-\eta/2$, and when $p \leq 1/(4t^2)$ the probability of $\widetilde{p} = 0$ is at least $\frac{(1-\eta/2)^2}{(1-\eta/2)^2+(\eta/2)^2} \geq 1-\eta/2$. □

Appendix C Volume estimation of convex bodies

In this appendix, we describe a quantum algorithm for estimating the volume of convex bodies that provides an improvement over the algorithm given in [CCH+19]. The problem statement is as follows. Let $K \subseteq \mathbb{R}^d$ be a convex body, and let $R > 0$ such that $B(1) \subseteq K \subseteq B(R)$, where $B(r) = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ is the ball of radius r centered at 0. We wish to estimate the volume $\mathrm{Vol}(K)$ of K up to multiplicative precision ϵ, i.e., to output $\widetilde{V}$ such that $(1-\epsilon)\,\mathrm{Vol}(K) \leq \widetilde{V} \leq (1+\epsilon)\,\mathrm{Vol}(K)$. We assume to have access to the convex body by means of a membership oracle O_K, i.e., for any point $x \in \mathbb{R}^d$, we can query whether $x \in K$. For the purpose of this paper, we are interested in the query complexity of this problem, i.e., we wish to minimize the number of queries made to this membership oracle. We note here that previous work also consider time-efficient implementations of algorithms that solve this problem [CCH+19]. We expect that similar techniques could also be used to implement our algorithm time-efficiently, but we leave this for future work.

First, observe that if the precision ϵ is smaller than $\epsilon \leq (3/4)^d$ then any polylogarithmic overhead in $1/\epsilon$ is polynomial in the dimension d. Thus, in this regime we can run a simple adaption of approximate counting on a suitably dense discretization of $B(R)$ – this will run with $O(1/\epsilon)$ queries, which when adding polylogarithmic overhead in already achieves the desired $\widetilde{O}(d^3 + d^{2.25}/\epsilon)$ query complexity. Thus, without loss of generality, we can focus on the regime in which $\epsilon > (3/4)^d$.

Previous works make use of the *pencil construction*, introduced in [LV06], where the idea is to define a new convex body $K' \in \mathbb{R}^{d+1}$ with one extra dimension, as

$$K' = \Big\{\mathbf{x} = (x_0, x) \in \mathbb{R}^{d+1} : x \in K \wedge x_0 \in [0, 2R] \wedge \|x\|_2 \leq x_0\Big\}.$$

The algorithm now consists of two steps. First, the volume of K' is estimated up to multiplicative precision $\epsilon/2$ using the Markov Chain Monte Carlo framework, with the partition function defined as

$$Z(\beta) = \int_{K'} e^{-\beta x_0} \, \mathrm{d}\mathbf{x}.$$

For sufficiently large values of β, the value of the partition function is almost completely determined by the tip of the pencil, whose shape is known to us a priori. On the other hand, for sufficiently small values of β, the partition function essentially captures the volume of K'. These claims are made more precise in [CCH+19] and the references therein, resulting in Algorithm 4 in said paper which uses a cooling schedule of length $m = \widetilde{\Theta}(\sqrt{d})$. We can speed up this part by using our unbiased product estimator (Theorem 3.3) in Step 2 of said algorithm. This reduces the number of steps of the quantum walk by a factor of $\sqrt{m} = \widetilde{\Theta}(d^{1/4})$. Hence, we can estimate the volume of K' with $\widetilde{O}(d^3 + d^{2.25}/\epsilon)$ calls to the membership oracle.

The second step relates the volumes of K and K'. It relies on the observation that $R\,\mathrm{Vol}(K) \leq \mathrm{Vol}(K') \leq 2R\,\mathrm{Vol}(K)$ since $[R, 2R] \times K \subseteq K' \subseteq [0, 2R] \times K$. The idea is then to use *rejection sampling* to obtain an $\epsilon/2$-precise multiplicative estimate of the ratio between $\mathrm{Vol}(K')$ and $\mathrm{Vol}(K)$. The procedure outlined in [CCH+19, Page 29] prepares approximately uniform samples from K in $\widetilde{O}(d^{2.5})$ calls to the membership oracle. These samples can be trivially converted to samples from $[0, 2R] \times K$ by sampling uniformly from the interval $[0, 2R]$ and adding the result as an extra dimension. Classically, one could now take $O(1/\epsilon^2)$ uniform samples from $[0, 2R] \times K$, and count the fraction that is contained in K'. This yields an ϵ-precise multiplicative estimate of the ratio between $\mathrm{Vol}(K)$ and $\mathrm{Vol}(K')$. Quantum approximate counting speeds up this step and only requires $O(1/\epsilon)$ quantum samples, yielding a total of $\widetilde{O}(d^{2.5}/\epsilon)$ queries to the membership oracle in this step of the algorithm.

We claim that the second step can be done with $\widetilde{O}(d^2/\epsilon)$ instead of $\widetilde{O}(d^{2.5}/\epsilon)$ queries, essentially because sampling from K requires in fact only $\widetilde{O}(d^2)$ quantum queries. For simplicity in the proof, we describe a slightly different rejection sampling strategy achieving this complexity. Instead of sampling from $[0, 2R] \times K$ and checking whether the sample is contained in K', we sample from K' and check whether the resulting sample is contained in $[R, 2R] \times K$. Using the same analysis as above, we obtain an $\epsilon/2$-precise multiplicative estimate of the ratio between $\mathrm{Vol}(K)$ and $\mathrm{Vol}(K')$, using $O(1/\epsilon)$ quantum samples from K'. Sampling approximately uniformly from K' can be achieved with the simulated annealing schedule from [CCH+19, Algorithm 3] that requires $\widetilde{O}(d^2)$ calls to the membership oracle. It remains to check that the obtained samples from K' are sufficiently close to uniform, which they are if we marginally increase the length of the cooling schedule:

Let $\epsilon_1 > 0$ be fixed later, and let $\beta' \leq \epsilon_1/(2R)$. We show that the Gibbs state at this inverse temperature, denoted by $|\pi'\rangle = |\pi'_{\beta'}\rangle$, is close to the uniform distribution over K', denoted by $|\pi\rangle$. To that end, recall that $\mathrm{Vol}(K') \geq Z(\beta')$, and so

$$\begin{aligned}\mathrm{Vol}(K')(1 - \langle\pi|\pi'\rangle) &= \int_{K'} \left(1 - \sqrt{\frac{\mathrm{Vol}(K')}{Z(\beta')}} e^{-\beta' x_0}\right) \,\mathrm{d}\mathbf{x} \leq \int_{K'} \left(1 - e^{-\beta' x_0}\right) \,\mathrm{d}\mathbf{x} \\ &\leq \int_{K'} \beta' x_0 \,\mathrm{d}\mathbf{x} \leq 2\beta' R \,\mathrm{Vol}(K') \leq \epsilon_1 \,\mathrm{Vol}(K'),\end{aligned}$$

which implies that $\||\pi\rangle - |\pi'\rangle\| = 2\sqrt{1 - \langle\pi|\pi'\rangle} \leq \sqrt{\epsilon_1}$.

Next, we can prepare $|\pi'\rangle$ by choosing $m = \lceil\sqrt{d}\log(4dR/\epsilon_1)\rceil$ in [CCH+19, Algorithm 3], where m denotes the length of the cooling schedule. The final inverse temperature then satisfies $\beta' \leq 2d(1-1/\sqrt{d})^{\sqrt{d}\log(4dR/\varepsilon_1)} \leq 2de^{-\log(4dR/\epsilon_1)} = \epsilon_1/(2R)$. Since in every annealing step we are performing N Gibbs state reflections, where $N = O(1)$, we can implement these reflections with precision $\sqrt{\epsilon_1}/(mN)$, to ensure that the total error amounting from the imperfections of these Gibbs state reflections amounts to $\sqrt{\epsilon_1}$ as well. Thus, we end up preparing a uniform quantum sample up to precision $2\sqrt{\epsilon_1}$, with a total number of $O(m\log(mN/\sqrt{\epsilon_1})) = \widetilde{O}(\sqrt{d}\log(1/\epsilon_1))$ reflections through Gibbs states.

Now, with $\epsilon_1 = 1/A^2(2N')^2 = \Theta(\epsilon^2)$, where $N' = \Theta(1/\epsilon)$ is the number of calls the state-preparation unitary in the approximate counting algorithm, we obtain that the total accumulated error throughout the whole procedure is at most $1/A$. With A a large enough constant, we ensure that the error probability of this step is negligible. Thus, we obtain that the total number of Gibbs state reflections in this second part of the procedure becomes $\widetilde{O}(N'\sqrt{d}\log(1/\epsilon_1)) = \widetilde{O}(\sqrt{d}/\epsilon)$. Since every reflection through a Gibbs state can be performed with $\widetilde{O}(d^{1.5}\mathrm{polylog}(1/\epsilon))$ membership queries, the total number of calls to O_K becomes $\widetilde{O}(d^2/\epsilon)$. Thus, this second step is less costly than the first step, and the resulting total query complexity becomes $\widetilde{O}(d^3 + d^{2.25}/\epsilon)$.