arXiv:2510.08518v3 [quant-ph] 1 Oct 2026

Randomized truncation of quantum states

Aram W. Harrow * Angus Lowe * Freek Witteveen †

October 5, 2026

Abstract

A fundamental task in quantum information is to approximate a pure quantum state in terms of sparse states or, for a bipartite system, states of bounded Schmidt rank. The optimal deterministic approximation in each case is straightforward, and maximizes the fidelity: keep the largest entries or singular values. On the other hand, random mixtures of sparse states can achieve quadratically improved trace distances, and yield nontrivial bounds on other distance measures like the robustness.

In this work, we give efficient algorithms for finding mixtures of sparse states that optimally approximate a given pure state in either trace distance or robustness. These algorithms also yield descriptions of efficiently samplable ensembles of sparse, or less-entangled, states that correspond to these optimal mixed approximations. This can be used for the truncation step of algorithms for matrix product states, improving their accuracy while using no extra memory and no asymptotic increase in runtime, and we demonstrate this improvement numerically.

Our proofs rely on tools from convex optimization and zero-sum games, as well as rigorous gurarantees for computing and sampling from maximum-entropy distributions. In particular, a key step in our construction is *conditional Poisson sampling*, which is the well-studied problem of drawing fixed-size subsets from a population with known inclusion probabilities. We give improved algorithms for carrying out this sampling, which may be of independent interest.

Contents

1 Introduction **2**
- 1.1 Simple example . . . 3
- 1.2 Overview of results . . . 4
- 1.3 Related work . . . 6
- 1.4 Proof techniques and intuition . . . 8
- 1.5 Open questions . . . 10
- 1.6 Notation . . . 10

2 Quantifying coherence and entanglement **11**

3 Computational hardness for mixed states **14**

*Center for Theoretical Physics – a Leinweber Institute, MIT, Cambridge, MA, 02139, USA
†QuSoft and CWI, Science Park 123, 1098 XG Amsterdam, Netherlands

4 Approximations in trace distance **15**
4.1 Computing the trace distance and optimal measurement . . . 16
4.2 Subset sampling . . . 21
4.3 Computing and sampling the optimal mixed approximation . . . 23
4.4 Coherence versus entanglement for the trace distance . . . 25
4.5 Variance estimates for ensembles . . . 27

5 Approximations in robustness **28**
5.1 Computing and sampling the mixed approximation in robustness . . . 28

6 Examples **30**
6.1 Additional simple examples . . . 30
6.2 Power laws . . . 32

7 Application to matrix product states **36**

8 Improved algorithms for conditional Poisson sampling **38**
8.1 Maximum-entropy distributions . . . 39
8.2 Sampling from the maximum-entropy distribution . . . 40
8.3 Inference . . . 43
8.4 Heuristics and conjecture for faster inference . . . 48

A Background material **49**
A.1 Minimax and zero-sum games . . . 49
A.2 Convex analysis . . . 51
A.3 Spectral stability . . . 52

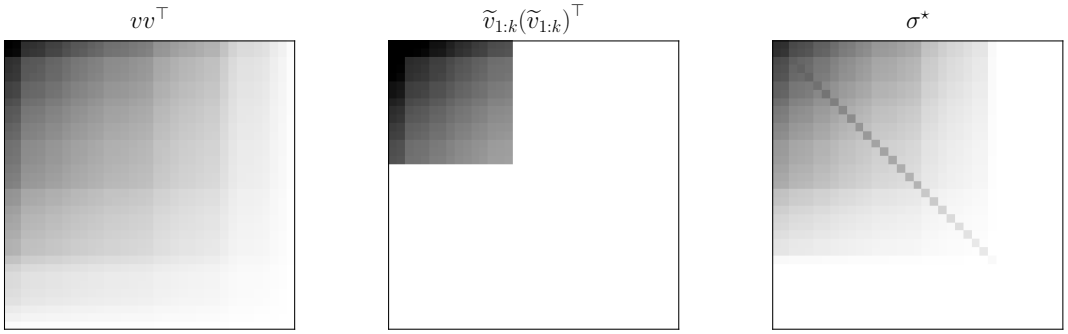
B Proof of Lemma 2.11 **52**

Editorial note: This is arxiv v3 of our paper. It is posted for the benefit of the QIP PC to meet the arxiv requirement for QIP 2027. We anticipate further updates soon to incorporate the fact that a proof has recently been posted of Conjecture 8.10. As a preview of those changes, we anticipate further runtime improvements to the algorithms in Section 8.

1 Introduction

This paper is motivated by a deceptively simple question: given a target unit vector, how well can we approximate it with a sparse, but potentially random, unit vector? We say a vector is k-sparse if it has at most k nonzero entries in the standard basis. Given a d-dimensional target vector v and a sparsity parameter $k \leq d$, it is clear that keeping just the k entries of v with the largest absolute values gives the best approximation among all deterministic choices for the k-sparse vector. Similarly, it is well known that the optimal low-rank approximation to a matrix, as measured by any unitarily invariant norm, is obtained by truncating (keeping just a few of) its singular values. Another instance of this type of problem arises when simulating quantum many-body systems using tensor networks, where multipartite states are represented using states that have low Schmidt rank with respect to several partitions of the subsystems. In each case—vectors, matrices, and tensors—we refer to the problem of finding a sparse approximations as *truncation*.

We focus on the setting of quantum information, although truncation has far-reaching applications in other areas of computer science as well. Here, a unit vector v corresponds to a pure quantum state



(a) (b) (c)

Figure 1: With $d = 35$, $k = 15$, and darker cells representing larger values: (a) Quantum state given by a random unit vector $v \in \mathbb{R}^d$ with positive entries sorted in decreasing order. (b) Approximation obtained by keeping the first k entries of v and normalizing, which yields $\widetilde{v}_{1:k}$. (c) Optimal approximation $\sigma^\star$ of $vv^\top$, with respect to trace distance, by a convex combination of states given by k-sparse unit vectors.

$\rho = vv^\dagger$, where v is interpreted as a column vector[1]. The quantum setting suggests a few important features of this problem. First, observables are quadratic in v (or linear in ρ) and appoximation error is defined primarily in terms of deviations of these observables from their true values. This can be measured using the standard *trace distance* if we meaure additive error, or the less common *robustness* if we consider multiplicative error. Second, there are many natural physical settings in which superposition or entanglement are expensive resources, but randomness is cheap. Thus it is natural to define the truncation task as approximating ρ with a mixture of k-sparse states, or analogously a mixture of states with Schmidt rank at most k.

The main result of our paper is that when we define truncation in this way, the simple answer of taking the top k entries of v is no longer optimal: Even when we are approximating pure states, the optimal truncation for either trace distance or robustness will generally be a mixed state. As one might expect, finding these optimal mixed truncations is more complicated than simply sorting the vector and keeping the top k entries; nevertheless, we show it can be done efficiently (see Theorem 1.1).

A key application of our result is the truncation of matrix product states (MPSs). Traditionally this is done by running a series of singular value decompositions (SVDs) and keeping the top k singular values at each step. Our randomized method improves the accuracy obtained in each of these truncations as measured by, say, the trace distance, at the cost of some overhead. However, as we show in Section 8.3, the asymptotic runtime is no worse than that of the SVD on its own. We remark that this question is really about classical optimization and computation; the input v is given as a classical description of the quantum state and the algorithms are classical.

1.1 Simple example

To make the benefits of mixed approximations concrete, we will describe a simple but instructive example. Let $v = (\sqrt{1-2\varepsilon}, \sqrt{\varepsilon}, \sqrt{\varepsilon})^\top$ and suppose we want to approximate this by a 2-sparse state. The best approximation in fidelity is given by either $w_1 \propto (\sqrt{1-2\varepsilon}, \sqrt{\varepsilon}, 0)^\top$ or $w_2 \propto$

[1]We eschew bra-ket notation in this paper, partly to emphasize that our quantum results apply more generally to any normalized complex vectors.

$(\sqrt{1-2\varepsilon}, 0, \sqrt{\varepsilon})^\top$, both giving a fidelity $F(v, w_i) = \sqrt{1-\varepsilon}$, and hence a trace distance of $T(v, w_i) = \sqrt{\varepsilon}$. If we look at the density matrix

$$vv^\top = \begin{pmatrix} 1-2\varepsilon & \sqrt{\varepsilon(1-2\varepsilon)} & \sqrt{\varepsilon(1-2\varepsilon)} \\ \sqrt{\varepsilon(1-2\varepsilon)} & \varepsilon & \varepsilon \\ \sqrt{\varepsilon(1-2\varepsilon)} & \varepsilon & \varepsilon \end{pmatrix} \tag{1}$$

then we see that when truncating to one of the w_i, the dominant error in trace distance comes from the off-diagonal terms on the order of $\sqrt{\varepsilon}$. We can mitigate this by taking a uniform mixture of states $\tilde{w}_1 \sim (\sqrt{1-2\varepsilon}, 2\sqrt{\varepsilon}, 0)^\top$ and $\tilde{w}_2 \sim (\sqrt{1-2\varepsilon}, 0, 2\sqrt{\varepsilon})^\top$. This way, we see that the mixed state

$$\sigma = \frac{1}{2}\tilde{w}_1\tilde{w}_1^\top + \frac{1}{2}\tilde{w}_2\tilde{w}_2^\top = \frac{1}{1+2\varepsilon}\begin{pmatrix} 1-2\varepsilon & \sqrt{\varepsilon(1-2\varepsilon)} & \sqrt{\varepsilon(1-2\varepsilon)} \\ \sqrt{\varepsilon(1-2\varepsilon)} & 2\varepsilon & 0 \\ \sqrt{\varepsilon(1-2\varepsilon)} & 0 & 2\varepsilon \end{pmatrix} \tag{2}$$

only differs from v in trace distance to order ε instead of $\sqrt{\varepsilon}$. Intuitively, the scarce resource here is the superposition between the 1st entry and the 2nd and 3rd entries, and we can boost this by increasing the amplitude of the 2nd and 3rd entries. This introduces inaccuracies of order ε along the diagonals while fixing the error of order $\sqrt{\varepsilon}$. The tradeoff between these two sources of error is also visible in the larger example illustrated in Figure 1.

The approximation in Eq. (2) turns out to achieve the optimal robustness of 2ε, meaning $vv^\top \preceq (1+2\varepsilon)\sigma$. In Section 5 we will see more generally how to construct this sort of robustness-optimal ensemble.

1.2 Overview of results

Efficient algorithms. We study the task of optimally approximating a pure state by a mixture of k-sparse states. This is both a computational and a mathematical problem. It is important that our target states be pure since the more general question of approximating mixed states with mixtures of k-sparse states — or even determining whether a given mixed state can be decomposed into k-sparse states — is NP-hard. These, and other hardness results, are discussed in Section 3.

Our main result is a collection of efficient algorithms that can find these randomized approximations when the state is pure. As we explain in further detail below, the runtime is dominated by the time required to find the right parametrization of a certain Gibbs distribution.

Theorem 1.1 (Informal)**.** *Given a pure state $v \in \mathbb{C}^d$ and a sparsity bound $k \leq d$, there is an efficient algorithm to find a description of the optimal randomized approximation to v in terms of either trace distance T or robustness R. This description includes:*

1. *the optimal value of T or R;*
2. *the corresponding density matrix (with some additional polynomial-time overhead); and*
3. *a procedure to sample k-sparse pure states from this distribution in time $O(d)$.*

The same results hold when approximating a pure bipartite state $v \in \mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ by a mixture of states with Schmidt rank at most k, the relevant task for MPS truncation.

These algorithms are given in Sections 4 and 5, building on sampling algorithms described in Section 8. Specifically, the trace distance version of this statement follows from Theorems 4.1, 4.8, and 4.9. For the robustness, the optimal value computation appears in prior work [Reg17, JLP+18],

and we add to this by giving an efficient method to describe and sample from the optimal solution in Theorems 5.1 and 5.2.

We now describe the form of our approximating ensembles. We find that for both trace distance and robustness, the approximating sparse states involve keeping some of the top entries of v and sampling from some or all of the remaining entries. Concretely, assume without loss of generality that in the standard basis

$$v = \sum_{i=1}^{d} v_i e_i, \tag{3}$$

where e_i are the standard basis vectors, the entries v_i are nonnegative and ordered non-increasingly. Then the optimal approximation to v samples a subset S with probability $p(S)$ and takes a state v_S, corresponding to a density matrix

$$\rho = \sum_{S} p(S) w(S) w(S)^\dagger \tag{4}$$

Here,

$$w(S) = \underbrace{\alpha \sum_{i=1}^{r} v_i e_i}_{\text{keep top } k-r-1} + \underbrace{\frac{\beta}{\sqrt{r+1}} \sum_{i \in S} e_i}_{\substack{\text{uniform superposition} \\ \text{on } S}} \tag{5}$$

where α and β are some appropriately chosen values, and $r \leq k-1$ an index that depends on v. The set S is sampled from the indices that are not kept deterministically $\{k-r, k-r+1, \ldots, d\}$. For the trace distance approximation, S is sampled from a subset $\{k-r, k-r+1, \ldots, \ell\}$ for $\ell \leq d$, so some indices are truncated deterministically. The probability distribution on subsets S of size $r+1$ needs to be such that the marginals $q_i = \Pr(i \in S)$ and $Q_{ij} = \Pr(i, j \in S)$ satisfy certain constraints. The precise values of α, β and the distribution p depend on whether one targets an optimal approximation for trace distance or robustness. Let us give a complete description for the robustness; a precise specification for the optimal approximation for the trace distance can be found in Theorem 4.8 and Eq. (69). For the robustness, let r be the unique index such that (recall we assumed $v_i \geq 0$ and non-increasing)

$$v_{k-r-1} > \frac{s_{k-r}}{r+1} \geq v_{k-r} \qquad \text{where} \qquad s_{k-r} := \sum_{i=k-r}^{d} v_i \tag{6}$$

and let

$$R_k = \sum_{i=1}^{k-r-1} v_i^2 + \frac{s_{k-r}^2}{r+1} - 1. \tag{7}$$

Then R_k is the optimal robustness approximation. The states in Eq. (5) are written as

$$w(S) = \frac{1}{\sqrt{1+R_k}} \left(\sum_{i=1}^{r} v_i e_i + \frac{s_{k-r}}{r+1} \sum_{i \in S} e_i \right) \tag{8}$$

and the probability distribution over subsets needs to be such that the marginals satisfy $q_i = v_i(r+1)/s_{k-r}$ and $Q_{ij} \leq q_i q_j$.

The type of sampling we need is known as *conditional Poisson sampling* in the survey sampling literature. We also refer to it as *importance sampling of sets* in this work and discuss the task in more detail in Section 8. After defining the problem mathematically, there are several algorithmic questions; primarily *inference*, or determining the distribution parameters given the marginals, and *sampling* given the distribution parameters. Since we would like our randomized truncation algorithms to be of use for numerical linear algebra tasks, such as MPS calculations, we give new and faster algorithms for inference, sampling and other tasks, which may be of independent interest.

Quantifying the advantage of randomized truncation. While our results give efficient algorithms for finding the optimal approximations in trace distance and robustness, we would also like to have an intuitive explanation of when these improve significantly over the deterministic truncation. As we explain in Section 2, the best possible improvement in terms of the accuracy is quadratic. To get a sense of when the quadratic improvement in trace distance is or is not possible, we discuss a range of representative examples in Section 6. This turns out to depend on the decay rate of the sorted entries of the target vector, and specifically on the relation between the ℓ_1 and ℓ_2 norms of its tails.

We also note that, while the improvement in trace distance is at most quadratic, in order to facilitate a comparison along the lines of recent work in randomized compiling (see Section 1.3), which aims to reduce gate count, one may wish to determine the advantage of randomized truncation in terms of the relevant resource being conserved; namely, the sparsity parameter k. This advantage depends on the functional form of k with $\sqrt{\varepsilon}$, the best deterministic trace distance, and T, the best randomized trace distance, through $k(\sqrt{\varepsilon})/k(T)$, and may in principle be super-quadratic.

Numerics and application to MPS. An important motivation for our work is the application to randomized methods for simulating quantum systems, which we explore numerically in Section 7. In particular, MPS are a widely used class of states where our randomized truncation schemes serve as a drop-in replacement for the usual deterministic approach to reducing bond dimension. Code for our paper is available at https://github.com/angusjlowe/rtrunc. It includes a package for computing optimal approximations for pure states as well as code to reproduce the plots in this paper.

1.3 Related work

Randomized approximations to classical and quantum objects have appeared in several different contexts.

Randomized compiling. A related question to the one we study in this work is how well unitary operators can be approximated by mixtures of unitaries. Here, the power of randomness is somewhat greater for channels than for states, which we discuss further in Section 1.4. Prior work by Campbell [Cam17] and Hastings [Has16] first showed (nonconstructively) that mixing unitaries can quadratically reduce diamond-norm error under general conditions. This led to randomized product formulas for Hamiltonian simulation [COS19, Cam19, OWC20, HW23], rotations [Low21, KMB24], and Quantum Signal Processing [MR25]. The general question of approximating the set of all unitaries by convex combinations of an allowed set of unitaries was studied by Akibue, Kato and Tani [AKT24b], who found that mixing quadratically improves the worst-case error, i.e., given an $\sqrt{\varepsilon}$-net of unitaries, taking convex combinations yields an $O(\varepsilon)$-net in diamond norm.

Probabilistic state synthesis. Randomized approximations to states have also been considered previously in [AKT22, AKT24a]. Given a set of allowed pure states in the random mixture, these

works show a quadratic improvement in accuracy for the problem of *worst-case approximation* (i.e. the hardest possible target state v) as well as when certain symmetry conditions are met. Their setting does not substantially overlap with ours since we consider approximations for arbitrary pure states, which in general do not obey their symmetry conditions. Indeed, in Section 6, we give examples where the quadratic accuracy improvement of [AKT24a] does not apply. Refs. [AKT22, AKT24a] also provide algorithms based on semidefinite programming for finding the optimal randomized approximation, whose runtime scales polynomially in the number of allowed states in the mixture. In our setting, this number is at least exponential in the sparsity parameter k. See Section 3 for further discussion on this point. Although this prior work does not resolve our problem, their findings of widespread quadratic advantage serve as inspiration for pursuing efficient algorithms applicable to our setting. Finally, we note that simultaneous work [WZYZ25] studies the application of a randomized approach to state preparation on quantum devices, and considers a similar set of examples to those appearing in Section 6.

Resource theory of coherence. The idea of a "resource theory" is to define a set of "free" states and/or operations, and then to use this to define distances from these objects which quantify some resource of interest [CG19]. The original example of this was to study entanglement as a resource when local operations and classical communication (LOCC) is free [BBPS96]. A natural way to quantify a resource is the robustness (related to the max-relative entropy) of a given state ρ with respect to the free states [VT99, LBT19]. The resource theory that directly relates to our work is that of "k-incoherence" (also known as "k-coherence"), which treats mixtures of k-sparse states as free, and quantifies the distance of a given state from this set; see [SAP17] for a review of this topic. The $k = 1$ case was studied in [CGJ+16], using both the trace distance and robustness as resource measures. The $k > 1$ case was analyzed in detail in [Reg17, JLP+18]. Among other results, these works show that for pure states, different notions of robustness are equal to each other and can be expressed in terms of a norm called the k-support norm (see Definition 2.9). Our result on the robustness in Section 5 can be seen as a constructive version of the upper bound on the robustness given in Theorem 1 of [JLP+18].

The k-support norm. Both the optimal fidelity and the optimal robustness can be understood in terms of two norms which are dual to one another: the top-k norm, and the k-support norm. These are defined in [AFS12] and reviewed in our paper in Section 2. In [AFS12] these norms were used to develop convex relaxations for the task of finding sparse solutions to regression problems. They also gave an expression for the proximal operator for the k-support norm [AFS12, Section 4] which is similar to our trace distance optimization in Section 4. Gauge norms are also studied in optimization settings [CRPW12] where the goal of finding a simple model can be formalized through constraints on sparsity or rank, or more generally, being a linear combination of a small number of atoms.

Tensor network Monte Carlo. An important application of truncation, both deterministic and randomized, is to reduce the bond dimension in matrix product states and tensor networks. An immediate application of our work is to approximate a tensor by a lower-rank tensor, as we will discuss in more detail in Section 7. Ferris [Fer15] has proposed a tensor truncation scheme with some similarities to ours, and this idea has also been studied more recently in [Tod24]. We defer the detailed comparison of our work with [Fer15] to Section 7. As a brief summary of the comparison, our methods provably minimize robustness or trace distance for bipartite states, while Ferris' method samples indices using a heuristic which is not based on the solution to an optimization problem.

Randomized SVD. Truncating pure entangled states can be thought of as finding a low-rank approximation to a matrix, which is a basic task in numerical linear algebra. The standard method to solve this problem is to compute the singular value decomposition (SVD) and keep only the top singular values. This is precisely the deterministic truncation scheme that we generally compare our methods against. In defining error for randomized methods, there are two main possibilities: we can evaluate the error either before or after the random choices are made. Existing algorithms evaluate the error *after* the random choices are made. If the original matrix is M and the random low-rank approximation is $\tilde{M}$, then they evaluate something like $\mathbb{E}[\mathrm{dist}(M, \tilde{M})]$ for some distance measure. In this framework, randomized algorithms are, by definition, no more accurate than deterministic algorithms, although they may improve the efficiency. This is the setting typically considered for randomized SVDs; see [KT24] for a review.

Our framework corresponds instead to evaluating the error *before* the random choices are made. Suppose that $\mathrm{dist}(M, \tilde{M}) = \sup_\varphi \varphi(M - \tilde{M})$ where φ ranges over some set of linear functionals. Norms such as operator norm or Schatten norms fall into this framework. Our approach then consists of choosing a distribution to minimize

$$\sup_{\varphi} \mathop{\mathbb{E}}_{\tilde{M}} [\varphi(M - \tilde{M})]. \tag{9}$$

This setting is more realistic if φ represents a test that is unknown in advance, but is not chosen adversarially based on our choice of approximation $\tilde{M}$. As a result of our framework, it is possible to be more accurate (in terms of bias) than the low-rank deterministic approximation found using the SVD. We believe this is likely to have applications beyond quantum information.

1.4 Proof techniques and intuition

Our work shows that in order to approximate a pure state, given a library of simpler pure states, the best accuracy comes from introducing randomness. This may at first seem counterintuitive, but there are a few reasons why we might expect this to be the case.

One explanation comes from zero-sum games. The optimal trace distance approximation can be viewed as such a game where, given a target pure state v, we seek the value of

$$\min_{\sigma \in \mathcal{I}_k} \max_{0 \preceq M \preceq \mathbb{1}} \mathrm{tr}\Big(M(vv^\dagger - \sigma)\Big). \tag{10}$$

Here, $\mathcal{I}_k$ is the set of mixtures of k-sparse states. This has the form of a zero-sum game, with the σ player trying to minimize the distinguishability and the M player trying to maximize it. A common feature of equilibria is that the optimal strategies are mixed. In game terms, the σ player often does better with a mixed strategy because they may then perform simultaneously well against many possible responses by the M player. This picture also tells us that a fixed M obviates the need for a mixed strategy by the σ player. This happens when the goal is to minimize an observable such as the energy of a system, or when maximizing the fidelity with a fixed state. In these cases, optimal pure strategies exist, while optimal mixed strategies may or may not also exist.

This principle also explains why randomness helps with approximating unitaries. The diamond norm involves a maximization over both the final distinguishing measurement *and* the choice of input state. The choice of final measurement is similar to the state case that we consider in this paper, and leads to analogous up-to-quadratic improvements. The maximum over initial states can lead to more dramatic, dimension-dependent, improvements. For example, the d-dimensional unitaries $U_i = I - 2e_i e_i^\dagger$ are each far from the identity but the average channel $\mathcal{E}(\rho) = \mathbb{E}_{i \in [d]} U_i \rho U_i^\dagger$ is $O(1/d)$ close to the identity. Similar ideas are used in qDRIFT and multiqubit Toffoli gates [Cam19, CHKT21, GKZ25].

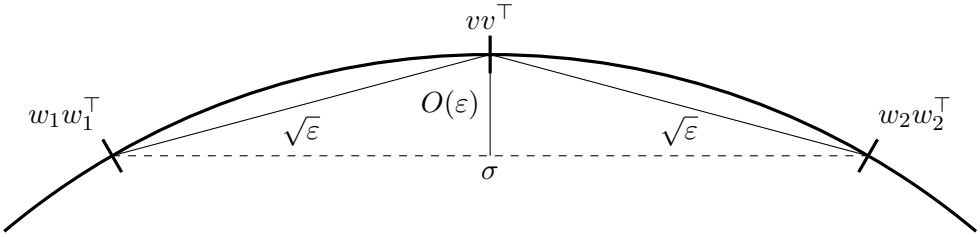


Figure 2: Illustration of the geometric intuition underlying cases where randomized truncation offers a significant improvement in approximating a pure state $vv^\top$.

As we saw in the state case, the advantage from randomness comes from achieving good performance against many different tests simultaneously. Thus if we fix both the input state and the final observable, then again deterministic approximation can achieve the optimal error.

Another way to understand the quadratic improvement in trace distance comes from the geometry of the state space, an idea which is due to [AKT22, AKT24b, AKT24a]. This is illustrated in Figure 2, where the best pure approximations to v are w_1, w_2, each at distance $\sqrt{\varepsilon}$, but a convex combination of these states has distance $O(\varepsilon)$.

Finally, our results illustrate the value of conditional Poisson sampling, the main idea for which we now sketch. Suppose $v_1, \ldots, v_d$ and $x_1, \ldots, x_d$ are positive reals with $\sum_i v_i = V$ and the goal is to estimate $s = \sum_{i=1}^d v_i x_i$. A natural unbiased estimator, familiar from Monte Carlo algorithms, is obtained by sampling i with probability v_i/V and outputting $\hat{s}_i \coloneqq V x_i$. This procedure can alternatively be viewed as constructing a randomized 1-sparse approximation to a vector $v = (v_1, \ldots, v_d)$. Namely, if $(e_j : j = 1, \ldots, d)$ are the standard basis vectors and i is drawn randomly with probability v_i/V then the vector-valued random variable Ve_i is an unbiased estimator of v.

If we wish to now generalize to a k-sparse approximation then we need to sample a random subset of the indices $S = \{i_1, \ldots, i_k\} \subset [d]$. As before, we can choose to impose that this k-sparse estimator be unbiased for the target vector v. This means that if $i \in S$ with probability $q_i \in (0, 1]$ then the random vector ought to be of the form

$$w(S) = \sum_{i \in S} \frac{v_i}{q_i} e_i. \tag{11}$$

Since $|S| = k$ and the q_i represent marginal probabilities, we must have $\sum_{i=1}^d q_i = k$. The task at hand — say, minimizing the variance of some function of this random vector — informs the desired values of the marginal probabilities $q_1, \ldots, q_d$.

For the randomized approximations in robustness and trace distance, we do not explicitly enforce that the k-sparse states be unbiased in the sense described above, but rather are of the form in Eq. (5), where we keep some terms deterministically, and sample from the other indices. This can be seen as a refinement of the approximation by importance sampling of sets in Eq. (11).

Conditional Poisson sampling, which is needed to compute and sample from the optimal ensemble, is a nontrivial algorithmic challenge. Algorithms for this task are known from the literature on maximum entropy sampling. We draw on this past work and improve upon it as well in Section 8, which may be of independent interest.

1.5 Open questions

We give a comprehensive analysis of the computational problem of finding optimal randomized approximations of pure quantum states by k-sparse states or states with Schmidt rank k. However, there are various interesting questions for future work.

Application to tensor networks. We propose tensor network truncation as a natural application of randomized truncation. How useful this is in practice, and for *which* applications precisely, will require detailed numerical study, which we leave to follow-up work.

Computing the sampling weights. We give a provable runtime bound for finding μ in Theorem 8.6 but we conjecture that this can be improved significantly. One plausible conjecture which would imply this is Conjecture 8.10. These algorithms are discussed in more detail in Section 8.3

Applications beyond quantum states. We focus on the case where the vector we try to approximate is a quantum state. A natural follow-up direction is to investigate applications beyond this interpretation. This may require generalizing to approximations in norms besides the (Schatten) 1-norm, such as the operator norm or general p-norms. Moreover, it may not be natural to impose a normalization condition on the approximating vectors, which changes the optimization problem.

Relation to Hamiltonian simulation. As described in Section 1.3, there are several randomized Hamiltonian simulation methods that can be viewed as solving a related problem. Namely, given a Hamiltonian $H = \sum_{i=1}^{L} v_i h_i$, can we approximate e^{-iHt} with a random mixture of $\exp(-i\sum_i w_i h_i)$ for a collection of k-sparse vectors w? The qDRIFT algorithm can be viewed as doing this for 1-sparse w. Applied to a starting state $|\psi\rangle$, and assuming $h_i^2 \propto I$, this means approximating $e^{-iHt}|\psi\rangle$ with a mixture of superpositions of $|\psi\rangle$ and $h_i|\psi\rangle$. Applying our robustness-optimal scheme with $k = 2$, and neglecting higher-order terms, turns out to yield precisely qDRIFT. Improving this to higher values of k is a compelling open problem. One of the key difficulties is that qDRIFT addresses only first-order and second-order error while modern Hamiltonian simulation methods generally control higher-order errors as well. An exploration of methods for partially randomized Hamiltonian simulation is initiated in recent work [GWS+25].

1.6 Notation

For a positive integer d, we write $[d] = \{1, 2, \ldots, d\}$. The standard basis of $\mathbb{C}^d$ is denoted by $\{e_i : i \in [d]\}$. For any operator X on $\mathbb{C}^d$ we let $X^\dagger$ denote the Hermitian conjugate. We adopt the physics convention where the Hermitian inner product $\langle \cdot, \cdot \rangle$ is linear in the second argument, so that $\langle u, v \rangle = u^\dagger v$ for any $u, v \in \mathbb{C}^d$. We write $v = \sum_{i=1}^{d} v_i e_i$, where v_i denotes the coordinate of v in the standard basis. We let $|v_i^\downarrow|$ denote the ith largest element (in nonincreasing order) of the sequence $(|v_j| : j \in [d])$, and define $v^\downarrow := \sum_{i=1}^{d} |v_i^\downarrow| e_i$. For $v \in \mathbb{C}^d$ and $1 \le j \le d$, set

$$s_j^v := \sum_{i=j}^{d} |v_i^\downarrow|. \tag{12}$$

The set of linear operators on $\mathbb{C}^d$ is denoted $\mathcal{L}(\mathbb{C}^d)$, and the set of density matrices on $\mathbb{C}^d$ (positive semidefinite with unit trace) is denoted $\mathcal{D}(\mathbb{C}^d)$. We use $v, w, \ldots$ for vectors in $\mathbb{C}^d$ (which also represent pure states), and $\rho, \sigma, \ldots$ for density matrices in $\mathcal{D}(\mathbb{C}^d)$. Matrix norms are Schatten p-norms throughout. Given $v \in \mathbb{C}^d$ and $\mathcal{A} \subseteq [d]$, we let $v_\mathcal{A}$ denote the vector obtained by keeping

coordinates v_i for $i \in \mathcal{A}$ and setting the rest to zero. If $\mathcal{A}$ is a contiguous block $\mathcal{A}_{i,j} = \{i, i+1, \ldots, j\}$, we also write $v_{i:j} \coloneqq v_{\mathcal{A}_{i,j}}$. For $p \in \mathbb{Z}_{>0}$ we denote the ℓ_p norm of v by $\|v\|_p$, and $\|v\|_0$ is the number of nonzero coordinates of v in the standard basis. Finally, for a set S and scalar a, the rescaled set is $aS = \{as : s \in S\}$, and for a point g we write $g + S = \{g + s : s \in S\}$. Other notation will be introduced as needed.

2 Quantifying coherence and entanglement

Diagonal density matrices represent classical states and can be viewed as mixtures of 1-sparse pure states: the standard basis vectors. By considering mixtures of k-sparse states, we can generate increasingly non-classical states, as measured by the resource theory of coherence.

Definition 2.1 (k-sparse mixture, $\mathcal{I}_k$)**.** A pure state $v \in \mathbb{C}^d$ is *k-sparse* if $\|v\|_0 \leq k$. A mixed state $\rho \in \mathcal{D}(\mathbb{C}^d)$ is a *k-sparse mixture* if $\rho = \sum_j p_j u_j u_j^\dagger$ for some k-sparse pure states $u_j \in \mathbb{C}^d$ and probabilities $p_j > 0$, $\sum p_j = 1$[2]. We denote the set of all k-sparse mixtures by $\mathcal{I}_k(\mathbb{C}^d)$, or simply $\mathcal{I}_k$ when the Hilbert space is evident.

The notation $\mathcal{I}_k$ is in reference to previous work which refers to this as the set of "incoherent" states [JLP+18]. Although this definition is basis dependent, it is closely related to other basis-independent objects. For example, another familiar resource in quantum information is bipartite entanglement, which can be measured by the Schmidt rank. A bipartite state $v \in \mathbb{C}^a \otimes \mathbb{C}^b$ has a Schmidt decomposition given by

$$v = \sum_{i=1}^{r} \lambda_i \, x_i \otimes y_i, \tag{13}$$

for some positive integer $r \leq \min\{a, b\}$, orthonormal bases $\{x_i\} \subset \mathbb{C}^a$ and $\{y_i\} \subset \mathbb{C}^b$, and Schmidt coefficients $\lambda_i > 0$, $\sum \lambda_i^2 = 1$.

Definition 2.2 (Schmidt number k, $\mathcal{S}_k$)**.** A pure state v has *Schmidt rank k* if it has k nonzero Schmidt coefficients. A mixed state $\rho \in \mathcal{D}(\mathbb{C}^a \otimes \mathbb{C}^b)$ has *Schmidt number at most k* if $\rho = \sum_j p_j u_j u_j^\dagger$ for some pure states $u_j \in \mathbb{C}^a \otimes \mathbb{C}^b$ with Schmidt rank at most k and p a probability distribution. We denote the set of mixed states with Schmidt number at most k by $\mathcal{S}_k(\mathbb{C}^a, \mathbb{C}^b)$ or $\mathcal{S}_k$ when the Hilbert spaces are evident.

The Schmidt number for density matrices was originally introduced and studied in [TH00]. We next briefly recall standard quantum distance measures; see e.g. [NC12, Ch. 5]. The *fidelity* between states ρ, σ is $F(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$. If $\sigma = vv^\dagger$ is pure, then $F(\rho, \sigma) = \sqrt{\langle v, \rho v\rangle}$, and if $\rho = ww^\dagger$ is also pure we have $F(vv^\dagger, ww^\dagger) = |\langle v, w\rangle|$. The *trace distance* is $T(\rho, \sigma) = \frac{1}{2}\|\rho - \sigma\|_1$. When convenient, we abbreviate $F(v, w) := F(vv^\dagger, ww^\dagger)$ and similarly for T. We also use $F(v, \rho)$ and $T(v, \rho)$ for comparisons between a mixed state ρ and a pure state given as a unit vector v. For two pure states v, w these are related by $F(v, w)^2 + T(v, w)^2 = 1$. More generally, the Fuchs-van de Graaf inequalities relate F and T by

$$1 - F \leq T \leq \sqrt{1 - F^2}. \tag{14}$$

If one of the two states is pure, the lower bound can be improved, and we have

$$1 - F^2 \leq T \leq \sqrt{1 - F^2}. \tag{15}$$

[2] In more conventional quantum notation we write $\rho = \sum_j p_j \, |u_j\rangle\langle u_j|$.

We are interested in the maximum fidelity (or minimum trace distance) achievable by approximating a state with one that has limited coherence or entanglement. The lower bound in Eq. (15) implies that, for pure states, we cannot hope to attain more than a quadratic advantage using a mixed approximation. Here and throughout, we let $d, a, b \in \mathbb{Z}_{>0}$ such that $a, b \leq d$.

Definition 2.3 (Fidelity with $\mathcal{I}_k$ or $\mathcal{S}_k$)**.** For $\rho \in \mathcal{D}(\mathbb{C}^d)$, let $F_k(\rho) := \max_{\sigma \in \mathcal{I}_k} F(\rho, \sigma)$. For $\rho \in \mathcal{D}(\mathbb{C}^a \otimes \mathbb{C}^b)$, let $F_k^{(E)}(\rho) := \max_{\sigma \in \mathcal{S}_k} F(\rho, \sigma)$.

As in the case of $F(\cdot, \cdot)$, if v is a unit vector corresponding to a pure state we abbreviate $F_k(vv^\dagger)$ by $F_k(v)$ and similarly for $F_k^{(E)}(\cdot)$. These measures are straightforward to compute for pure states.

Lemma 2.4. *If $v \in \mathbb{C}^d$ is pure then for any positive integer $k \leq d$ it holds that $F_k(v)^2 = \sum_{i=1}^k |v_i^\downarrow|^2$.*

Similarly, if $v \in \mathbb{C}^a \otimes \mathbb{C}^b$ is a bipartite state with a Schmidt decomposition of the form in Eq. (13), we have the following.

Lemma 2.5. *If v is of the form given in Eq. (13) then for any positive integer $k \leq r$ it holds that $F_k^{(E)}(v)^2 = \sum_{i=1}^k \lambda_i^2$.*

Proof. By definition, $F_k^{(E)}(v)^2 = \max_{\sigma \in \mathcal{S}_k} \langle v, \sigma v \rangle$. The function $\sigma \mapsto \langle v, \sigma v \rangle$ is linear and the set $\mathcal{S}_k$ is convex so it suffices to consider a maximization over the extreme points of $\mathcal{S}_k$. Thus, $F_k^{(E)}(v)^2 = \max_\varphi |\langle \varphi, v \rangle|^2$ where the maximization is over pure states on $\mathbb{C}^a \otimes \mathbb{C}^b$ having Schmidt rank at most k. The optimal solution to this problem is attained by the state proportional to keeping the top k Schmidt coefficients, by Eckart-Young Theorem [EY36, Mir60]. □

These facts agree with the intuition that if we want to maximize the overlap with a k-sparse state, we should keep the k largest coordinates. We can also phrase this in terms of an appropriate norm, whose appearance in the context of truncation is natural.

Definition 2.6 (Top-k norm)**.** For $v \in \mathbb{R}^d$ and $1 \leq k \leq d$, the *top-k norm* is

$$\|v\|_{(k)} := \Big(\sum_{i=1}^k |v_i^\downarrow|^2 \Big)^{1/2}. \tag{16}$$

With this terminology, Lemmas 2.4 and 2.5 state that $F_k(v) = \|v\|_{(k)}$ and, if v is bipartite, $F_k^{(E)}(v) = \|\lambda\|_{(k)}$ where λ is a vector of Schmidt coefficients. The story is more interesting for optimal trace-distance approximations.

Definition 2.7 (Trace distance with $\mathcal{I}_k$ or $\mathcal{S}_k$)**.** For $\rho \in \mathcal{D}(\mathbb{C}^d)$, let $T_k(\rho) := \min_{\sigma \in \mathcal{I}_k} T(\rho, \sigma)$. For $\rho \in \mathcal{D}(\mathbb{C}^a \otimes \mathbb{C}^b)$, let $T_k^{(E)}(\rho) := \min_{\sigma \in \mathcal{S}_k} T(\rho, \sigma)$.

Finally, we recall the *robustness* of coherence and entanglement [VT99, JLP+18], which can be thought of as measuring a multiplicative approximation error by states in $\mathcal{I}_k$ or $\mathcal{S}_k$.

Definition 2.8 (Robustness with $\mathcal{I}_k$ or $\mathcal{S}_k$)**.** For $\rho \in \mathcal{D}(\mathbb{C}^d)$,

$$R_k(\rho) := \min\Big\{ s \geq 0 : \frac{\rho + s\sigma}{1+s} \in \mathcal{I}_k,\ \sigma \in \mathcal{D}(\mathbb{C}^d), \Big\}. \tag{17}$$

For $\rho \in \mathcal{D}(\mathbb{C}^a \otimes \mathbb{C}^b)$,

$$R_k^{(E)}(\rho) := \min\Big\{ s \geq 0 : \frac{\rho + s\sigma}{1+s} \in \mathcal{S}_k,\ \sigma \in \mathcal{D}(\mathbb{C}^a \otimes \mathbb{C}^b) \Big\}. \tag{18}$$

The robustness gives a bound on the trace distance through

$$T_k(\rho) \leq \frac{R_k(\rho)}{1 + R_k(\rho)} \tag{19}$$

which follows by taking the trace distance between ρ and the optimum in Eq. (17). There is an alternative notion of the robustness which arises when one restricts the feasible set in the minimization in Definition 2.8 to density matrices σ which are themselves in $\mathcal{I}_k$, or $\mathcal{S}_k$. However, in [JLP+18] (building on [AFS12, Reg17]) it was proven that for pure states, these notions coincide. Moreover, in this case the robustness can be computed using the dual norm to the top-k norm. We now define this norm, and collect some of its relevant properties. Below, we write $\binom{\mathcal{A}}{k}$ for the family of k-element subsets of a finite set $\mathcal{A}$.

Definition 2.9 (k-support norm)**.** The *k-support norm* is the dual of the top-k norm. An equivalent definition is

$$\|v\|_{(k,*)} \;:=\; \min\Big\{ \sum_{S \in \binom{[d]}{k}} \|w(S)\|_2 \;:\; v = \sum_{S \in \binom{[d]}{k}} w(S) \Big\} \tag{20}$$

for any $v \in \mathbb{R}^d$ and $1 \leq k \leq d$. Here we minimize over collections of vectors $\{w(S)\}_S$ such that the support of each $w(S)$ is contained in S.

This norm was introduced in the context of sparse prediction problems in [AFS12]. See also [Reg17] and [Bha97, Ch. 4] for additional background. That $\|\cdot\|_{(k)}$ and $\|\cdot\|_{(k,*)}$ are dual to one another follows from the observation that the k-support norm is a gauge function for the set

$$C_k \;:=\; \mathrm{conv}\big\{\, w \in \mathbb{R}^d :\; \|w\|_0 = k,\; \|w\|_2 \leq 1 \,\big\}, \tag{21}$$

This implies C_k is the unit ball for the k-support norm, and the top-k norm is indeed the maximum inner product with a vector from this set. For $k = 1$ and $k = d$ one recovers the ℓ_1 and ℓ_2 norms, respectively. Eqs. (20) and (21) do not suggest an efficient algorithm for computing $\|v\|_{(k,*)}$. However, there is a closed-form expression for $\|v\|_{(k,*)}$ which can be computed in $O(d \log d)$ time.

Lemma 2.10 ([AFS12, Prop. 2.1])**.** *For every $v \in \mathbb{R}^d$,*

$$\|v\|^2_{(k,*)} \;=\; \sum_{i=1}^{k-r-1} |v^{\downarrow}_i|^2 \;+\; \frac{\big(s^v_{k-r}\big)^2}{r+1}, \tag{22}$$

where $r \in \{0, 1, \ldots, k-1\}$ is the unique[3] index satisfying

$$|v^{\downarrow}_{k-r-1}| \;>\; \frac{s^v_{k-r}}{r+1} \;\geq\; |v^{\downarrow}_{k-r}| \tag{23}$$

(with the convention $|v^{\downarrow}_0| := +\infty$), and where s^v_j is the partial sum defined in Eq. (12).

The following optimization view underlies Lemma 2.10. For $v \in \mathbb{R}^d$, the optimal value for

$$\begin{aligned} \max.\quad & \langle v, x \rangle - \tfrac{1}{2}\|x\|^2_{(k)} \\ \text{s.t.}\quad & x \in \mathbb{R}^d \end{aligned} \tag{24}$$

is equal to $\frac{1}{2}\|v\|^2_{(k,*)}$. This is a standard fact about dual norms, stated explicitly in Lemma A.8. (See Section A.2 for a brief overview of the convex analysis relevant in this work.) Lemma 2.10 is then an easy consequence of the following.

[3]The uniqueness of the choice of r satisfying the condition in Eq. (23) is stated as Lemma 3.1 in [MF85].

Lemma 2.11. *Let $v \in \mathbb{R}^d$ and let ℓ be the smallest index with $v^\downarrow_\ell = 0$ (set $\ell = d+1$ if none). Assume $\ell \geq k+1$. A vector x is optimal for the maximization in Eq. (24) if and only if, with r as in Lemma 2.10,*

$$x^\downarrow_j = \begin{cases} v^\downarrow_j, & j = 1, \ldots, k-r-1, \\ \dfrac{s^v_{k-r}}{r+1}, & j = k-r, \ldots, \ell-1. \end{cases} \tag{25}$$

The proof is essentially shown in the proof of Proposition 2.1 in [AFS12]. We give the full argument in Appendix B.

For pure states, the robustness is directly related to the k-support norm. This was proven in [Reg17] (see also [JLP+18]) and our construction in Section 5 also yields a simple alternate proof.

Theorem 2.12 (R_k for pure states [Reg17])**.** *If $v \in \mathbb{R}^d$ is a pure state,*

$$R_k(v) = \|v\|^2_{(k,*)} - 1. \tag{26}$$

Also, if $v \in \mathbb{C}^a \otimes \mathbb{C}^b$ has a vector of Schmidt coefficients λ, $R^{(E)}_k(v) = \|\lambda\|^2_{(k,*)} - 1$.

For our applications we will usually want to calculate $\|\cdot\|_{(k,*)}$ precisely. But an $O(1)$-approximation is possible using the 1- and 2-norms. From [AFS12] we have that for a unit vector v,

$$\frac{\|v\|_1}{\sqrt{k}} \leq \|v\|_{(k,*)} \leq \sqrt{2}\frac{\|v\|_1}{\sqrt{k}}, \tag{27}$$

which in particular gives $R_k(v) \geq \frac{\|v\|_1^2}{k} - 1$.

Sorting entries. A useful symmetry is that the distance measures we use, as well as our notions of sparsity, are invariant under permutation and phases. Specifically, for the three distance measures introduced in this section and, for any $v \in \mathbb{C}^d$,

$$F_k(v) = F_k(v^\downarrow), \quad T_k(v) = T_k(v^\downarrow), \quad \text{and} \quad R_k(v) = R_k(v^\downarrow). \tag{28}$$

Thus in all computations, we can assume without loss of generality that the entries are real, nonnegative, and ordered. Obtaining the optimal solutions for the optimization problems defining each of these quantities is then a straightforward matter of tracking the phases and ordering of entries in the original vector v, and we neglect this detail throughout this work for the sake of clarity.

3 Computational hardness for mixed states

How hard is it to compute approximations by sparse states or states with small Schmidt numbers? For an arbitrary state $\rho \in \mathcal{D}(\mathbb{C}^d)$ the problem is computationally intractable. In fact, we will see that even weak membership of $\mathcal{I}_k$ and $\mathcal{S}_k$ is NP-hard[4]. As a consequence, computing $F_k(\rho)$, $F^{(E)}_k(\rho)$, $T_k(\rho)$, and $T^{(E)}_k(\rho)$ is NP-hard in general.

For entanglement, weak membership of $\mathcal{S}_1$ is precisely the separability problem, which is known to be NP-complete [Gur03, Gha08]. In contrast, checking 1-sparsity is trivial: $\rho \in \mathcal{I}_1$ if and only if ρ is diagonal in the standard basis. However, for larger k, the problem again becomes NP-hard.

[4]Weak membership of a convex set S is the task of deciding whether ρ is in the ε-interior or ε-exterior of S in Euclidean distance. Weak linear optimization is the task of approximately solving $\max_{\sigma \in S} \operatorname{tr}[M\sigma]$ up to additive error ε. These terms are defined in detail in [GLS12].

Proposition 3.1. *Given $\rho \in \mathcal{D}(\mathbb{C}^d)$ and k, the problem of weak membership of ρ in $\mathcal{I}_k$ is* NP*-complete.*

Proof. By Carathéodory's theorem, any element in $\mathcal{I}_k$ can be written as a mixture of at most d^2+1 k-sparse states. This shows the existence of an efficiently verifiable witness for weak membership, so the problem is in NP. To prove hardness, we use the fact that weak membership, weak separation, and weak linear optimization are polynomial-time equivalent for convex sets [GLS12]. Thus it suffices to consider weak linear optimization over $\mathcal{I}_k$, i.e.

$$m_* \;=\; \max_{\sigma \in \mathcal{I}_k} \operatorname{tr}[M\sigma]. \tag{29}$$

The maximum is achieved at an extreme point of $\mathcal{I}_k$, namely a pure k-sparse state. Equivalently,

$$m_* \;=\; \max_{\substack{S \subseteq [d] \\ |S|=k}} \|M_S\|_\infty, \tag{30}$$

where M_S is the $k \times k$ principal submatrix of M indexed by S.

We consider the case where $M = A_G$ is the adjacency matrix of a graph G on d vertices. Then each M_S is the adjacency matrix A_H of the induced subgraph $H = G[S]$. For any k-vertex graph H, we have $\|A_H\|_\infty \leq k-1$, with equality if and only if H is the complete graph K_k. Moreover Proposition 1.3.10 of [CRS10] implies that if H is not complete then $\|A_H\|_\infty$ is bounded away from $k-1$ by an inverse polynomial in k. Thus the k-Clique problem reduces to weak linear optimization over $\mathcal{I}_k$, proving NP-hardness. □

We emphasize that this hardness requires k to be a variable. If k is fixed then $\mathcal{I}_k$ can be expressed as a semidefinite program of size $d^{O(k)}$, since any $\sigma \in \mathcal{I}_k$ can be expressed as a mixture of states in $\mathcal{D}(\mathbb{C}^S)$ for S ranging over all subsets of $[d]$ of size k. Here $\mathbb{C}^S$ means the $|S|$-dimensional subspace of states with support contained in S. The hardness results shows that this SDP is essentially optimal if finding k-cliques in d-vertex graphs takes time $d^{\Omega(k)}$. This latter conjecture is implied by the Exponential Time Hypothesis and other plausible complexity-theory conjectures [ABW15].

For *pure* states, membership in $\mathcal{I}_k$ and $\mathcal{S}_k$ is easy to decide: in the first case by support size, in the second by Schmidt rank. We have already seen that for pure v, the fidelities $F_k(v)$ and $F_k^{(E)}(v)$ are easy to compute. In Section 4 we will show that the trace distances $T_k(v)$ and $T_k^{(E)}(v)$ can also be computed efficiently, although these algorithms are considerably more complicated than those for the fidelity.

Our algorithms for $T_k(v)$ and $R_k(v)$ thus rely crucially on v being a pure state and involve details of the properties of trace distance and robustness. For a general distance measure on quantum states $D(\rho, \sigma)$, one can define $D(v, \mathcal{I}_k) := \min_{\sigma \in \mathcal{I}_k} D(vv^\dagger, \sigma)$. We do not have a general reduction from $D(v, \mathcal{I}_k)$ to the pairwise distances $D(\rho, \sigma)$, nor do we know examples where the pairwise distances are easy but calculating $D(v, \mathcal{I}_k)$ is hard.

4 Approximations in trace distance

In this section we give efficient algorithms to compute the optimal randomized truncation to a pure state with respect to trace distance. More specifically, given a target unit vector $v \in \mathbb{C}^d$, our first algorithm, given in Section 4.1, computes the value

$$T_k(v) := \min_{\sigma \in \mathcal{I}_k} \frac{1}{2}\|vv^\dagger - \sigma\|_1 = \min_{\sigma \in \mathcal{I}_k} \max_{0 \preceq M \preceq \mathbb{1}} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big). \tag{31}$$

It also returns a description of a measurement $M^\star = mm^\dagger$ with $m \in \mathbb{C}^d$ such that $(\sigma^\star, M^\star)$ is a saddle point, for some $\sigma^\star \in \mathcal{I}_k$. A second algorithm, given in Section 4.3, then returns a description of the optimal solution $\sigma^\star$, either explicitly in the standard basis or in the form of an ensemble of k-sparse pure states $\{(p_i, v_i)\}_i$ where the distribution p can be efficiently sampled.

4.1 Computing the trace distance and optimal measurement

The main result of this section is:

Theorem 4.1. *Let $k, d \in \mathbb{Z}_{>0}$ such that $k \leq d$ and $v \in \mathbb{C}^d$ be a unit vector. There is an algorithm (Algorithm 1) running in time $O(dk + d \log d)$ that computes $T_k(v)$ as well as a measurement operator $M^\star$ such that $(\sigma^\star, M^\star)$ is a saddle point of the minimax in Eq. (31), for some $\sigma^\star \in \mathcal{I}_k$.*

Our proof of Theorem 4.1 will proceed by first showing that the optimal M can be taken to be of the form $mm^\dagger$, i.e. a rank-1 measurement. Then we derive the structure of m, proving that it consists of two pieces that are proportional to the entries of v with a flat region in the middle (when sorted). At this point, there is only a single parameter and it can be found by solving a cubic equation. Then in Section 4.3 we will use this to find an optimal approximating ensemble of k-sparse states. The three regions in m correspond to three regions in our random k-sparse approximation w: the first region is proportional to the entries of v, the second region is uniform over a set (from which we will later importance sample), and the third region is zero.

The fact that M can be taken without loss of generality to be rank-one is due to the following lemma.

Lemma 4.2. *For any $d \in \mathbb{Z}_{>0}$, density matrix $\sigma \in \mathcal{D}(\mathbb{C}^d)$, and unit vector $v \in \mathbb{C}^d$, the optimum in $\max_{0 \preceq M \preceq \mathbb{1}} \operatorname{tr}\big(M(vv^\dagger - \sigma)\big)$ is attained by a rank-one projection operator.*

Proof. By Fact A.10, which is a standard consequence of Weyl's Inequality, the largest eigenvalue of the matrix $vv^\dagger - \sigma$ lies in the interval $[0, 1]$, while the second largest eigenvalue lies in $[-1, 0]$. If $\sigma = vv^\dagger$ then the optimal value of the maximization is zero, and $M = mm^\dagger$ is an optimal solution for any unit vector $m \in \mathbb{C}^d$. Otherwise, $vv^\dagger - \sigma$ has just a single positive eigenvalue, and an optimal solution is given by the rank-one projector onto this eigenspace. □

We also have the following related lemma.

Lemma 4.3. *For any $d \in \mathbb{Z}_{>0}$, density matrix $\sigma \in \mathcal{D}(\mathbb{C}^d)$, and unit vector $v \in \mathbb{C}^d$, it holds that*

$$\max_{0 \preceq M \preceq \mathbb{1}} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big) = \max_{\rho \in \mathcal{D}(\mathbb{C}^d)} \operatorname{tr}\Big(\rho(vv^\dagger - \sigma)\Big). \tag{32}$$

Moreover, if $\sigma \neq vv^\dagger$, then the optimum for the maximization in the right-hand side of Eq. (32) is rank-one and unique.

Proof. The $\geq$ direction of Eq. (32) holds because $\mathcal{D}(\mathbb{C}^d) \subset \{M \in \mathbb{C}^{d \times d} : 0 \preceq M \preceq \mathbb{1}\}$. For the $\leq$ direction, the left-hand side of Eq. (32) is maximized by a rank-one solution, due to Lemma 4.2, and this is also a feasible solution for the maximization on the right-hand side. This proves Eq. (32).

The uniqueness part of the claim follows from the fact that, as shown in the proof of Lemma 4.2, if $\sigma \neq vv^\dagger$ then the matrix $A := vv^\dagger - \sigma$ has a single positive eigenvalue $\lambda_1(A) > 0$ which is the optimal value for the maximization. Indeed, let P be the orthogonal projection onto this eigenspace, and set $Q := \mathbb{1} - P$. Then for any $\rho \in \mathcal{D}(\mathbb{C}^d)$ we have

$$\operatorname{tr}(\rho A) = \operatorname{tr}(\rho P A) + \operatorname{tr}(\rho Q A) \leq \operatorname{tr}(\rho P A) = \lambda_1(A) \operatorname{tr}(\rho P) \tag{33}$$

and the right-hand side of the above is strictly less than $\lambda_1(A)$ unless $\rho = P$. □

The proof of the next lemma uses the above results as well as some of the facts in Section A.1, following a similar reasoning to the proof of Lemma 3 in [AKT24a].

Lemma 4.4. *Let $k, d \in \mathbb{Z}_{>0}$ such that $k \leq d$ and $v \in \mathbb{R}^d$ be a unit vector such that $v_1 \geq v_2 \geq \cdots \geq v_d > 0$. Then $T_k(v)$ is equal to*

$$\begin{aligned} \max \quad & \langle m, v\rangle^2 - \|m\|_{(k)}^2 \\ \text{s.t.} \quad & m \in \mathbb{R}^d \\ & \|m\|_2 = 1. \end{aligned} \tag{34}$$

Moreover, the minimax problem defining $T_k(v)$ in Eq. (31) has a saddle point $(\sigma^\star, M^\star)$ for some $\sigma^\star \in \mathcal{I}_k$ and $M^\star = m^\star (m^\star)^\dagger$ where $m^\star$ is the optimal solution to the maximization in Eq. (34).

Proof. If $k = d$ then $T_k(v) = 0$ and the objective function in Eq. (34) is just $m \mapsto \langle m, v\rangle^2 - 1$, which clearly has an optimal value of zero. Also, the statement about saddle points is trivial in this case. Hence, from now on we assume $k < d$. Let $\mathcal{M}$ denote the set of Hermitian matrices $M \in \mathbb{C}^{d\times d}$ such that $0 \preceq M \preceq \mathbb{1}$ and $\mathcal{M}_1$ the subset of such matrices which are rank-one projection operators. We have

$$\max_{M\in\mathcal{M}} \min_{\sigma\in\mathcal{I}_k} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big) = \min_{\sigma\in\mathcal{I}_k} \max_{M\in\mathcal{M}} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big) \tag{35}$$

$$= \min_{\sigma\in\mathcal{I}_k} \max_{\rho\in\mathcal{D}(\mathbb{C}^d)} \operatorname{tr}\Big(\rho(vv^\dagger - \sigma)\Big) \tag{36}$$

$$= \max_{\rho\in\mathcal{D}(\mathbb{C}^d)} \min_{\sigma\in\mathcal{I}_k} \operatorname{tr}\Big(\rho(vv^\dagger - \sigma)\Big), \tag{37}$$

where in the first and third lines we used Sion's Minimax Theorem (Theorem A.1), and in the second line we used Lemma 4.3. By Lemma A.3 there exists a saddle point $(\sigma^\star, \rho^\star) \in \mathcal{I}_k \times \mathcal{D}(\mathbb{C}^d)$ for the minimax in Eq. (36) such that

$$\operatorname{tr}\Big(\rho^\star(vv^\dagger - \sigma^\star)\Big) = \max_{\rho\in\mathcal{D}(\mathbb{C}^d)} \operatorname{tr}\Big(\rho(vv^\dagger - \sigma^\star)\Big). \tag{38}$$

Appealing once again to Lemma 4.3, the maximum in Eq. (38) is uniquely attained by a rank-one projection operator because $k < d$, which implies $\sigma^\star \neq vv^\dagger$. Hence, $\rho^\star \in \mathcal{M}_1$ and the right-hand side of Eq. (37) is equal to

$$\max_{M\in\mathcal{M}_1} \min_{\sigma\in\mathcal{I}_k} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big) = \max_{m\in\mathbb{R}^d:\ \|m\|_2=1} \left\{ |\langle m, v\rangle|^2 - \max_{\sigma\in\mathcal{I}_k} \langle m, \sigma m\rangle \right\} \tag{39}$$

$$= \max_{m\in\mathbb{R}^d:\ \|m\|_2=1} \left\{ |\langle m, v\rangle|^2 - \sum_{i=1}^k |m_i^\downarrow|^2 \right\}, \tag{40}$$

where the second line follows from Lemma 2.4. This proves the first part of the claim.

For the second part, it suffices to show that $(\sigma^\star, \rho^\star)$ is also a saddle point for the minimax in Eq. (35). But this is clear by comparing the left-hand side of Eq. (39) with that of Eq. (35), which we have shown to be equal. Explicitly, we have

$$\rho^\star \in \arg\max_{M\in\mathcal{M}} \min_{\sigma\in\mathcal{I}_k} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big) \quad \text{and} \quad \sigma^\star \in \arg\min_{\sigma\in\mathcal{I}_k} \max_{M\in\mathcal{M}} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big) \tag{41}$$

where the second inclusion follows from Lemma 4.2. □

The optimization in Eq. (34) bears some resemblance to the computation of the proximal operator for the k-support norm, as studied in [AFS12], and can be solved in a similar way. To begin executing this strategy, we require the following fact.

Lemma 4.5. *A point $m \in \mathbb{R}^d$ is an optimal solution for the optimization problem in Eq. (34) with $k < d$ if and only if there exists a $\lambda \geq 0$ such that*

$$\|u\|_{(k,*)}^2 + \|m\|_{(k)}^2 = 2\langle u, m\rangle \quad \text{where} \quad u \coloneqq \langle v, m\rangle v - \lambda m. \tag{42}$$

Proof. We make use of basic facts from convex analysis, given in Section A.2. The optimization problem in Eq. (34) is equivalent to minimizing $f(m)$ over all $m \in \Omega$, where $f(m) \coloneqq -\frac{1}{2}\langle m, v\rangle^2 + \frac{1}{2}\|m\|_{(k)}^2$ and $\Omega \coloneqq \{m \in \mathbb{R}^d : \|m\|_2 \leq 1\}$. Relaxing the norm constraint here is valid by homogeneity of the objective function in Eq. (34), along with the fact that the optimal value is greater than zero when $k < d$. By Lemma A.7 the point m is an optimal solution if and only if $0 \in \partial f(m) + \mathcal{N}_\Omega(m)$, where $\partial f(m)$ is the subdifferential of f at m and $\mathcal{N}_\Omega(m)$ is the normal cone at m. When m is a unit vector, the normal cone $\mathcal{N}_\Omega(m)$ is just the ray generated by the vector m. Also, the subdifferential can be computed as $\partial f(m) = -\langle m, v\rangle v + \partial\big(\frac{1}{2}\|\cdot\|_{(k)}^2\big)(m)$. Therefore the optimality condition for a given unit vector m is equivalent to

$$\langle v, m\rangle v - \lambda m \in \partial\big(\frac{1}{2}\|\cdot\|_{(k)}^2\big)(m) \tag{43}$$

for some $\lambda \geq 0$. The equivalence between this condition and Eq. (42) is a direct consequence of Corollary A.9. □

We are now ready to show that the optimal measurement vector m has a simple functional form. An example of such an m is given in Figure 3a. This restricted form will allow us to efficiently compute it, as well as the optimal trace distance value. We first establish some notation.

Definition 4.6. Let $r < k < \ell \leq d$ be nonnegative integers and $v \in \mathbb{C}^d$ be a unit vector. For any $\lambda \geq 0$ we set

$$\theta_{r,\ell}^v(\lambda) \coloneqq \frac{s_{k-r}^v - s_\ell^v}{r + 1 + (\ell - k + r)\lambda} \tag{44}$$

where s_j^v is defined as in Eq. (12).

This definition excludes the case where $k = d$, since in that case the v is already d-sparse and the trace distance is 0. Also define $v_0 \coloneqq +\infty$ and $v_{d+1} \coloneqq 0$.

Lemma 4.7. *Let $k, d \in \mathbb{Z}_{>0}$ such that $k < d$ and $v \in \mathbb{R}^d$ be a unit vector such that $v_1 \geq v_2 \geq \cdots \geq v_d > 0$. A solution $m \in \mathbb{R}^d$ to the optimization problem in Eq. (34) is optimal if and only if:*

1. *$m = \frac{\widetilde{m}}{\|\widetilde{m}\|_2}$ where*

$$\widetilde{m}_i = \begin{cases} \frac{v_i}{1+\lambda}, & i = 1, \ldots, k-r-1 \\ \theta_{r,\ell}^v(\lambda), & i = k-r, \ldots, \ell-1 \\ \frac{v_i}{\lambda}, & i = \ell, \ldots, d \end{cases} \tag{45}$$

for some $r \in \{0, 1, \ldots, k-1\}$, $\ell \in \{k+1, \ldots, d+1\}$, and $\lambda > 0$ such that

$$\theta_{r,\ell}^v(\lambda) \in \frac{1}{1+\lambda}[v_{k-r}, v_{k-r-1}) \cap \frac{1}{\lambda}(v_\ell, v_{\ell-1}], \tag{46}$$

and

2. $\langle v, \widetilde{m} \rangle = 1$.

Moreover, for such a unit vector m, it holds that $\lambda = T_k(v)$.

Proof. Since $k < d$ and v has no entries equal to zero then complementary slackness implies that $\lambda > 0$. Specifically, since λ is a Lagrange multiplier for the constraint $\|m\|_2 \leq 1$, if $\lambda = 0$ then there is an m with norm less than one which attains the optimal value. But this can only happen if that value is 0.

Next, by Lemma 4.5, m is optimal for Eq. (34) if and only if $\|m\|_2 = 1$ and there is a $\lambda > 0$ such that $\|u\|^2_{(k,*)} + \|m\|^2_{(k)} = 2\langle u, m \rangle$ where $u = \langle v, m \rangle v - \lambda m$. Let us assume that m is nonnegative and sorted in nonincreasing order; we will justify this assumption shortly. By Corollary A.9, the unit vector m is optimal for the problem

$$\begin{aligned} \max \quad & \langle u, m' \rangle - \frac{1}{2}\|m'\|^2_{(k)} \\ \text{s.t.} \quad & m' \in \mathbb{R}^d. \end{aligned} \tag{47}$$

Here, we can assume u is also nonnegative and sorted without loss of generality. Indeed, if $u_j < u_{j+1}$ for some index $j \in [d-1]$, we must have $m_j = m_{j+1}$. Otherwise, m is not optimal for the maximization in Eq. (34). But then $v_j < v_{j+1}$ from the definition of u, which contradicts the assumption that v is sorted.

Let us first prove necessity of the two conditions in Lemma 4.5. Suppose ℓ is the least positive integer such that $u_\ell = 0$, with $\ell := d+1$ if such an integer does not exist, and r is the unique integer such that

$$u_{k-r-1} > \frac{s^u_{k-r}}{r+1} \geq u_{k-r}. \tag{48}$$

Observe that we must have $\ell \geq k+1$; otherwise, $u_j = m_j$ for all $j \in [d]$ is an optimal solution to the problem in Eq. (47). But then this implies $u_j = \langle v, u \rangle v_j/(1+\lambda)$ for all $j \in [d]$, and therefore $v_\ell = 0$, contradicting the assumption that v has no entries equal to zero. Next, by the definition of ℓ, we have $m_j = \langle v, m \rangle v_j/\lambda$ for every $j \in \{\ell, \ldots, d\}$. By Lemma 2.11, m must also satisfy $m_j = u_j$ for every $j \in \{1, \ldots, k-r-1\}$ and $m_j = s^u_{k-r}/(r+1)$ for every $j \in \{k-r, \ldots, \ell-1\}$. The first of these relations is equivalent to $m_j = \langle v, m \rangle v_j/(1+\lambda)$ for every $j \in \{1, \ldots, k-r-1\}$. The second implies

$$m_{k-r} = \frac{1}{r+1}\left(\sum_{j=k-r}^{\ell-1} \langle v, m \rangle v_j - \lambda m_j\right) = \frac{1}{r+1}\left(\langle v, m \rangle (s^v_{k-r} - s^v_\ell) - \lambda(\ell - k + r) m_{k-r}\right) \tag{49}$$

which is true only if $m_j = \langle m, v \rangle \theta^v_{r,\ell}(\lambda)$ for each $j \in \{k-r, \ldots, \ell-1\}$. Furthermore, $\theta^v_{r,\ell}(\lambda)$ must satisfy the condition in Eq. (46) for m to be sorted in nonincreasing order. (Some of the inequalities are strict because of the definitions of r and ℓ.) Thus, we necessarily have $m = \langle v, m \rangle \widetilde{m}$ at the optimum, where $\widetilde{m}$ is of the same form as in Eq. (45). Additionally, since m is a unit vector, we obtain $\langle v, m \rangle = \|\widetilde{m}\|_2^{-1}$, which implies the second condition.

We now return to the assumption that m is sorted. We can justify this assumption in an identical manner to the final step in the proof of Lemma 2.11 in Appendix B. Suppose there is an optimal m for which $m_j < m_{j+1}$ for some $j \in [d-1]$. Then we must have $v_j = v_{j+1}$ and, consequently, the solution obtained by permuting the indices j and $j+1$ in m is also optimal; otherwise, the objective function in Eq. (34) is less than that obtained by such a permutation. But this contradicts the necessary condition on the form of an optimal sorted m derived above. For

example, for $j \in [k-r-2]$ we have shown above that $m_j = m_{j+1}$ when $v_j = v_{j+1}$. Other cases may be handled similarly.

To see that the two conditions are sufficient, we compute

$$s^u_{k-r} = \sum_{i=k-r}^{\ell-1} \left(\langle v, m \rangle v_i - \lambda \frac{\widetilde{m}_i}{\|\widetilde{m}\|_2} \right) \tag{50}$$

$$= \langle v, m \rangle \left(s^v_{k-r} - s^v_\ell - \lambda(\ell - k + r)\theta^v_{r,\ell}(\lambda) \right) \tag{51}$$

$$= \langle v, m \rangle (r+1)\theta^v_{r,\ell}(\lambda). \tag{52}$$

Thus,

$$u_{k-r} = \langle v, m \rangle (v_{k-r} - \lambda\theta^v_{r,\ell}) \leq \langle v, m \rangle \frac{v_{k-r}}{1+\lambda} \leq \langle v, m \rangle \theta^v_{r,\ell}(\lambda) = \frac{s^u_{k-r}}{r+1} \tag{53}$$

where the first equality uses Condition 2 in the statement of the lemma, both inequalities follow from Eq. (46) and the final equality is implied by Eq. (52). Similarly, from $\theta^v_{r,\ell}(\lambda) < \frac{1}{1+\lambda} v_{k-r-1}$, we have

$$u_{k-r-1} = \langle v, m \rangle \left(v_{k-r-1} - \lambda \frac{v_{k-r-1}}{1+\lambda} \right) = \langle v, m \rangle \frac{v_{k-r-1}}{1+\lambda} > \langle v, m \rangle \theta^v_{r,\ell}(\lambda) = \frac{s^u_{k-r}}{r+1}. \tag{54}$$

Noting that the constraints in Eq. (46) also ensure that m is sorted in nonincreasing order, we may write

$$\|u\|^2_{(k,*)} = \sum_{i=1}^{k-r-1} u_i^2 + \frac{(s^u_{k-r})^2}{r+1} \tag{55}$$

$$= \langle v, m \rangle^2 \left(\frac{\|v_{1:k-r-1}\|_2^2}{(1+\lambda)^2} + (r+1)\theta^v_{r,\ell}(\lambda)^2 \right) \tag{56}$$

$$= \|\widetilde{m}\|_2^{-2} \|\widetilde{m}\|^2_{(k)} \tag{57}$$

$$= \|m\|^2_{(k)}, \tag{58}$$

where in the second line we used the fact that $u_i = m_i = \frac{\langle v, m \rangle v_i}{1+\lambda}$ for every $i \in [k-r-1]$. By the above and Lemma 4.5, if $\|m\|^2_{(k)} = \langle u, m \rangle$ then m is optimal. Since $m_i = \langle v, m \rangle \theta^v_{r,\ell}(\lambda)$ for $i \in \{k-r, \ldots, \ell-1\}$ we have

$$\langle u, m \rangle = \sum_{i=1}^{k-r-1} m_i^2 + \langle v, m \rangle \theta^v_{r,\ell}(\lambda) s^u_{k-r} = \sum_{i=1}^{k-r-1} m_i^2 + \langle v, m \rangle^2 (\theta^v_{r,\ell}(\lambda))^2 (r+1) = \|m\|^2_{(k)} \tag{59}$$

where the second equality follows from Eq. (52).

Finally, it is easy to see that $\lambda = T_k(v)$ for an optimal solution m of this form, since

$$\langle v, m \rangle^2 - \|m\|^2_{(k)} = \langle v, m \rangle^2 - \langle u, m \rangle = \lambda. \tag{60}$$

This concludes the proof. □

Algorithm 1 Optimal trace distance & measurement

Input: Description of unit vector $v \in \mathbb{C}^d$; $k \in \{1, \ldots, d\}$
Output: $T_k(v)$ (optimal trace distance); m (optimal distinguishing vector)

1: **if** $k = d$ **then**
2: **return** $T_k \leftarrow 0$ and $m \leftarrow v$
3: **end if**
4: **for** $r \in \{0, \ldots, k-1\}$, $\ell \in \{k+1, \ldots, d+1\}$ **do**
5: $\Lambda \leftarrow$ positive solutions to the cubic equation in λ implied by Eq. (61)
6: **for** $\lambda \in \Lambda$ **do**
7: **if** Eq. (46) holds for these values of r, ℓ, and λ **then**
8: $T_k \leftarrow \lambda$
9: Set $\widetilde{m}$ as in Eq. (45)
10: $m \leftarrow \widetilde{m}/\|\widetilde{m}\|_2$
11: Go to Line 15
12: **end if**
13: **end for**
14: **end for**
15: Re-order and re-introduce phases to m
16: **return** T_k and m

Algorithm description. Condition 2 in Lemma 4.7, which pertains to the norm of the vector $\widetilde{m}$ in that lemma, is equivalent to

$$1 = \frac{\|v_{1:k-r-1}\|_2^2}{1+\lambda} + \frac{\|v_{k-r:\ell-1}\|_1^2}{r+1+(\ell-k+r)\lambda} + \frac{\|v_{\ell:d}\|_2^2}{\lambda}, \tag{61}$$

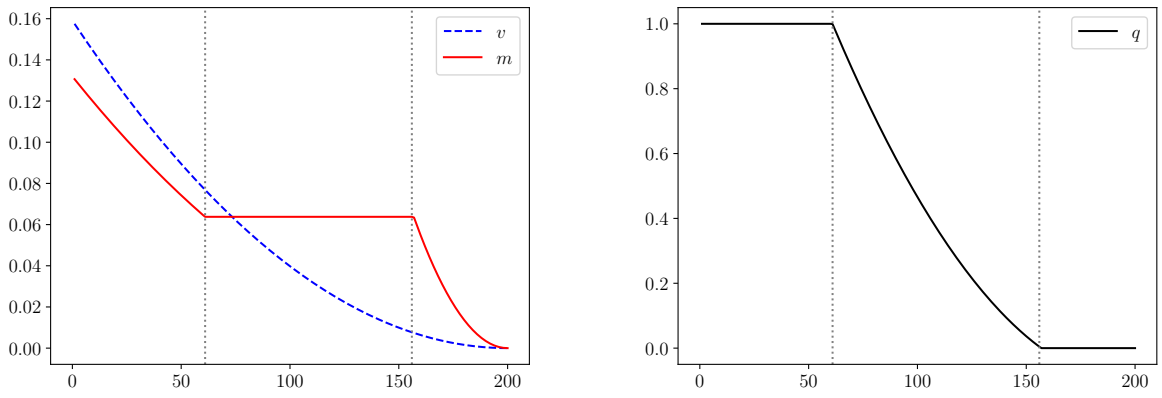
a cubic equation in terms of λ. This leads naturally to an algorithm for computing the trace distance and optimal measurement, given in Algorithm 1.

A direct inspection of Algorithm 1 leads to a runtime on the order of $d \log d + k(d-k)d$. The first term is due to the runtime of sorting the vector v. The second term is due to the $k(d-k)$ for-loop iterations. Within each of these iterations, the runtime is dominated by the computation of the constant (not λ) terms involved in Eq. (61), each of which can be computed in time on the order of d. By a straightforward memoization of the partial sums and norms — i.e., keeping track of the intermediate values of the sums between the for-loop iterations — we can remove this order d overhead in the second term. This proves Theorem 4.1.

An interesting question is whether the runtime could be further improved to $\tilde{O}(d)$, perhaps by restricting the range of possible r, ℓ. However, the current runtime is already less than what is needed for other steps: the singular value decomposition is used for truncating entangled states or matrix-product states and needs time d^3 or $d^{2.37\cdots}$, and even the heuristic version of Theorem 8.5 needs time dk^2.

4.2 Subset sampling

When we want to compute and sample the optimal density matrix, we need to solve a certain sampling problem. We here describe this problem, and postpone a detailed discussion to Section 8. Consider a set $[n] = \{1, 2, \ldots, n\}$. We want to sample to subset of ℓ distinct integers from $[n]$. What we will need for computing the density matrix and sampling from it, both for the optimal trace distance and robustness approximations, is the following. In each case, the input is the alphabet



(a) (b)

Figure 3: With $d = 200$ and $k = 100$: (a) Example of a positive unit vector $v \in \mathbb{R}^d$ and the corresponding optimal $m \in \mathbb{R}^d$ derived in Lemma 4.7. This m has all three regions described in Eq. (45), but for other choices of v and k, one might see only one or two of these regions. Vertical dotted lines are at $k - r$ and $\ell - 1$, where in this case $r = 39$ and $\ell = 157$. (b) Marginal probability of drawing a k-sparse vector with a nonzero weight on the i^{th} index, as a function of i, with the optimal ensemble of states described in Section 4.3. In this case, the first 60 entries are kept deterministically.

size n, the size of the set ℓ, and the marginal probabilities $q_1, \ldots, q_n$ that each element is included in the set. Normalization requires that $\sum_{j=1}^n q_j = \ell$. Given these inputs, we want a distribution on subsets $S \subset [n]$, with $|S| = \ell$ which can be represented efficiently and satisfies four desired criteria:

1. marginal probability constraints are satisfied, i.e., $q_j = \sum_{S, j \in S} p(S)$ for every $j \in [n]$;
2. two-element marginals $Q_{i,j} := \sum_{S, i, j \in S} p(S)$ can be computed in time $\text{poly}(n)$;
3. the distribution can be sampled from in time $\text{poly}(n)$; and
4. negative correlation (of indicator random variables): $Q_{i,j} \leq q_i q_j$ for all $i, j \in [n]$.

Note that there are $\binom{n}{\ell}$ subsets of size ℓ, so it is not obvious that one can efficiently represent, find and sample from such a distribution. Sampling subsets with constrained marginals is known as *conditional Poisson sampling*. In Section 8, we explain that the *maximum-entropy distribution* is a solution to this problem. This distribution is a distribution of the form

$$p_\mu(S) := \frac{\mathrm{e}^{-\sum_{j \in S} \mu_j}}{Z} \; \forall S \subset [n], |S| = \ell \quad Z_\mu := \sum_{S \subset [n], |S| = \ell} \mathrm{e}^{-\sum_{j \in S} \mu_j} \tag{62}$$

for some $\mu \in \mathbb{R}^n$, where the maximum-entropy distribution is given by $p^\star := p_{\mu^\star}$ with $\mu^\star$ the solution of an optimization problem. Note that these probability distributions are fully specified by μ, so they have an efficient representation. To satisfy the requirements, we need to be able to perform the following algorithmic tasks:

- *Inference:* given q, compute $\mu^\star$.

- *Sampling:* given μ, sample from p_μ.
- *Computing marginals:* given μ, compute Z_μ, q_i and Q_{ij}.

We give efficient algorithms for these tasks in Section 8.

4.3 Computing and sampling the optimal mixed approximation

In this section, we describe the density matrix $\sigma^\star \in \mathcal{I}_k$ which attains the minimum value $T_k(v)$ in the optimization problem in Eq. (31), i.e., the optimal randomized truncation of v. We then show that this density matrix corresponds to an ensemble of k-sparse pure states which can be efficiently sampled, upon obtaining $T_k(v)$ and m from Algorithm 1. The ensemble consists of pure states which are proportional to the vector obtained by keeping the top $k-r-1$ entries of v deterministically, truncating indices greater than $\ell-1$, and drawing $r+1$ nonzero indices at random from the remaining $\ell-k+r$ indices. An example of the marginal inclusion probabilities q_j for $j = k-r, \ldots, \ell-1$ is given in Figure 3b and, as we will see in the proof of Theorem 4.8, it holds that

$$q_j = \frac{v_j}{\theta} - \lambda \quad \text{for every} \quad j \in \{k-r, \ldots, \ell-1\} \tag{63}$$

where $\lambda = T_k(v)$ and $\theta = \theta^v_{r,\ell}(\lambda)$. The algorithm can also straightforwardly be applied to approximate an entangled quantum state by sampling states having Schmidt-rank at most k.

In Theorems 4.8 and 4.9 below, the stated runtimes inherit the runtime of the inference problem for the maximum-entropy sampling. Hence, $\mathrm{poly}(d, \log(1/\eta))$ can be further refined using any of the runtimes quoted in Table 2, e.g., $\widetilde{O}(d^{1.5}k)$ for constant ε using the accelerated projected gradient descent.

Theorem 4.8. *Let $k, d \in \mathbb{Z}_{>0}$ satisfy $k \leq d$, let $\eta \in (0,1)$, and let $v \in \mathbb{C}^d$ be a unit vector. There is an algorithm, requiring at most $\mathrm{poly}(d, \log(1/\eta))$ steps, which computes the entries in the standard basis of a density matrix $\sigma \in \mathcal{I}_k$ such that $T(v,\sigma) \leq T_k(v) + \eta$.*

Proof. We use the shorthand $T_k = T_k(v)$ and assume that v is real, nonnegative, and sorted in nonincreasing order. As explained at the end of Section 2, the relevant phases and ordering can easily be handled upon solving the problem with this simplifying assumption. We make use of the minimax equality

$$T_k = \min_{\sigma \in \mathcal{I}_k} \max_{x \in \mathbb{R}^d : \|x\|_2 = 1} \langle x, v\rangle^2 - \langle x, \sigma x\rangle = \max_{x \in \mathbb{R}^d : \|x\|_2 = 1} \min_{\sigma \in \mathcal{I}_k} \langle x, v\rangle^2 - \langle x, \sigma x\rangle. \tag{64}$$

The first equality is a direct consequence of Lemma 4.2 while the second was established in the proof of Lemma 4.4. Let $(\sigma^\star, m)$ be a saddle point for the this minimax problem. Observe that we must have

$$\max_{\sigma \in \mathcal{I}_k} \langle m, \sigma m\rangle = \langle m, \sigma^\star m\rangle. \tag{65}$$

Otherwise,

$$T_k = \min_{\sigma \in \mathcal{I}_k} \langle m, v\rangle^2 - \langle m, \sigma m\rangle = \langle m, v\rangle^2 - \max_{\sigma \in \mathcal{I}_k} \langle m, \sigma m\rangle < \langle m, v\rangle^2 - \langle m, \sigma^\star m\rangle \tag{66}$$

which contradicts $(\sigma^\star, m)$ being a saddle point. Eq. (65) implies that $\sigma^\star$ is a mixture of pure states attaining the maximum in

$$\max_w \{|\langle m, w\rangle|^2 : \|w\|_2 = 1,\ \|w\|_0 \leq k\}. \tag{67}$$

Additionally, by Lemma 4.7, $m \propto \widetilde{m}$ with

$$\widetilde{m}_i = \begin{cases} \frac{v_i}{1+\lambda}, & i = 1, \dots, k-r-1 \\ \theta, & i = k-r, \dots, \ell-1, \\ \frac{v_i}{\lambda}, & i = \ell, \dots, d \end{cases} \tag{68}$$

for some $r \in \{0, \dots, k-1\}$, $\ell \in \{k+1, \dots, d+1\}$, and $\theta \in \mathbb{R}$ such that $v_{k-r-1}/(1+\lambda) > \theta > v_\ell/\lambda$. Here, $\lambda = T_k$, and all these quantities can be efficiently computed. Eqs. (67) and (68) then imply that each state in the mixture comprising $\sigma^\star$ is of the form

$$w(S) \coloneqq \frac{1}{\|\widetilde{m}\|_{(k)}} \left(\sum_{i=1}^{k-r-1} \widetilde{m}_i e_i + \theta \sum_{j \in S} e_j \right) \tag{69}$$

for some $S \subseteq \{k-r, \dots, \ell-1\}$ satisfying $|S| = r+1$. Denote by Ω the collection of all such S. Then we can write $\sigma^\star = \mathbb{E}_S \, w(S) w(S)^\dagger$ for some set-valued random variable $S \in \Omega$. Next, observe that

$$T_k = \max_{x : \|x\|_2 = 1} x^\dagger \left(vv^\dagger - \sigma^\star \right) x \tag{70}$$

and m is an optimal solution for this problem, which leads to the eigenvector equation $(vv^\dagger - \sigma^\star)m = T_k m$. Consequently,

$$\begin{aligned} T_k \widetilde{m} &= \frac{T_k m}{\langle m, v \rangle} && (71) \\ &= \frac{1}{\langle m, v \rangle} \left[\langle v, m \rangle v - \mathop{\mathbb{E}}_S \langle w(S), m \rangle w(S) \right] && (72) \\ &= v - \mathop{\mathbb{E}}_S \left[\sum_{i=1}^{k-r-1} \widetilde{m}_i e_i + \theta \sum_{j \in S} e_j \right] && (73) \\ &= v - \sum_{i=1}^{k-r-1} \widetilde{m}_i e_i - \theta \sum_{j=k-r}^{\ell-1} q_j e_j && (74) \end{aligned}$$

where $q_j \coloneqq \Pr_S [j \in S]$ is the marginal probability that the index j is included in the random subset S, for each $j \in \{k-r, \dots, \ell-1\}$. Hence, for each $j \in \{k-r, \dots, \ell-1\}$ it holds that $T_k \theta = v_j - \theta q_j$, or equivalently,

$$q_j = \frac{v_j}{\theta} - T_k. \tag{75}$$

One may verify that $\sum_{j=k-r}^{\ell-1} q_j = r+1$ by making use of the fact that $\theta = \theta^v_{r,\ell}(\lambda)$, as in Definition 4.6, which implies these are valid single-element marginal probabilities.

We have successfully derived the fact that $\sigma^\star$ is a mixture of pure states $w(S)$ where S is a random subset of $\{k-r, \dots, \ell-1\}$ of fixed size $r+1$, whose marginal inclusion probabilities satisfy Eq. (75). Since each quantity on the right-hand side of that equation can be efficiently computed (see Algorithm 1) we can also compute these marginal probabilities efficiently.

Next, we claim that sampling S from the max-entropy distribution $p^\star$ subject to these marginal constraints results in an optimal density matrix $\sigma^\star \in \mathcal{I}_k$. This follows from the observation that

for any two distributions $p_1, p_2 : \Omega \to [0,1]$ with the same single-element marginal probabilities, we necessarily have

$$\left(\mathop{\mathbb{E}}_{S_1 \sim p_1} w(S_1) w(S_1)^\top - \mathop{\mathbb{E}}_{S_2 \sim p_2} w(S_2) w(S_2)^\top \right) m = \|m\|_{(k)} \left(\mathop{\mathbb{E}}_{S_1 \sim p_1} w(S_1) - \mathop{\mathbb{E}}_{S_2 \sim p_2} w(S_2) \right) = 0, \quad (76)$$

and this holds in particular for $p_1 = p^\star$ and p_2 any choice of distribution over the subsets $S \in \Omega$ which results in an optimal density matrix.

It remains to argue that we can compute the entries of an approximation σ efficiently, and this is a consequence of the algorithms in Section 8 for conditional Poisson sampling. We saw that taking S to have the max-entropy distribution $p^\star$ subject to the marginal constraints given by the q_j, results in a density matrix $\sigma^\star = \mathbb{E}_{S \sim p^\star} w(S) w(S)^\top$ which is optimal. By Theorem 8.5, since $\ell - k + r \leq d$, using at most $\mathrm{poly}(d, \log(1/\eta))$ operations we can find a set of weights $\mu \in \mathbb{R}^{\ell - k + r}$, specifying a distribution p_μ such that $\mathrm{d_{TV}}(p^\star, p_\mu) \leq \eta$. Setting $\sigma \coloneqq \mathbb{E}_{S \sim p_\mu} w(S) w(S)^\top$, this implies $T(\sigma, \sigma^\star) \leq \eta$ and, by the triangle inequality, $T(v, \sigma) \leq T_k + \eta$. Moreover, setting $y \coloneqq m_{1:k-r-1}$ we have

$$\sigma = \frac{1}{\|m\|_{(k)}^2} \left[y y^\top + \theta \sum_{i=k-r}^{\ell-1} q(\mu)_i (y e_i^\top + e_i y^\top) + \theta^2 \sum_{j,j'=k-r}^{\ell-1} Q(\mu)_{jj'} e_j e_{j'}^\top \right] \quad (77)$$

where $q(\mu)_i \coloneqq \Pr_{S \sim p_\mu}[i \in S]$ and $Q(\mu)_{ij} \coloneqq \Pr_{S \sim p_\mu}[i \in S \wedge j \in S]$ for every $i, j \in \{k-r, \ldots, \ell - 1\}$ are the single- and two-element marginal inclusion probabilities. By Lemma 8.2, each term in the square brackets can be computed efficiently, and the claim follows. □

Theorem 4.9. *Let $\sigma \in \mathcal{I}_k(\mathbb{C}^d)$ be the η-accurate approximation to the optimal density matrix obtained by the algorithm in Theorem 4.8. There is an algorithm running in time $\mathrm{poly}(d, \log(1/\eta))$ which samples a random k-sparse pure state $w \in \mathbb{C}^d$ from a distribution such that $\mathbb{E}\, ww^\dagger = \sigma$. Moreover, subsequent samples can be obtained in $O(d)$ steps.*

Proof. Take the ensemble of pure states to be the random pure state $w(S)$ defined in Eq. (69), where S is a random subset of indices having a distribution which is η-accurate in total variation distance to the max-entropy distribution. As stated in the proof of Theorem 4.8, it takes $\mathrm{poly}(d, \log(1/\eta))$ operations to find the parameters μ corresponding to this distribution. The claim follows from Lemma 8.2, which implies that it takes $O(d)$ steps to sample from that distribution once μ is found. □

4.4 Coherence versus entanglement for the trace distance

In this section we show that for approximating pure states by density matrices with low Schmidt number, the results on the k-sparse approximations carry over directly. Recall that given any state $\rho \in \mathcal{D}(\mathbb{C}^a \otimes \mathbb{C}^b)$, the optimal approximation by a state with Schmidt rank at most k is the following optimization problem:

$$T_k^{(E)}(\rho) := \min_{\sigma \in \mathcal{S}_k} T(\rho, \sigma). \quad (78)$$

We prove the following.

Lemma 4.10. *Let $v \in \mathbb{C}^a \otimes \mathbb{C}^b$ have a Schmidt decomposition*

$$v = \sum_{i=1}^{r} \lambda_i \, x_i \otimes y_i \quad \text{with} \quad \lambda = (\lambda_1, \ldots, \lambda_r)^\top. \quad (79)$$

It holds that $T_k^{(E)}(v) = T_k(\lambda)$. *Moreover, an optimal solution* $\sigma^\star \in \mathcal{S}_k$ *to the minimization in Eq.* (78) *is given by applying the linear map with the action* $e_i \mapsto x_i \otimes y_i$ *for each* $i \in [r]$ *to the optimum in the problem defining* $T_k(\lambda)$.

Proof. Without loss of generality, choose the Schmidt basis to be the standard basis, so

$$v = \sum_{i=1}^{r} \lambda_i \, e_i \otimes e_i, \tag{80}$$

and define the k-sparse states with respect to the basis of states $e_i \otimes e_i$. First of all it is clear that $T_k^{(E)}(v) \leq T_k(v)$ since, with this definition of sparsity, $\mathcal{I}_k \subseteq \mathcal{S}_k$.

In order to conclude that we have equality, we need to show that we can restrict the optimization over $\mathcal{S}_k$ which defines $T_k^{(E)}(v)$ to $\mathcal{I}_k$. To this end, observe that the argument of Lemma 4.4 applies with $\mathcal{I}_k$ replaced by $\mathcal{S}_k$ without alteration, which shows that

$$T_k^{(E)}(v) = \max_{M:\|M\|_\infty \leq 1} \min_{\sigma \in \mathcal{S}_k} \operatorname{tr}\Big(M(vv^\dagger - \sigma)\Big) \tag{81}$$

$$= \max_{m:\|m\|_2 \leq 1} |\langle m, v\rangle|^2 - \|m\|_{(E,k)}^2, \tag{82}$$

where $\|m\|_{(E,k)}^2 \coloneqq \max_{\sigma \in \mathcal{S}_k} \operatorname{tr}\big(mm^\dagger \sigma\big)$ equals the sum of the squares of the k largest Schmidt coefficients of m, by Lemma 2.5. Once we show that the optimal m is of the form $m = \sum_i m_i e_i \otimes e_i$, we can conclude that $T_k^{(E)}(v) = T_k(v)$, since then it is clear we can optimize over $\mathcal{I}_k$ instead of $\mathcal{S}_k$. Consider an arbitrary $m = \sum_{ij} m_{ij} e_i \otimes e_j$. We denote by e_{ij} the matrix $e_i e_j^\top$. Let $P = \sum_i e_{ii} \otimes e_{ii}$ and let $\tilde{m} = \sum_i m_{ii} e_i \otimes e_i = Pm$. Let $\Phi : \mathcal{L}(\mathbb{C}^a \otimes \mathbb{C}^b) \to \mathcal{L}(\mathbb{C}^a \otimes \mathbb{C}^b)$ be the quantum channel with the action $\Phi : X \mapsto \mathbb{E} UXU^\dagger$ where $U \in \mathcal{L}(\mathbb{C}^a \otimes \mathbb{C}^b)$ is a random unitary operator of the form

$$U = \sum_{i,j} z_i z_j^\star e_{ii} \otimes e_{jj} \tag{83}$$

with uniformly random phases $z_i \in \mathbb{C}$, $|z_i| = 1$. This implies that Φ acts on matrix elements through

$$e_{ij} \otimes e_{kl} \mapsto \begin{cases} e_{ij} \otimes e_{kl} & \text{if } i = j \text{ and } k = l \text{ or } i = k \text{ and } j = l \\ 0 & \text{otherwise.} \end{cases} \tag{84}$$

Furthermore, the channel is self-adjoint $\Phi = \Phi^\dagger$ and preserves Schmidt number. Acting on the rank-1 projection operator $mm^\dagger$ gives

$$\Phi(mm^\dagger) = \tilde{m}\tilde{m}^\dagger + \sum_{i \neq j} |m_{ij}|^2 e_{ii} \otimes e_{jj} =: \tilde{m}\tilde{m}^\dagger + B \tag{85}$$

for some operator B which is positive semidefinite. In particular, for any $w \in \mathbb{C}^a \otimes \mathbb{C}^b$ with Schmidt rank at most k, we have

$$|\langle \tilde{m}, w\rangle|^2 = \operatorname{tr}\Big(mm^\dagger \Phi(ww^\dagger)\Big) - w^\dagger B w \leq \|m\|_{(E,k)}^2 \tag{86}$$

where the inequality follows from the fact that $\Phi(ww^\dagger) \in \mathcal{S}_k$ and the second term is at most zero. Maximizing over $w \in \mathcal{S}_k$ we arrive at $\|\tilde{m}\|_{(E,k)} \leq \|m\|_{(E,k)}$. Since $|\langle m, v\rangle|^2 = |\langle \tilde{m}, v\rangle|^2$ we conclude that

$$|\langle m, v\rangle|^2 - \|m\|_{(E,k)}^2 \leq |\langle \tilde{m}, v\rangle|^2 - \|\tilde{m}\|_{(E,k)}^2 \tag{87}$$

and hence we can maximize over m of the form $m = \sum_i m_i e_i \otimes e_i$. □

4.5 Variance estimates for ensembles

A randomized truncation yields a k-sparse state $w(S)$ with some probability. Given an observable M (along with a method to compute its expectation with respect to a target state) one may use the random variable $w(S)^\dagger M w(S)$ as an estimator for the original expectation value of interest, incurring some bias and variance in the process. (For a deterministic truncation, there is no variance.) The trace distance value bounds the bias, while the variance can be related to the quality of an approximation using the individual states $w(S)$ comprising the ensemble. In this section, we record bounds on the variance in such a process, which may inform the feasibility of using these methods in numerical simulations, e.g., using MPS. (See Section 7.) In fact, these bounds hold for generic ensembles of pure states.

Recall the trace distance is given by

$$T(v,\sigma) = \max_{\|M\|_\infty \leq 1} \operatorname{tr}\left(M(vv^\dagger - \sigma)\right) = \max_{\|M\|_\infty \leq 1} \mathbb{E} \operatorname{tr}\left(M(vv^\dagger - w(S)w(S)^\dagger)\right). \tag{88}$$

We have the following.

Lemma 4.11. *Let S be a random variable, let v be an arbitrary pure state, and let $w(S)$ be a random pure state such that $\mathbb{E}\, w(S)w(S)^\dagger = \sigma$.*

1. *The expected value of the trace distance between $w(S)$ and v is bounded by*

$$\mathbb{E}\, T(v, w(S)) \leq \sqrt{T(v,\sigma)}. \tag{89}$$

2. *Let M be any observable with $\|M\|_\infty \leq 1$. It holds that*

$$\operatorname{Var}[w(S)^\dagger M w(S)] \leq T(v,\sigma)\left(1 + \sqrt{T(v,\sigma)}\right)^2. \tag{90}$$

Proof. We can bound the second moment of the random variable $T(v, w(S))$ through

$$\mathbb{E}\, T(v, w(S))^2 = 1 - \mathbb{E}\, |\langle v, w(S)\rangle|^2 = 1 - F(v,\sigma)^2 \leq T(v,\sigma) \tag{91}$$

where the second equality uses $\sigma = \mathbb{E}\, w(S)w(S)^\dagger$ and the inequality follows from Eq. (15). Then the first bound in the lemma is a consequence of the relationship between L_1 and L_2 norms for random variables. For the second bound, note that

$$\left|\operatorname{tr}\left(M(vv^\dagger - w(S)w(S)^\dagger)\right)\right| \leq T(v, w(S)) \tag{92}$$

which implies

$$\mathbb{E} \operatorname{tr}\left(M(vv^\dagger - w(S)w(S)^\dagger)\right)^2 \leq \mathbb{E}\, T(v, w(S))^2 \leq T(v,\sigma). \tag{93}$$

Let $\|X\|_{L_2}$ denote the L_2 norm of a random variable X. By the triangle inequality and Eq. (93),

$$\left\| w(S)^\dagger M w(S) - \operatorname{Tr}(\sigma M)\right\|_{L_2} \leq \left|\operatorname{Tr}\left(M(\sigma - vv^\dagger)\right)\right| + \sqrt{T(v,\sigma)}. \tag{94}$$

The claim follows upon noting that the first term on the right-hand side above is at most $T(v,\sigma)$, and $\mathbb{E}\, w(S)^\dagger M w(S) = \operatorname{Tr}(\sigma M)$. □

These bounds are fairly constraining: suppose that $F_k(v) = \sqrt{1-\varepsilon}$ and we are in a case where randomized truncation is close to the optimal value allowed by the Fuchs-van de Graaf inequalities, i.e., $T(v, \sigma) = O(\varepsilon)$. Then the above argument shows that *on average* the trace distance $T(v, w(S))$ must be $O(\sqrt{\varepsilon})$. But we also know that the optimal pure state approximation has trace distance $\sqrt{\varepsilon}$, so the states $w(S)$ must be close to the optimal deterministic approximations with high probability. Additionally, in this "good" regime the standard deviation of the estimator is at most $O(\sqrt{\varepsilon})$, which is on the same order as the error from the best deterministic truncation even after a single sample. Obtaining multiple samples using the randomized truncation then decreases the variance, from this already-competitive starting point.

5 Approximations in robustness

Recall that a simple expression for the robustness with $\mathcal{I}_k$ of a pure state is derived in [Reg17, JLP+18]. (See Theorem 2.12.) Namely, for any unit vector $v \in \mathbb{C}^d$, we have $R_k(v) = \|v\|^2_{(k,*)} - 1$. By the characterization of the k-support norm from Lemma 2.10, this value is efficiently computable [AFS12], in time $O(d \log d)$. In this section, we show that the optimal density matrix in robustness is also efficiently computable and samplable, via the same max-entropy construction used in Section 4.3.

5.1 Computing and sampling the mixed approximation in robustness

The following result can be shown using a simpler version of the construction leading to Theorem 4.8. As in Section 4.3, below in Theorems 5.1 and 5.2 the runtime $\mathrm{poly}(d, \log(1/\eta))$ can be further refined using any of the runtimes quoted in Table 2.

Theorem 5.1. *Let $k, d \in \mathbb{Z}_{>0}$ satisfy $2 \le k \le d$, let $\eta \in (0,1)$, and let $v \in \mathbb{C}^d$ be a unit vector. There is an algorithm, requiring at most $\mathrm{poly}(d, \log(1/\eta))$ steps, which computes the entries in the standard basis of a density matrix $\tau \in \mathcal{I}_k$ such that $T(\tau, \tau^\star) \le \eta$ where $\tau^\star \in \mathcal{I}_k$ satisfies*

$$\tau^\star = \frac{vv^\dagger + R_k(v)\sigma}{1 + R_k(v)} \quad \text{for some} \quad \sigma \in \mathcal{I}_2. \tag{95}$$

Proof. We use the shorthand $R_k = R_k(v)$ and, as in the proof of Theorem 4.8, assume that v is real, nonnegative, and sorted in nonincreasing order.

Let $\mathcal{S}$ denote the set of all subsets $S \subseteq \{k-r, \dots, d\}$ satisfying $|S| = r+1$. For any $S \in \mathcal{S}$, define the state

$$u(S) \coloneqq \sum_{i=1}^{k-r-1} v_i e_i + \frac{s^v_{k-r}}{r+1} \sum_{j \in S} e_j. \tag{96}$$

These states are not generally unit vectors; in fact, $\langle u(S), u(S)\rangle = 1 + R_k$ for any $S \in \mathcal{S}$. We use them to construct the (not necessarily normalized) mixture

$$\widetilde{\tau} = \mathop{\mathbb{E}}_S u(S) u(S)^\dagger \tag{97}$$

for a yet-to-be-specified distribution over $\mathcal{S}$. Then $\mathrm{tr}(\widetilde{\tau}) = 1 + R_k$ and

$$\widetilde{\tau} = v_{1:k-r-1} v_{1:k-r-1}^\top + \frac{s^v_{k-r}}{r+1} \sum_{i=k-r}^{d} q_i (v_{1:k-r} e_i^\top + e_i v_{1:k-r}^\top) + \left(\frac{s^v_{j-r}}{r+1}\right)^2 \sum_{j,j'=k-r}^{d} Q_{j,j'} e_j e_{j'}^\top \tag{98}$$

where $q_i = \Pr_S[i \in S]$ and $Q_{ij} \coloneqq \Pr_S[i \in S \wedge j \in S]$ for every $i, j \in \{k-r, \ldots, d\}$, are the single- and two-element marginal inclusion probabilities. Now, suppose that the distribution over $S \in \mathcal{S}$ is the max-entropy distribution whose single-element marginals are given by $q_i = v_i(r+1)/s^v_{k-r}$. Evidently, $\sum_{i=k-r}^{d} q_i = r+1$, so this is a valid set of marginal probabilities. Then

$$\Delta \coloneqq \widetilde{\tau} - vv^\dagger = \left(\frac{s^v_{k-r}}{r+1}\right)^2 \sum_{i,j=k-r}^{d} (Q_{ij} - q_i q_j)\, e_i e_j^\top. \tag{99}$$

Since, for every $i = k-r, \ldots, d$ we have $Q_{ii} = q_i$, the matrix Δ has diagonal elements

$$\Delta_{ii} = v_i\left(\frac{s^v_{k-r}}{r+1} - v_i\right), \tag{100}$$

which are positive by the condition in Eq. (23). On the other hand, a basic property of the max-entropy distribution (see Lemma 8.1 below) is that $Q_{ij} < q_i q_j$ for every $i, j \in \{k-r, \ldots, d\}$ with $i \neq j$. Finally, we can compute the i^{th} row sum of Δ as

$$\sum_{j=k-r}^{d} \Delta_{i,j} \propto \sum_{j=k-r}^{d} (Q_{ij} - q_i q_j) = \mathop{\mathbb{E}}_{S} \mathbf{1}_{i \in S}\left(\sum_{j=k-r}^{d} \mathbf{1}_{j \in S}\right) - q_i(r+1) = 0. \tag{101}$$

In summary, Δ has (i) nonnegative diagonals, (ii) nonpositive off diagonals, and (iii) rows summing to 0. Such matrices are known as diagonally dominant and are proportional to an element of $\mathcal{I}_2$, by Lemma 1 in [JLP+18]. Thus, we have established:

$$\widetilde{\tau} = vv^\dagger + R_k \sigma \qquad \text{for some} \qquad \sigma \in \mathcal{I}_2. \tag{102}$$

Setting $\tau^\star = \widetilde{\tau}/(1+R_k)$ gives the desired optimal density matrix. Of course, we cannot compute exactly the one- and two-element marginals corresponding to the max-entropy distribution. But we can do so for an approximation that is η close in total variation distance in time $\mathrm{poly}(d, \log(1/\eta))$. This leads to an approximation τ to $\tau^\star$ which is η-close in trace distance, using Lemmas 8.2 and 8.5, in exactly the same way as in the proof of Theorem 4.8. □

The construction above also tells us how to efficiently sample from the ensemble corresponding to the robustness-optimal density matrix, as in the case for the trace distance.

Theorem 5.2. *Let $\tau \in \mathcal{I}_k(\mathbb{C}^d)$ be the η-accurate approximation to the optimal density matrix obtained by the algorithm in Theorem 5.1. There is an algorithm running in time $\mathrm{poly}(d, \log(1/\eta))$ which samples a random k-sparse pure state $u \in \mathbb{C}^d$ from a distribution such that $\mathbb{E}\, uu^\dagger = \sigma$. Moreover, subsequent samples can be obtained in $O(d)$ steps.*

Proof. The proof is identical to the proof of Theorem 4.9 upon taking the random pure state to be proportional to $u(S)$, defined in Eq. (96), rather than $w(S)$. □

Remark 5.3. Our construction also yields a simple alternate proof of the fact from [Reg17, JLP+18] that $\|v\|^2_{k,*} = 1 + R_k(v) = 1 + R^s_k(v)$ where $R^s_k(v)$ is the smallest $t \geq 0$ for which $\exists \sigma \in \mathcal{I}_k$ such that $\frac{vv^\dagger + t\sigma}{1+t} \in \mathcal{I}_k$. It is straightforward to show that

$$\|v\|^2_{(k,*)} \leq 1 + R_k(v) \leq 1 + R^s_k(v). \tag{103}$$

The second inquality is because R^s_k imposes additional restrictions on σ that do not appear in R_k. For the first inequality, from the definition of $\|\cdot\|_{(k,*)}$ there exists a vector m with $\langle v, m\rangle = \|v\|_{(k,*)}$

and $\|m\|_{(k)} \leq 1$. And from the definition of $R_k(v)$, there exists $\tau \in \mathcal{I}_k$ with $vv^\dagger \leq (1 + R_k(v))\tau$. Since τ is a mixture of k-sparse states, $\langle m, \tau m\rangle \leq 1$. Combining these facts yields the first inequality in Eq. (103).

Finally, Theorem 5.1 constructs $\sigma \in \mathcal{I}_2 \subseteq \mathcal{I}_k$ and $\tau \in \mathcal{I}_k$ such that

$$vv^\dagger + R\sigma = (1 + R)\tau, \tag{104}$$

where $R = R_k(v)$. Examining the proof of Theorem 5.1, we see that in fact it proves Eq. (104) for $R = \|v\|^2_{(k,*)} - 1$. Since Eq. (104) is also a valid decomposition for R^s_k, this proves that $R^s_k(v) \leq \|v\|^2_{(k,*)} - 1$. Thus all the inequalities in Eq. (103) are equalities.

Our proof is short and constructive. However, it does rely on the existence of distributions over subsets of indices with the desired single-element marginals and with negative correlation.

6 Examples

We have derived the *optimal* k-sparse trace distance approximations to pure states. If v is a pure state, and v_k is the optimal pure k-sparse approximation, then we know that $F(v, v_k)^2 + T(v, v_k)^2 = 1$. In other words, if $F_k(v) = F(v, v_k) = \sqrt{1 - \varepsilon}$, we have $T(v, v_k) = \sqrt{\varepsilon}$. On the other hand, when we allow approximation by mixed states, the Fuchs-van de Graaf inequalities as in Eq. (15) only constrain the optimal trace distance as $T_k(v) = T(v, \rho) \geq 1 - F(v, \rho)^2 \geq 1 - F_k(v)^2 = \varepsilon$. In Section 1.1, we already saw an example where indeed $T_k(v) = O(\varepsilon)$. However, there are also cases where $T_k(v) = \Theta(\sqrt{\varepsilon})$.

This raises the question: is such a quadratic improvement in approximation by allowing randomized truncation a very special case, or is it common? We will now give a broad set of examples, which shows that a quadratic improvement happens in many cases. The examples are discussed from the perspective of k-sparsity, but also all have bipartite entanglement equivalents. In these examples, we will denote by d the dimension of the Hilbert space, and we let ε denote truncation error in fidelity, so $F_k(v) = \sqrt{1 - \varepsilon}$. If we fix an allowed error ε, this also implicitly defines a value of k that achieves this error (so we can write $\varepsilon(k)$ or $k(\varepsilon)$).

The easiest way to upper bound $T_k(v)$ is by either giving an explicit approximation $\rho \in \mathcal{I}_k$, or by using the robustness, which is computed by the k-support norm, using Lemma 2.10. Lower bounding $T_k(v)$ can either be done by the Fuchs-van de Graaf inequalities, or by explicitly choosing a vector $m \in \mathbb{C}^d$ which lower bounds the optimization problem in Lemma 4.4.

6.1 Additional simple examples

Uniform state. Suppose $u = d^{-1/2} \sum_{i\in[d]} e_i$ (i.e. in the bipartite case, this corresponds to a maximally entangled state). We have $F_k(u) = \sqrt{k/d}$ so $k = d(1-\varepsilon)$. When computing the k-support norm of u, for the condition in Eq. (23) we have $s_{k-r} = (d - k + r)/\sqrt{d}$, which is never smaller than $(r + 1)u_{k-r-1} = (r + 1)/\sqrt{d}$ for $r < k - 1$ and $k < d$. So in Lemma 2.10, we get $r = k - 1$ and

$$\|u\|^2_{(k,*)} = \frac{s_1^2}{k} = \frac{d}{k} = \frac{1}{1 - \varepsilon}. \tag{105}$$

This yields

$$\begin{aligned} R_k(u) &= \frac{d}{k} - 1 = O(\varepsilon) \\ T_k(u) &\leq \frac{R_k(u)}{1 + R_k(u)} = 1 - \frac{k}{d} = O(\varepsilon). \end{aligned} \tag{106}$$

We actually have $T_k(u) = 1 - \frac{k}{d}$, as can be seen by taking $m = u$ in Lemma 4.4.

The uniform state is the state furthest away from $\mathcal{I}_k$ for any k, when measured in fidelity, robustness or trace distance. To see this, first $T(\rho, \mathcal{I}_k)$ is a convex function of ρ, so is maximized by a ρ pure. Let v be any pure state. We have $\|v\|_{(k)} \geq \sqrt{k/d}$, with equality for the uniform state. This shows the uniform state u is furthest from $\mathcal{I}_k$ in fidelity. Additionally, it implies that the dual norm satisfies $\|v\|_{(k,*)} \leq \sqrt{d/k}$ and hence

$$R_k(v) = \|v\|^2_{(k,*)} - 1 \leq \frac{d}{k} - 1, \tag{107}$$

which further implies that for any state v

$$T_k(v) \leq \frac{R_k(v)}{1 + R_k(v)} \leq 1 - \frac{k}{d}. \tag{108}$$

Both bounds are saturated by the uniform state u.

A state for which there is no advantage. We start with a qubit example, with almost all of the weight on a single basis state. Let $v = (\sqrt{1-\varepsilon}, \sqrt{\varepsilon})^\top$ and take $k = 1$, so $F_1(v) = \sqrt{1-\varepsilon}$. The best 1-sparse approximation in trace distance is

$$\rho = \begin{pmatrix} 1-\varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}. \tag{109}$$

It has $F(\rho, v) = \sqrt{1 - 2\varepsilon + 2\varepsilon^2}$, and gives a trace distance $T_1(v) = \sqrt{\varepsilon(1-\varepsilon)}$, which is only slightly better than the pure state approximation. We can also compute $\|v\|_{(1,*} = s_1 = \sqrt{1-\varepsilon} + \sqrt{\varepsilon}$ and hence

$$R_1(v) = \|v\|^2_{(1,*)} - 1 = 2\sqrt{\varepsilon(1-\varepsilon)}. \tag{110}$$

We can generalize this to states of arbitrary dimension d. Fix $k \leq d$ and ε. Consider v with

$$v_i = \sqrt{\frac{1-\varepsilon}{k}} \text{ for } i = 1, \ldots, k \quad \text{and} \quad v_i = \sqrt{\frac{\varepsilon}{d-k}} \text{ for } i = k+1, \ldots, d. \tag{111}$$

Then $F_k(v) = \sqrt{1-\varepsilon}$. For the trace distance, let

$$m_i = \begin{cases} \frac{1}{\sqrt{2k}} & 1 \leq i \leq k, \\ \frac{1}{\sqrt{2(d-k)}} & k < i \leq d. \end{cases} \tag{112}$$

Then $\|m\|_2 = 1$ and by Lemma 4.4

$$T_k(v) \geq |\langle m, v\rangle|^2 - \|m\|^2_{(k)} = \sqrt{\varepsilon(1-\varepsilon)} \tag{113}$$

so the trace distance does not improve much beyond the pure state approximation. For the robustness, $s_{k-r} = (r+1)\sqrt{\frac{1-\varepsilon}{k}} + \sqrt{(d-k)\varepsilon}$ is large for $d \gg k$, and we get $r = k - 1$ and

$$\|v\|^2_{(k,*)} = \frac{s_1^2}{k} = \frac{1}{k}(\sqrt{(1-\varepsilon)k} + \sqrt{\varepsilon(d-k)})^2 \approx \frac{d}{k}\varepsilon. \tag{114}$$

In this example, for fixed k and ε, the robustness $R_k(v)$ can get arbitrarily larger with increasing d. That is, even if the deterministic truncation error ε is small, $R_k(v)$ can be large.

The case $k = 1$. A different generalization of the previous example is to consider a general $v \in \mathbb{C}^d$ and to take $k = 1$, as studied in [CGJ+16]. If $F_1(v) = \sqrt{1-\varepsilon}$ then $v_1 = \sqrt{1-\varepsilon}$ and so $\|v_{2:d}\|_2 = \sqrt{\varepsilon}$.

There is an explicit but complicated formula for T_1 in [CGJ+16]. But without giving an exact expression, we can see that $T_1 = O(\sqrt{\varepsilon})$ by taking

$$m = \frac{1}{\sqrt{2}} e_1 + \frac{1}{\sqrt{2\varepsilon}} v_{2:k}. \tag{115}$$

From Lemma 4.4, we have

$$T_1(v) \geq \langle m, v\rangle^2 - \|m\|_{(1)}^2 = \left(\sqrt{\frac{1-\varepsilon}{2}} + \sqrt{\frac{\varepsilon}{2}}\right)^2 - \frac{1}{2} = \sqrt{\varepsilon(1-\varepsilon)}. \tag{116}$$

Thus for $k = 1$ we cannot asymptotically improve the trace distance using mixed states. Of course, even in this case, mixed states are the only way to achieve nontrivial robustness, which here equals $\|v\|_1^2 - 1$. And mixed states can still yield a non-asymptotic improvement in trace distance for $k = 1$.

Intuitively the weakness of the $k = 1$ approximation is because σ has no off-diagonal entries, so there is not much space for interesting mixed strategies.

6.2 Power laws

An informative set of examples will be states in which the coefficients of v decay as a power law. Here we will see that the asymptotic improvement in trace distance (or lack thereof) will depend on the exponent of the power law.

For $\gamma > 0$ choose

$$v_i = \frac{i^{-\gamma}}{\sqrt{Z}} \quad \text{and} \quad Z = \sum_{i=1}^{d} i^{-2\gamma} \tag{117}$$

so the coefficients of the state have a power law decay. Does using randomized truncation help? Recall that randomized truncation is closely related to the k-support norm, which interpolates between the ℓ_2 and ℓ_1 norms. There are three relevant regimes. Consider the unnormalized vector w with $w_i = i^{-\gamma}$, so $Z = \|w\|_2$.

(A) For $\gamma < \frac{1}{2}$, both $\|w\|_2$ and $\|w\|_1$ diverge with increasing d.

(B) For $\frac{1}{2} < \gamma < 1$, $\|w\|_2$ converges, but $\|w\|_1$ diverges with increasing d.

(C) For $\gamma > 1$, both $\|w\|_2$ and $\|w\|_1$ converge with increasing d.

We will see that for cases (A) and (C), we get a quadratic advantage, and for (B) we get a smaller improvement.

For the following, we will assume that both d and k are large. We can approximate[5] the normalization Z by using the following asymptotic expansion of generalized harmonic numbers:

$$H_d^{(\alpha)} := \sum_{j=1}^{d} j^{-\alpha} = \zeta(\alpha) + \frac{1}{1-\alpha} d^{1-\alpha} + O(d^{-\alpha}). \tag{118}$$

[5]Note that the intention of this section is to illustrate a qualitative behaviour, rather than prove a rigorous claim. Hence, for brevity, we omit some details pertaining to the accuracy of approximating sums by integrals.

	γ	Z	ε	k	s_{k-r}
(A):	$(0, 1/2)$	$\frac{d^{1-2\gamma}}{1-2\gamma}$	$1-\left(\frac{k}{d}\right)^{1-2\gamma}$	$d(1-\varepsilon)^{\frac{1}{1-2\gamma}}$	$\frac{d^{1-\gamma}-(k-r-1)^{1-\gamma}}{d^{\frac{1}{2}-\gamma}}$
(B):	$(1/2, 1)$	$\zeta(2\gamma)$	$\frac{k^{-(2\gamma-1)}}{\zeta(2\gamma)}$	$\left(\frac{\zeta(2\gamma)}{\varepsilon}\right)^{\frac{1}{2\gamma-1}}$	$\frac{d^{1-\gamma}-(k-r-1)^{1-\gamma}}{\sqrt{\zeta(2\gamma)}(1-\gamma)}$
(C):	$(1, \infty)$	same	same	same	$\frac{(k-r+1)^{1-\gamma}-d^{1-\gamma}}{\sqrt{\zeta(2\gamma)}(\gamma-1)}$

Table 1: Approximations of relevant quantities for target states obeying power laws at different values of the parameter γ.

In particular, for $\gamma < \frac{1}{2}$ the constant factor is small for large d and we can approximate

$$Z \approx \frac{d^{1-2\gamma}}{1-2\gamma} \tag{119}$$

and for $\gamma > \frac{1}{2}$ and large d we can approximate $Z \approx \zeta(2\gamma)$. We can also approximate

$$\varepsilon = \sum_{i=k+1}^{d} |v_i|^2 \approx \frac{1}{Z} \int_k^d x^{-2\gamma} \mathrm{d}x. \tag{120}$$

Finally, we can estimate

$$s_{k-r} = \frac{1}{\sqrt{Z}} \sum_{i=k-r}^{d} i^{-\gamma} \approx \frac{1}{\sqrt{Z}} \int_{k-r-1}^{d} x^{-\gamma} \mathrm{d}x. \tag{121}$$

Ignoring some terms which are small as d grows, this gives the approximations shown in Table 1.

We will now determine for these three cases how $T_k(v)$ scales with ε, where $\sqrt{\varepsilon}$ is the deterministic truncation error. We take γ to be a fixed constant value (which we absorb into the constants in the big-O notation). We will see that we get a quadratic advantage for small ε for cases (A) and (C), and a smaller advantage in between for (B). This is illustrated in Figure 4.

Case (A): For this case, where $0 < \gamma < \frac{1}{2}$ we take k as in the table, and use the general bound in Eq. (108) to get

$$T_k(v) \leq 1 - \frac{k}{d} = 1 - (1-\varepsilon)^{\frac{1}{1-2\gamma}} = O(\varepsilon), \tag{122}$$

so for large k (and small ε) randomized truncation gives a quadratic improvement in trace distance.

Case (B): Here $\frac{1}{2} < \gamma < 1$. Let $\beta = 1/(3-2\gamma)$, let $\ell = k^{2\beta}$, and define w to be the deterministic truncation of v to ℓ terms. The error in trace distance can be bounded as

$$T(v, w) = \sqrt{1 - F_\ell(v)^2} = \sum_{i=\ell+1}^{d} v_i^2 = O\left(\ell^{\frac{1}{2\gamma-1}}\right) = O(\varepsilon^\beta). \tag{123}$$

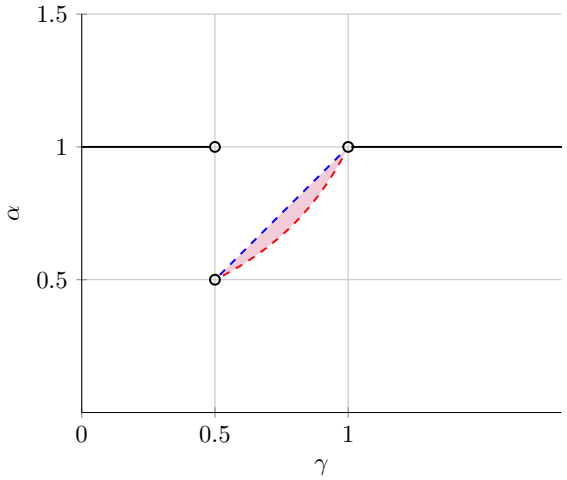


Figure 4: Leading order exponent α of the optimal randomized trace distance as a function of $\varepsilon \equiv 1 - F_k(v)^2$ for a d-dimensional power law state v, as described in the main text. In particular, the state v is defined as in Eq. (117) with parameter γ, and the value of k is chosen to keep ε fixed as $d \to \infty$. Then the resulting trace distance is $c\varepsilon^\alpha$ to leading order, for some constant c. For $\gamma \notin (1/2, 1)$ the exponent is equal to 1, corresponding to the maximum possible advantage using randomized truncation. For $\gamma \in (1/2, 1)$ the value of the exponent lies within the shaded region.

We now bound $R_k(w)$. Using the quantities listed in Table 1 for this case, we see

$$\begin{aligned} \|w\|_{(k,*)}^2 &= \sum_{i=1}^{k-r-1} |w_i|^2 + \frac{s_{k-r}^2}{r+1} \\ &\leq 1 + \frac{s_1^2}{k} \\ &= 1 + O\left(\frac{\ell^{2(1-\gamma)}}{k}\right) \\ &= 1 + O\left(k^{4\beta(1-\gamma)-1}\right) \\ &= 1 + O\left(k^{-\frac{2\gamma-1}{3-2\gamma}}\right) \\ &= 1 + O(\varepsilon^\beta), \end{aligned} \tag{124}$$

which gives

$$R_k(w) = O(\varepsilon^\beta) \quad \text{and} \quad T_k(w) = O(\varepsilon^\beta). \tag{125}$$

Therefore,

$$T_k(v) \leq T(v, w) + T_k(w) = O(\varepsilon^\beta). \tag{126}$$

Since $\beta > \frac{1}{2}$ this is an improvement over the pure state approximation, which achieves trace distance $\sqrt{\varepsilon}$.

We now show that one can not do much better in this case, and that $T_k(v) = \Omega(\varepsilon^\gamma)$, which is worse than the optimal scaling with ε since $\gamma < 1$. We give a lower bound by choosing a unit vector m and bounding

$$T_k(v) \geq |\langle m, v\rangle|^2 - \|m\|_{(k)}^2 \tag{127}$$

by Lemma 4.4. We let $\ell > k$ and $0 < a < 1$, which we choose later, and let

$$m_i = \begin{cases} a v_i & \text{for } i = 1, \dots, k \\ \sqrt{\frac{1-a^2(1-\varepsilon)}{\ell}} & \text{for } i = k+1, \dots, k+\ell \end{cases} \tag{128}$$

which is chosen such that

$$\|m\|_2^2 = a^2 \|v\|_{(k)}^2 + 1 - a^2(1-\varepsilon) = 1. \tag{129}$$

We now have

$$\begin{aligned} |\langle m, v \rangle|^2 &\geq a^2 \|v\|_{(k)}^4 + 2a \sqrt{\frac{1-a^2(1-\varepsilon)}{\ell}} \sum_{i=k+1}^{k+\ell} v_i \\ &= a^2(1-\varepsilon)^2 + 2a \sqrt{\frac{1-a^2(1-\varepsilon)}{\ell}} \sum_{i=k+1}^{k+\ell} v_i \,. \end{aligned} \tag{130}$$

Also, as long as $m_{k+1} \leq m_k$,

$$\|m\|_{(k)}^2 = a^2 \|v\|_{(k)}^2 = a^2(1-\varepsilon). \tag{131}$$

This gives

$$T_k(v) \geq |\langle m, v \rangle|^2 - \|m\|_{(k)}^2 = 2a \sqrt{\frac{1-a^2(1-\varepsilon)}{\ell}} \sum_{i=k+1}^{k+\ell} v_i - O(\varepsilon). \tag{132}$$

We choose $a = \frac{1}{2}(1-\sqrt{\varepsilon})^{-\frac{1}{2}}$, which gives $1 - a^2(1-\varepsilon) = \frac{1}{4}(3-\varepsilon)$, so and $ab = \Theta(1)$. We choose $\ell = \Theta(k^{2\gamma})$, which is good enough to make sure $\Omega(k^{-\gamma}) = m_k \geq m_{k+1} = \Theta(1/\sqrt{\ell})$, and for which $v_{k+1} + \cdots + v_{k+\ell} = \Omega(\ell^{1-\gamma})$. We conclude that

$$T_k(v) = \Omega(\ell^{\frac{1}{2}-\gamma}) = \Omega(k^{-\gamma(2\gamma-1)}) = \Omega(\varepsilon^{\gamma}). \tag{133}$$

Case (C): For large d, we have

$$\varepsilon = \Theta(k^{1-2\gamma}). \tag{134}$$

When computing the k-support norm, we have $s_{k-r} = \Theta((k-r)^{1-\gamma})$, since for $\gamma > 1$, the $d^{1-\gamma}$ term is subleading. The condition that $(r+1)v_{k-r} = \Theta(r(k-r)^{-\gamma})$ should be approximately equal to s_{k-r} in Eq. (23) then gives $r = \Theta(k)$ and as before we can bound

$$\|v\|_{(k,*)}^2 = \sum_{i=1}^{k-r-1} |v_i|^2 + \frac{s_{k-r}^2}{r+1} \leq 1 + O(k^{-1}(k^{1-\gamma})^2) = 1 + O(k^{1-2\gamma}) = 1 + O(\varepsilon). \tag{135}$$

Hence, $R_k(v) = O(\varepsilon)$, and also $T_k(v) = O(\varepsilon)$.

Summary. We can summarize the state of affairs for power law states with a diagram, given in Figure 4. For v a fixed d-dimensional power law state, as in Eq. (117), consider the trace distance $T_d(\varepsilon)$ defined as $T_k(v)$ for k such that $\varepsilon = 1 - F_k(v)^2$. In other words, this is the optimal randomized trace distance as a function of ε, where $\sqrt{\varepsilon}$ is the optimal deterministic trace distance. Then the arguments above show that $\lim_{d\to\infty} T_d(\varepsilon) = \Theta(\varepsilon^{\alpha})$ for some $\alpha \in [1/2, 1]$, the exponent of the leading-order contribution to the trace distance. The case where $\alpha = 1$ represents the maximal possible advantage allowed by the bound in Eq. (15), while $\alpha = 1/2$ corresponds to no advantage over deterministic truncation.

7 Application to matrix product states

In this section, we discuss the application of randomized truncation to tensor network algorithms. We focus on matrix product states (MPS), but the principle applies to other tensor networks as well.

Definition 7.1. Let $\mathcal{H} = (\mathbb{C}^d)^{\otimes n}$, let $\{e_{i_1 \dots i_n} = e_{i_1} \otimes \cdots \otimes e_{i_n} : i_1, \dots, i_n \in [d]\}$ be the standard basis vectors, and let $r_0, r_1, \dots, r_n \in \mathbb{Z}_{>0}$ with $r_0 = r_n$. A state $v \in \mathcal{H}$ is a *matrix product state* with bond dimensions $(r_0, \dots, r_n)$ if it can be written as

$$v = \sum_{i_1, \dots, i_n \in [d]} \operatorname{tr}\left(M^{(1,i_1)} M^{(2,i_2)} \cdots M^{(n,i_n)}\right) e_{i_1 \cdots i_n}, \tag{136}$$

where for each $\alpha \in [n]$ and $i \in [d]$ we have $M^{(\alpha,i)} \in \mathbb{C}^{r_{\alpha-1} \times r_\alpha}$ are some matrices.

We restrict our attention to MPS with open boundary conditions, where $r_0 = r_n = 1$. In many applications, the goal is to simulate a quantum many-body system (e.g., spin degrees of freedom on a lattice, or shallow quantum circuits) by representing its state as an MPS. In order to limit the computational and memory resources used in these simulations, a common subroutine is to truncate a bond, which usually refers to keeping the largest Schmidt coefficients with respect to the bipartition corresponding to that bond.

We propose using an optimal randomized truncation of the bonds instead. Although the results from Sections 4 and 5 give efficient methods for this procedure, there are at least two unresolved issues in this setting. First, the trace distance is a worst-case error metric and therefore does not provide a guaranteed improvement over determinstic truncation for a specific observable of interest. Nevertheless, it is reasonable to expect quantum states which are closer in trace distance (or robustness) to lead to better estimates. The second issue is that performing multiple randomized truncations in succession may not yield a globally optimal approximation, i.e., an approximation of the target state as a convex combination of MPS with bounded bond dimensions. However, this observation applies equally well in the case where one uses deterministic truncations. In fact, finding the best product state approximation (corresponding to $\chi = 1$) to a multipartite state is NP-hard to compute, even for $n = 3$ parties [HL13], so one should not expect efficient algorithms for finding globally optimal MPS approximations.

Details of numerics. In light of the above discussion, we perform numerics to test the effectiveness of using randomized truncation for multipartite states. We begin by generating partially random 10-qubit states as follows. First, we sample a 10-qubit state from the uniform distribution. Next, we compute a Schmidt decomposition of the state with respect to the first bond and replace the resulting Schmidt coefficients with coefficients obeying a (normalized) power law as in Eq. (117), for fixed exponent γ. We then carry out an identical procedure on the remaining bonds in sequence from left to right. The result is a multipartite state whose entanglement spectrum with respect to any given bipartition is constrained depending on the value of γ.

To compare the deterministic and randomized approaches, we truncate the bonds of the resulting state (viewed as an MPS) and compute the expected value of the observable $Z_5 Z_6$, the Pauli-Z operators on the 5th and 6th qubits. Specifically, we once again perform Schmidt decompositions from left to right in sequence, truncating and normalizing each time. For the deterministic truncation, at each bond we keep the top χ Schmidt coefficients, where χ is the bond dimension cutoff. For the randomized method, at each bond we randomly sample the coefficients according to the optimal randomized truncation in trace distance with sparsity χ, as in Theorem 4.9. The result is a randomly

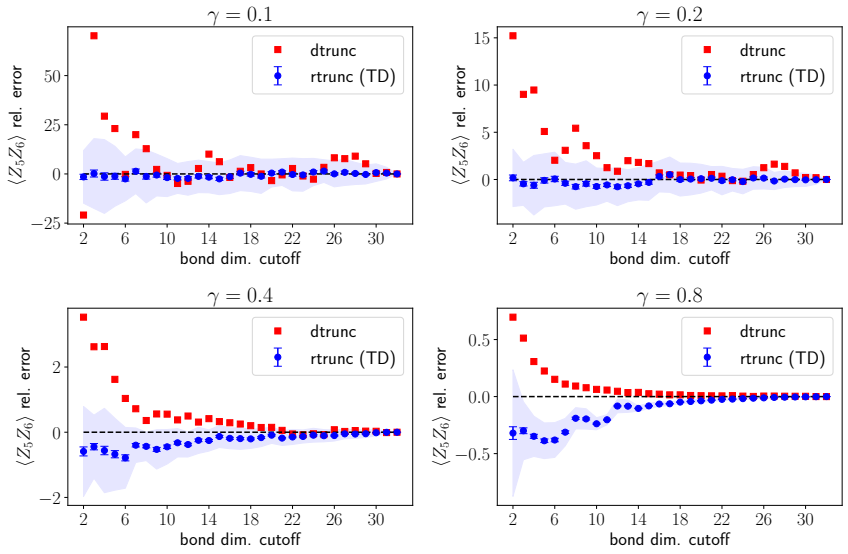


Figure 5: Estimating the observable $\langle Z_5 Z_6 \rangle$ with respect to 4 random states on 10 qubits, which are generated using the method described in the main text. The parameter γ determines the strength of the power law decay in the Schmidt coefficients across each bond. The dtrunc estimate is obtained from the standard, deterministic truncation of the bonds, while the rtrunc (TD) estimate is the sample mean of 100 samples obtained by applying the optimal (bipartite) randomized truncation in trace distance. The light blue region shows an estimate of the standard deviation σ, and error bars of length $\sigma/10$, corresponding to uncertainties in the empirical mean, are also displayed.

sampled MPS with bond dimensions at most χ, from which we can compute the expected value of $Z_5 Z_6$. To compute the final estimate, we take the sample mean of 100 samples. We plot the results for different values of γ in Figure 5.

Discussion. Figure 5 suggests that randomized truncation offers a significant improvement for heavier tails and at small bond dimension cutoffs, by an order of magnitude in the relative error in some cases. However, when the exponent for the power law $\gamma = 0.8$, there is little improvement over the deterministic approach, and both methods result in fairly accurate estimates. This is consistent with the observation that for, say, the middle bond, more than 70% of the entanglement spectrum is concentrated on just the top two squared Schmidt coefficients, so one should not expect more than a modest improvement by sampling smaller Schmidt coefficients. Randomized truncation can in principle be used as a subroutine in any MPS algorithm that requires bond dimension truncation, including TEBD and DMRG (see, e.g., [Sch11]). Whether it is beneficial in practice will depend on the application at hand and how well trace distance or robustness correlate with the relevant metrics for success. Crucially, the computational cost of the randomized truncation may be comparable to the standard method. Note that in either case we must perform Schmidt decompositions at each bond. However, for randomized truncation we must in addition apply the algorithms from Section 4.3 or Section 5, which require computing the weights in a max-entropy distribution. As shown in Section 8, this can be done efficiently and in practice has cost $O(\chi d)$ per iteration when using the Newton method with the diagonal approximation outlined in Section 8.3. Upon solving for the weights, obtaining a single sample of the Schmidt coefficients at a given bond takes $O(\chi)$ time. Each of these times is subleading compared to the amount of time to perform the SVDs in the

deterministic truncation. To fully evaluate the utility of the method, it remains to determine the number of iterations and samples needed, and then compare this with the cost of reaching the same precision with deterministic truncation and a larger bond dimension. We leave a detailed analysis of this tradeoff to future work.

Prior work on tensor network Monte Carlo. The use of a randomized tensor network truncation algorithm has previously been proposed in Ref. [Fer15]. There are two key differences between that work and the approach we have discussed above. First, we use explicitly normalized states for the randomized truncation, whereas the method from Ref. [Fer15] uses an unbiased estimate of the state vectors, and so each state is only normalized in expectation. Iterating the latter procedure enough times will lead to an exponentially growing variance [Tod24], while iterating approximations using normalized states will instead incur growing bias.

A more substantial difference is that our method may drop or keep a subset of the Schmidt coefficients deterministically, and chooses marginal inclusion probabilities for the remaining indices, which we have denoted by q thoughout this work, in a way that minimizes trace distance or robustness. In Section 8 we describe how to compute from q the max-entropy weights μ which lead to the optimal ensembles of states. In contrast, Ref. [Fer15] simply sets μ_i such that $v_i^2 = e^{-\mu_i}$ for all the indices i. This is a reasonable heuristic, since larger values of v_i then yield larger values of q_i, though it does not enjoy the theoretical guarantees established in the present work. How the method of [Fer15] compares to our randomized approach is an interesting question which could be explored numerically in future work.

8 Improved algorithms for conditional Poisson sampling

We make use of efficient methods to compute and sample from distributions over ℓ distinct integers drawn from the set $[n] = \{1, 2, \ldots, n\}$, with $n \geq \ell$. Equivalently, we are interested in random n-bit strings of Hamming weight ℓ, representing inclusion in the draw. We refer to this set as $H_\ell := \{x \in \{0,1\}^n : |x| = \ell\}$. The precise problem we consider is as follows. The inputs are the alphabet size n, the size of the set ℓ, and the marginal probabilities $q_1, \ldots, q_n \in (0, 1)$ that each element is included in the set[6]. Normalization requires that $\sum_{j=1}^n q_j = \ell$. Given these inputs our goal is to efficiently compute a description of a distribution $p : H_\ell \to [0, 1]$ that satisfies four desired criteria:

1. single-element marginal (inclusion) probability constraints are satisfied, i.e., $q_j = \sum_{x \in H_\ell} p(x) x_j$ for every $j \in [n]$;
2. two-element marginals $Q_{i,j} := \sum_{x \in H_\ell} p(x) x_i x_j$ can be computed in time poly(n);
3. the distribution can be sampled from in time poly(n); and
4. negative correlation (of indicator random variables): $Q_{i,j} \leq q_i q_j$ for all $i, j \in [n]$.

As we will see in Section 8.1 below, this problem reduces to solving for the maximum-entropy distribution subject to Item 1, the single-element marginal probability constraints. Such distributions over H_ℓ are a special case of Fisher's multivariable noncentral hypergeometric distribution and are ubiquitous throughout physics and mathematics; for example, when studying systems of ℓ fermions restricted to occupy n distinct modes, or in survey sampling, where the corresponding sampling task

[6]We exclude the possibility of $q_i \in \{0, 1\}$ without loss of generality, since in those cases we can deterministically leave out or include that element and do not need to involve it in the sampling.

is known as conditional Poisson sampling [Til06]. The latter terminology, which we adopt in this work, is due to the observation that (i) sampling *independent* but non-identical Bernoulli random variables $z_1, \ldots, z_n$ is called Poisson sampling, and (ii) if the resulting conditional distribution, conditioned on $\sum_i z_i = \ell$, satisfies the marginal constraints it is in fact the maximum-entropy distribution over H_ℓ. (See, e.g., [CDL94].) For randomized truncation the problem is relevant because our methods, described in Sections 4 and 5, involve sampling random fixed-size subsets of indices from the vector to be truncated.

8.1 Maximum-entropy distributions

We consider the following optimization problem:

$$\begin{aligned} \max_{p \in \mathbb{R}_{\geq 0}^{H_\ell}} \quad & \mathrm{H}(p) \coloneqq \sum_{x \in H_\ell} p(x) \log(1/p(x)) \\ \text{s.t.} \quad & \sum_{x \in H_\ell} p(x) x_j = q_j \qquad \forall j \in [n], \\ & \sum_{x \in H_\ell} p(x) = 1. \end{aligned} \tag{137}$$

Since there are exponentially many variables and only $n+1$ constraints, most work on the subject focuses on the dual optimization problem. First we observe that the optimal solution to Eq. (137) takes the form $p^\star(x) \equiv p_{\mu^\star}(x)$ for some $\mu^\star \in \mathbb{R}^n$, where

$$p_\mu(x) \coloneqq \frac{\mathrm{e}^{-\sum_{j=1}^n x_j \mu_j}}{Z} \ \forall x \in H_\ell, \quad Z_\mu \coloneqq \sum_{x \in H_\ell} \mathrm{e}^{-\sum_{j=1}^n x_j \mu_j} \tag{138}$$

for any set of parameters $(\mu_j \in \mathbb{R} \colon j \in [n])$. This is a standard fact from statistical physics, but for a self-contained proof of a slightly more general fact, see [SV14, Lemma A.1]. Due to Eq. (138), the optimal distribution has a succinct description in terms of just n parameters $\mu_1^\star, \ldots, \mu_n^\star$, which correspond to the optimal dual variables for the dual problem:

$$\begin{aligned} \text{min.} \quad & g_q(\mu) \coloneqq \sum_{i=1}^n \mu_i q_i + \log\left(\sum_{x \in H_\ell} \mathrm{e}^{-\sum_{j=1}^n x_j \mu_j} \right) \\ \text{s.t .} \quad & \mu_j \in \mathbb{R} \quad \forall j \in [n] \end{aligned} \tag{139}$$

where $q \coloneqq (q_1, q_2, \ldots, q_n)^\top$.

We introduce some additional notation before proceeding further. In the above, we considered q to be an input to the optimization problems Eqs. (137) and (139). However, we can also consider $q(\mu)$, a function of μ, where we first determine a distribution p_μ over H_ℓ which has the form given in Eq. (138), and then we use this to determine the single-element marginals $q(\mu)_i = \sum_x p_\mu(x) x_i$ and two-element marginals $Q(\mu)_{ij} = \sum_x p_\mu(x) x_i x_j$. In particular, $q(\mu^\star) = q$ for the dual optimal solution $\mu^\star$. The following is a useful fact about these marginals, recorded as Property 2 in [CDL94].

Lemma 8.1. *For any $\mu \in \mathbb{R}^n$ and $i, j \in [n]$ with $i \neq j$ we have $0 < Q(\mu)_{ij} < q(\mu)_i q(\mu)_j$.*

In other words, the x_i, x_j are negatively correlated for any $i \neq j$.

8.2 Sampling from the maximum-entropy distribution

Given the single-element marginal probabilities $q \in \mathbb{R}^n$ as input, the algorithmic tasks involved in a maximum-entropy approach to the conditional Poisson sampling problem are:

- Compute the dual optimal solution $\mu^\star$ from q.
- Given $\mu \in \mathbb{R}^n$, compute Z_μ, $q(\mu)$ and $Q(\mu)$.
- Given $\mu \in \mathbb{R}^n$, sample from p_μ.

The first problem is referred to as *inference* since we need to infer the parameters $\mu^\star$ from q, and we give algorithms for it in Section 8.3 below. In the remainder of this section we focus on the arithmetic complexity of the latter two problems, ignoring bit complexity. However, Theorem 8.5 gives us a polynomial bound on the largest value of $\|\mu\|_\infty$ which need be considered; hence taking bit complexity into account only adds a polylogarithmic overhead. We give algorithms that improve on those in [CDL94, Til06] for large enough ℓ, which computed Z_μ in time $O(n\ell)$, $q(\mu)$ in time $O(n\ell^2)$ and $Q(\mu)$ in time $O(n^2\ell)$, and sampled from p_μ in time $O(n\ell)$.

Lemma 8.2. *Given $\mu \in \mathbb{R}^n$ one can compute Z_μ in time $O(n \log^2 \ell \log\log \ell)$, $q(\mu)$ in time $O(n\ell)$, and $Q(\mu)$ in time $O(n^2)$.*

Proof. The partition function Z_μ is equal to the coefficient of t^ℓ in the polynomial

$$f(t) = \prod_{i=1}^{n} \left(1 + te^{-\mu_i}\right) \tag{140}$$

We can evaluate $f(t)$ with a multiplication tree of depth $\log n$, while keeping track only of intermediate polynomials modulo $t^{\ell+1}$. This involves $n/2^j$ degree-2^{j-1} multiplies for $j = 1, \ldots, \log \ell$ and then $O(n/\ell)$ degree-ℓ multiplies. Since degree-d polynomial multiplication can be done in time $O(d \log(d) \log\log(d))$ using Ref. [CK87], the final runtime is $O(n \log^2(\ell) \log\log(\ell))$.

The previous computation of Z_μ in fact yielded $Z_\mu^{(m)}$ for $m = 1, \ldots, \ell$ where we define

$$p_\mu^{(m)}(x) \coloneqq \frac{\mathrm{e}^{-\sum_{j\in[n]} x_j \mu_j}}{Z_\mu^{(m)}} \; \forall x \in H_m, \quad Z_\mu^{(m)} \coloneqq \sum_{x \in H_m} \mathrm{e}^{-\sum_{j\in[n]} x_j \mu_j} \tag{141}$$

Similarly, set $q_j^{(m)}(\mu) = \sum_{i\in[n]} p_\mu^{(m)}(x) x_i$ to make the m-dependence of the marginal probabilities explicit. Then Result 22(i) from [Til06] states that

$$q_j^{(m)}(\mu) = e^{-\mu_j}(1 - q_j^{(m-1)}(\mu)) \frac{Z_\mu^{(m-1)}}{Z_\mu^{(m)}}. \tag{142}$$

Note that $q_j^{(0)} = 0$. Then given knowledge of $Z_\mu^{(1)}, \ldots, Z_\mu^{(\ell)}$ from the previous step we can compute each $q_j^{(\ell)}(\mu) = q_j(\mu)$ in time $O(\ell)$. Computing all of q takes time $O(n\ell)$, which dominates the runtime compared to computing the required partition functions. (Result 22(iii) from [Til06] is another option, yielding the same runtime.)

For $Q_{ij}(\mu)$, when $\mu_i \neq \mu_j$ there is a formula that can be evaluated in constant time. In this case, Result 23(ii) from [Til06] states

$$Q_{ij}(\mu) = \frac{q_i(\mu)e^{-\mu_j} - q_j(\mu)e^{-\mu_i}}{e^{-\mu_j} - e^{-\mu_i}}. \tag{143}$$

The remaining entries of $Q(\mu)$ can be computed using the fact that $Q_{ii}(\mu) = q_i(\mu)$ and $\sum_j Q_{ij}(\mu) = q_i(\mu)n$. Concretely, when $\mu_i = \mu_j$, define $C := |\{j : j \neq i, \mu_j = \mu_i\}|$, so that

$$Q_{ij}(\mu) = \frac{1}{C} q_i(\mu)(n-1) - \frac{1}{C} \sum_{k:\mu_k \neq \mu_i} Q_{ik}(\mu) \tag{144}$$

when $i \neq j$ and $\mu_i = \mu_j$. This also costs $O(1)$ time per entry, if we assume that the μ vector is sorted. Thus $Q(\mu)$ can be computed in time $O(n^2)$. □

We now give an improved algorithm to sample from the distribution p_μ. A different algorithm with the same runtime is given in [HJJ20], but it only yields approximate samples and the runtime bound is not fully proven.

Lemma 8.3. *Given $\mu \in \mathbb{R}^n$, a sample from p_μ can be obtained in expected time $O(n \log n)$.*

The first step of our algorithm involves a rescaling of the weights which we separate into its own lemma.

Lemma 8.4. *Given $\mu \in \mathbb{R}^n$ and $\ell \leq n-1$, in time $O(n \log(n(1 + \|\mu\|_\infty)/\varepsilon))$ we can find $\Delta \in \mathbb{R}$ such that*

$$\sum_{i=1}^{n} (1 + e^{\mu_i + \Delta})^{-1} = \ell, \tag{145}$$

up to error ε. This implies that the product distribution $P_{\mu+\Delta} = \bigotimes_i Ber\big(1 + e^{\mu_i+\Delta}\big)^{-1}$ has expected Hamming weight ℓ (up to the approximation error). Additionally $P_{\mu+\Delta}|_{H_\ell} = p_\mu$.

We say that the $\mu + \Delta$ satisfying Eq. (145) are "centered."

Proof of Lemma 8.4. Let $g : \mathbb{R} \to [0, n]$ be defined through $g(\Delta) = \sum_{i=1}^n \frac{1}{1+e^{\mu_i+\Delta}}$. Our goal is to find Δ such that $g(\Delta) = \ell$. Since g is monotonically decreasing, binary search will work here. Write $R = \|\mu\|_\infty$. Then

$$\frac{n}{1 + e^{R+\Delta}} \leq g(\Delta) \leq \frac{n}{1 + e^{-R+\Delta}} \quad \text{and} \quad -n \leq g'(\Delta) \leq 0. \tag{146}$$

We can then set

$$a := -R + \log\left(\frac{n-\ell}{\ell}\right) \quad \text{and} \quad b := R + \log\left(\frac{n-\ell}{\ell}\right) \tag{147}$$

to ensure that $g(a) \geq \ell$ while $g(b) \leq \ell$. Performing a binary search over the interval $[a, b]$ yields Δ in time $O(n \log(R/\varepsilon))$, to accuracy ε. which will suffice for the rest of the algorithm.

Finally, the last claim of the lemma is that p_μ is not affected by the choice of Δ. This is because each $x \in H_\ell$ has the same $e^{\ell\Delta}$ dependence on Δ, so once we condition on $x \in H_\ell$, the effects of Δ go away. □

By Lemma 8.8, if $q_j \in [\eta, 1-\eta]$, we can easily shift μ so that $\|\mu\|_\infty \leq O(\log(1/\eta))$, and inverse-polynomial precision in ε will suffice, so this step can be done efficiently.

Proof of Lemma 8.3. Begin by using the algorithm Lemma 8.4 so that we can use a vector $\mu \in \mathbb{R}^n$ such that $\sum_{i=1}^n (1 + e^{\mu_i})^{-1} = \ell$. Define the product distribution P_μ where x_i is distributed according to $\mathrm{Ber}(\frac{1}{1+e^{\mu_i}})$, i.e., the Bernoulli random variable with $\Pr[x_i = 1] = \frac{1}{1+e^{\mu_i}}$ and $\Pr[x_i = 0] = \frac{e^{\mu_i}}{1+e^{\mu_i}}$. Our choice of Δ ensures that $\mathbb{E}_{P_\mu}[|x|] = \ell$. Furthermore, it is straightforward to check that p_μ is

equivalent to conditioning P_μ on the event $H_\ell := \{x : |x| = \ell\}$. Rather than naively sampling until the desired event occurs[7], out algorithm proceeds as follows:

1. Sample x from P_μ.
2. Repeat the following n times.
   - If $x \in H_\ell$ then output x and stop.
   - Choose $i \in [n]$ uniformly randomly and replace x_i with $\mathrm{Ber}(\frac{1}{1+e^{\mu_i}})$.
3. Repeat Step 1.

We claim that this procedure halts in Step 2 with probability $\Omega(1)$. This means the expected runtime is $O(n)$, and the runtime is at most $O(n \log(1/\varepsilon))$ with probability at least $1-\varepsilon$. Furthermore, conditioned on halting, the algorithm outputs a sample from p_μ. To prove this, let $x(0), x(1), x(2), \ldots$ denote the values of x produced by the algorithm. By construction, $x(0)$ is distributed according to P_μ and, since the update procedure is Glauber dynamics, the subsequent $x(t)$ are as well. Finally, the algorithm halts when $x(t) \in H_\ell$, thus producing a sample from P_μ conditioned on H_ℓ. This yields a sample from p_μ.

It remains only to analyze the success probability in Step 2. We can get intuition for this from the case when each q_i is equal to $1/2$, $\ell = n/2$ and n is large. Let $z = |x| - n/2$. In that case z starts typically around $\pm O(\sqrt{n})$ and then evolves according to n steps of a simple random walk. For concreteness, suppose z starts at $c\sqrt{n}$ for $c > 0$. The probability that a random walk will hit $-c\sqrt{n}$ after n steps is known to be $\approx \exp\bigl(-c^2/2\bigr)$, which is $\Omega(1)$ if c is $O(1)$. Extending this argument to nonequal q_i is more delicate but follows a similar intuition. We next analyze this part of the algorithm

Define $s(x)$ to be $+1$ if $|x| > \ell$, 0 if $|x| = \ell$, and -1 if $|x| < \ell$. A sufficient, but not necessary, condition for the algorithm to succeed is for $s(x(0))s(x(n))$ to equal 0 or -1. If the product is 0, then either $x(0)$ or $x(n)$ is in H_ℓ. If the product is -1 then the Hamming weight went from $> \ell$ to $< \ell$ (or vice versa) and so at some intermediate point we had $x(t) \in H_\ell$. It is possible for the algorithm to succeed even if $s(x(0))s(x(n)) = 1$, but we will not analyze this case. We now use the mixing properties of the walk to show that the signs of $s(x(0))$ and $s(x(n))$ have bounded correlations, and thus have a decent chance of either having opposite signs or of being zero.

Let $\pi \coloneqq P_\mu(H_\ell)$. For later use, we record that

$$\pi = \Omega(\ell^{-1/2}). \tag{148}$$

Indeed, this follows from Chebyshev's inequality together with

$$\mathbb{E}\,|x| = \sum_i q_i = l, \qquad \mathrm{Var}[|x|] = \sum_i q_i(1-q_i) \leq \sum_i q_i = l. \tag{149}$$

Theorem 3.2 in Ref. [JS68] states that the median for the Hamming weight of x is equal to ℓ, given our assumption that the mean is equal to ℓ. This implies the inequality:

$$\left| \mathop{\mathbb{E}}_{x\sim P_\mu}[s] \right| = \left| \Pr_{x\sim P_\mu}[|x| > \ell] - \Pr_{x\sim P_\mu}[|x| < \ell] \right| \leq \pi. \tag{150}$$

[7] In fact, this rejection sampling approach would achieve a runtime $O(n\sqrt{\ell})$, which is already an improvement over the $O(n\ell)$-time draw-by-draw algorithm from [CDL94], and has been described elsewhere in the survey sampling literature. However, as shown in the proof we can do better.

Let T denote the transition matrix of the update step in which we resample a random coordinate of x. This process is reversible and has stationary distribution P_μ. Furthermore, T has a spectral gap equal to $1/n$[8].

For $f, g : \{0,1\}^n \to \mathbb{C}$, define $\langle f, g\rangle \coloneqq \mathbb{E}_{x\sim P_\mu}[f(x)^* g(x)]$. Reversibility means that T is Hermitian with respect to this inner product and the gap means that

$$\langle f, (I-T)f\rangle \geq \frac{1}{n}\operatorname{Var}(f) = \frac{1}{n}\big(\langle f, f\rangle - \langle f, 1\rangle^2\big) \ . \tag{151}$$

(Here, the "1" in the inner product denotes the constant function equal to one.) Let $f = s - \mathbb{E}[s]$ so that $\langle f, 1\rangle = 0$. Rearranging the gap equation we obtain

$$\begin{aligned}\langle f, T^n f\rangle &\leq \left(1 - \frac{1}{n}\right)^n \langle f, f\rangle &(152)\\ &\leq e^{-1}(\langle s, s\rangle - \mathbb{E}[s]^2) &(153)\\ &\leq e^{-1}(1 - \mathbb{E}[s]^2) \ . &(154)\end{aligned}$$

On the other hand, $\langle f, T^n f\rangle = \langle s, T^n s\rangle - \mathbb{E}[s]^2$. Thus

$$\langle s, T^n s\rangle \leq e^{-1}(1-\mathbb{E}[s]^2) + \mathbb{E}[s]^2 \leq e^{-1} + (1-e^{-1})\pi^2 \ . \tag{155}$$

The left-hand side can also be related to the underlying walk as

$$\langle s, T^n s\rangle = \mathbb{E}[s(x(0))s(x(n))] \geq 2\Pr[s(x(0))s(x(n)) = 1] - 1. \tag{156}$$

Combining these we find that

$$\Pr[s(x(0))s(x(n)) = 1] \leq \frac{1}{2}(1+e^{-1})(1+\pi^2) \ . \tag{157}$$

Now suppose that $\pi \leq 1/2$. Then the right-hand side is strictly less than $7/8$ and there is a constant probability of success. On the other hand with probability π the first sample $x(0)$ is already in H_ℓ, so if $\pi > 1/2$ then we also obtain a constant probability of success. Either way, each iteration of the sampling algorithm succeeds with probability $\Omega(1)$, so the expected number of repetitions is $O(1)$. □

8.3 Inference

Finally we turn to the inverse problem of finding $\mu^\star \in \mathbb{R}^n$ such that $q(\mu^\star) \equiv \sum_{x\in H_\ell} p_{\mu^\star}(x)x = q$ given some input q, which we refer to as *inference*. This turns out to be the slowest part of our overall randomized truncation algorithm, and is accomplished by solving the dual optimization problem, Eq. (139). We discuss several algorithms for the task.

First we sketch why we should expect a polynomial-time solution in the first place. Naively Eq. (137) is a convex program with $|H_\ell| = \binom{n}{\ell}$ variables, while Eq. (139) has only n variables. This observation on its own, however, does not substantially reduce the complexity of the problem since the dual objective requires summing over $\binom{n}{\ell}$ terms. Fortunately, the exponentially large sum can be evaluated efficiently, due to Lemma 8.2, and this is the basis for efficient algorithms. One also

[8]To see this, observe that the entire nontrivial spectrum of T (not including 1) can be generated as an eigenvalue with corresponding eigenvector $\phi_S : \{0,1\}^n \to \mathbb{R}$ defined through $\phi_S(x) = \prod_{i\in S}(x_i - \frac{1}{1+e^{\mu_i}})$ for any $S \subseteq [n]$. The largest such eigenvalue comes from, say, $S = \{1\}$, and is equal to $1 - 1/n$.

needs to show that an accurate solution of Eq. (139) yields a good approximation of $p^\star$. For this, note that the dual objective function can be rewritten as

$$g_q(\mu) = \mathrm{H}(p^\star) + \mathrm{D}_{\mathrm{KL}}(p^\star \parallel p_\mu). \tag{158}$$

Then by Pinsker's inequality, an ε^2-accurate solution μ to the dual problem gives a solution p_μ to the primal problem which is $O(\varepsilon)$ in total variation distance from the optimum $p^\star$. For this reason we express our goal as ε^2-accuracy.

As in Section 8.2, we focus on arithmetic complexity. However, for the inference problem we also keep track of the complexity in terms of a boundary distance parameter η which satisfies

$$q_i \in [\eta, 1-\eta] \quad \text{for every } i \in [n]. \tag{159}$$

We will see later that the eigenvalue of the Hessian are related to $q_i(1-q_i)$, so we might expect that η affects the convergence of naive gradient descent. At the same time, we would ideally like the run time to scale polylogarithmically in η, since this is polynomial in the bit length of the input. We can also assume that $\eta \geq \varepsilon/n$, because one can round each q_i to be in $[\eta, 1-\eta]$ by incurring error $\leq n\eta$.

In Table 2, we summarize several algorithms for the task of minimizing Eq. (158). An algorithm with rigorous guarantees, running in time polynomial in n, $\log(1/\varepsilon)$, and $\log(1/\eta)$ can be obtained through the ellipsoid method, as shown by Straszak and Vishnoi [SV19]. This leads to the following, which we state without proof.

Theorem 8.5 (Implied by [SV19, Theorem 15])**.** *Assume $q_j \in [\eta, 1-\eta]$ for each $j \in [n]$, $\varepsilon \in (0,1]$, and $\sum_j q_j = \ell$. A variant of the ellipsoid algorithm runs in time $\mathrm{poly}(n, \log(1/\varepsilon), \log(1/\eta))$ and computes an ε^2-accurate solution $\mu \in \mathbb{R}^n$ to the dual problem in Eq. (139) such that $\|\mu\|_2 \leq \mathrm{poly}(n, \log(1/\varepsilon), \log(1/\eta))$.*

With this theorem in hand, we have enough to conclude the argument that a maximum-entropy construction satisfies the four desired criteria listed at the beginning of this section, and that a polynomial-size representation of this distribution, in the form of the parameters μ, can be computed in polynomial time. Nonetheless, one might hope for alternative methods with an explicit and favorable dependence on n, which we now provide.

In fact, we will present three such algorithms, along with performance guarantees. Each of these three algorithms outperforms the other two for some choice of parameters. We summarize their performance in Theorem 8.6 below. In particular, the box-constrained Newton algorithm (BN) runs asymptotically at least as quickly as the SVD used in MPS truncation, assuming that the SVD is done in time $O(n^3)$ instead of $O(n^\omega)$. Even if the SVD uses the $O(n^{2.807})$-time Strassen method, the runtime will be comparable, especially if ℓ is significantly smaller than n.

To get intuition for the optimization problem, observe that $g_q(\mu)$ is convex with gradient and Hessian given by

$$\nabla_\mu g_q(\mu) = q - q(\mu) \quad \text{and} \quad \nabla^2_\mu g_q(\mu) =: \Sigma(\mu) = Q(\mu) - q(\mu)q(\mu)^T, \tag{160}$$

which we can compute in time $O(n\ell)$ and $O(n^2)$ by Lemma 8.2. Letting $D(\mu)$ be the diagonal of $\Sigma(\mu)$, we will prove in Lemma 8.7 that $\Sigma(\mu) \preceq 2D(\mu)$. Since the entries of $D(\mu)$ are $q_i(\mu)(1-q_i(\mu))$, this means that the Hessian has eigenvalues that are always $\leq 1/2$ but approach zero as $q_i(\mu)$ approaches 0 or 1. If the Hessian were in fact $D(\mu)$ and not $\Sigma(\mu)$ then this would mean, in the language of convex optimization, that g_q would be $1/4$-smooth but only $O(\eta)$-strongly convex. (The comparison between $\Sigma(\mu)$ and $D(\mu)$ can be made more rigorous by Lemma 8.7 and Conjecture 8.10.)

As a result we can expect gradient descent to work but with some slowdowns depending on η (or equivalently $\|\mu\|_\infty$). These slowdowns can be dealt with using either Newton's method or a variant of gradient descent that does not require strong convexity.

APGD. The simplest algorithm we can use is gradient descent. A somewhat improved variant is the FISTA algorithm with constant step size from Section 4 of [BT09], also known as accelerated projected gradient descent (APGD). The algorithm starts with $\mu^{(0)} = \nu^{(1)} = 0$, $\beta_1 = 2$ and for $t = 1, \dots, T$ sets

$$\mu^{(t)} = \nu^{(t)} + \alpha(q - q(\nu^{(t)})) \tag{161}$$

$$\alpha \text{ is the largest value } \leq 2 \text{ s.t. } \left\|\mu^{(t)}\right\|_2 \leq R_2 \tag{162}$$

$$\beta_{t+1} = 1 + \sqrt{1 + \beta_t^2} \tag{163}$$

$$\nu^{(t+1)} = \mu^{(t)} + \frac{\beta_t - 2}{\beta_{t+1}}(\mu^{(t)} - \mu^{(t-1)}) \tag{164}$$

We will choose R_2 and T later. This iteration is dominated by the cost of computing $q(\cdot)$ which takes time $O(n\ell)$ by Lemma 8.2. The runtime guarantee requires that an approximate solution exists within a 2-norm ball of radius R_2. Lemma 8.8 below implies R_2 can be taken to be $O(\sqrt{n}\log(1/\eta))$, improving on the $R_2 \leq \mathrm{poly}(n)$ bound of [SV19]. Finally, Thm 4.4 of [BT09] states that

$$g_q(\mu^{(t)}) - g_q(\mu^*) \leq \frac{2\beta R_2^2}{(t+1)^2}, \tag{165}$$

implying that it suffices to take $T = \lceil \sqrt{2} R_2/\varepsilon \rceil \lesssim \sqrt{n}\log(1/\eta)\varepsilon^{-1}$ to achieve ε^2-accuracy. We believe that a bound of the form $T \leq \tilde{O}(\eta^{-1/2}\log(1/\varepsilon))$ is also likely, assuming a constrained version of Thm 3.18 of [Bub15] exists, but we do not explore this question further here.

BN. The cautious step sizes of APGD can be improved using Newton's method; specifically, the box-constrained Newton method from [CMTV17], for which the basic iterate is: given a function $f\colon \mathbb{R}^n \to \mathbb{R}$ to minimize and $k > 0$, set $x \to x + \frac{1}{k}\tilde{y}$ where $\tilde{y} \in \mathbb{R}^n$ is any vector such that $\|\tilde{y}\|_\infty \leq k$ and

$$\frac{1}{2}\tilde{y}^T A \tilde{y} + b^T \tilde{y} \leq \frac{1}{2} \cdot \min_{\|y\|_\infty \leq 1} \left(\frac{1}{2} y^T A y + b^T y\right) \tag{166}$$

for $A = \frac{\varepsilon^2}{k}\nabla^2 f(x)$ and $b = \frac{1}{k}\nabla f(x)$. In the case where A is symmetric and diagonally dominant (SDD), $\tilde{y}$ can be computed efficiently with $k = O(\log n)$, as shown in [CMTV17]. If, in addition, we have the *second-order robustness* condition

$$\frac{1}{c}\nabla^2 f(x) \preceq \nabla^2 f(x+h) \preceq c\nabla^2 f(x) \quad \text{for every } \|h\|_\infty \leq 1, \tag{167}$$

where $c > 0$ is some constant, then we can apply Theorem 3.2 in [CMTV17] to bound the runtime of the solver. Namely, an ε^2-approximate solution can be computed in time

$$\widetilde{O}\left(n^2(1 + R_\infty)\log\left(\frac{f(x_0) - f(x^\star)}{\varepsilon}\right)\right) \tag{168}$$

where $x_0 \in \mathbb{R}^n$ is some initial point, $x^\star \in \mathbb{R}^n$ is a minimizer of f, and R_∞ is a bound on the diameter of the level set at x_0. We will now show that if we define f through

$$f(x) = g_q\left(\frac{x}{4\ell}\right) + \frac{\varepsilon^2}{8n}\sum_{i=1}^n \cosh\left(\frac{x_i}{M}\right) \quad \text{where} \quad M := 8\ell\log(1/\eta), \tag{169}$$

then the above SDD condition and second-order robustness condition are satisfied, and $R_\infty = O(\ell \log(1/\eta) \log(n/\varepsilon))$ with the intial point $x_0 = 0$. This is similar to the regularization approach in Section 4 of [CMTV17]. Moreover, the choice of M allows us to conclude that an $\varepsilon^2/4$-minimizer x of f yields an $\varepsilon^2/2$-minimizer $\mu = x/(4\ell)$ of g_q, through

$$g_q(\mu) \le f(x) \le f(4\ell\mu^\star) + \frac{\varepsilon^2}{4} = g_q(\mu^\star) + \frac{\varepsilon^2}{8n} \sum_{i=1}^n \cosh\left(\frac{4\ell\mu_i^\star}{M}\right) + \frac{\varepsilon^2}{4} \le g_q(\mu^\star) + \frac{\varepsilon^2}{2} \tag{170}$$

where $\mu^\star$ is the minimizer of g_q, and we used Lemma 8.8 to assume $\|\mu^\star\|_\infty \le 2\log(1/\eta)$. In summary, this gives an approximate solution to our original optimization problem with the desired accuracy in time

$$\widetilde{O}\left(n^2 \ell \log(1/\eta) \log(1/\varepsilon)\right). \tag{171}$$

First, to see that $\nabla^2 f$ is SDD, write

$$\nabla^2 f(x) = \frac{1}{16\ell^2} \Sigma\Big(\frac{x}{4\ell}\Big) + \frac{\varepsilon^2}{8nM^2} \operatorname{diag}\Big(\cosh\Big(\frac{x_i}{M}\Big)\Big), \tag{172}$$

using the chain rule together with Eq. (160). The second term is a strictly positive diagonal matrix, and $\Sigma(\mu)$ is also SDD for every $\mu \in \mathbb{R}^n$, by Lemma 8.7.

Next, the fact that f is second-order robust with $c = \mathrm{e}$, follows from Lemma 8.9. In particular, this bounds the first term, while the second term can be bounded using hyperbolic trigonometric identities.

Finally, for the bound on the diameter of the level set of f at the initial point $x_0 = 0$, note that

$$f(0) = g_q(0) + \frac{\varepsilon^2}{8n} \sum_{i=1}^n \cosh(0) = \log\binom{n}{\ell} + \frac{\varepsilon^2}{8} \tag{173}$$

while, for any x such that $f(x) \le f(0)$ and any $i \in [n]$,

$$\frac{\varepsilon^2}{8n} \cosh\Big(\frac{x_i}{M}\Big) \le \frac{\varepsilon^2}{8n} \sum_{j=1}^n \cosh\Big(\frac{x_j}{M}\Big) \le f(0) - g_q\Big(\frac{x}{4\ell}\Big) \le \log\binom{n}{\ell} + \frac{\varepsilon^2}{8}. \tag{174}$$

Using $\cosh(t) \ge \mathrm{e}^{|t|}/2$ and $\log\binom{n}{\ell} \le n$, this implies $|x_i| \le M \cdot O\left(\log(n/\varepsilon)\right)$ for every i, giving $R_\infty = O(\ell \log(1/\eta) \log(n/\varepsilon))$ as claimed.

IPFF. This third algorithm and its analysis are entirely due to [CDL94], except for our faster algorithm for computing $q(\mu)$. IPFF improves a 1940 algorithm called Iterative Proportional Fitting [DS40] with the simple modification of holding the largest weight fixed. The algorithm assumes $q_n = \max_i q_i$ and initializes $w_i^{(0)} := q_i$. Let $z_{-i}(t)$ denote the partition function $Z_\mu^{(\ell-1)}$ where $e^{-\mu}$ is the vector length $n-1$ obtained by deleting the i-th entry of $w_i^{(t-1)}$. Then for $t = 1, 2, \ldots, T$ and $i \in [n-1]$ we set

$$w_i^{(t)} = q_i \frac{z_{-i}}{z_{-n}} \tag{175}$$

We set all $w_n^{(t)} = q_n$. The intuition behind Eq. (175) is that $w_i = e^{-\mu_i^\star}$ is the unique fixed point of the map and Theorem 2 of [CDL94] proves that the distance to this fixed point shrinks by at least q_n, implying that $T = O((1-q_n)^{-1} \log(1/\varepsilon)) \le O(\eta^{-1} \log(1/\varepsilon))$ steps suffice. To obtain the total runtime, our Lemma 8.2 shows how to carry out each step in time $O(n^2\ell)$.

We summarize the above discussion in the following theorem which compares the performance of all three algorithms. We also present these results with prior ones in Table 2.

Method	Upper Bound on Runtime	References
Ellipsoid	$\mathrm{poly}(n, \log(1/\varepsilon), \log(1/\eta))$	[SV19]
Box-constrained (BC) Newton	$n^2 \ell \log(1/\eta) \log(1/\varepsilon)$	Thm. 8.6, [CMTV17, CKV20]
Diagonal BC Newton	$n\ell^2 \log(1/\eta) \log(1/\varepsilon)$	assuming Conjecture 8.10
IPF	unknown	[DS40]
IPF with μ_n fixed	$n\ell^2\eta^{-1} \log(1/\varepsilon)$	[CDL94]
Accelerated PGD	$n^{1.5}\ell\varepsilon^{-1} \log(1/\eta)$	Thm. 8.6, [BT09]

Table 2: Comparison of different algorithms for the inference problem of finding $\mu^\star \in \mathbb{R}^n$ such that $q(\mu^\star) = q$, given $q \in \mathbb{R}^n$. Runtimes are written omitting the $\tilde{O}()$ and depend on n (length of bit string), ℓ (Hamming weight), ε (trace-distance accuracy), and $\eta = \min_i\{q_i, 1 - q_i\}$.

Theorem 8.6. *The following three algorithms each compute an ε^2-accurate solution $\mu \in \mathbb{R}^n$ to the dual problem in Eq. (139).*

1. Accelerated projected gradient descent (APGD) *in time* $O(n^{1.5}\ell\varepsilon^{-1} \log(\ell/\eta))$.
2. Box-constrained Newton's method (BN) *in time* $\widetilde{O}(n^2\ell \log(1/\eta) \log(1/\varepsilon))$.
3. Iterative Proportional Fitting with μ_n fixed (IPFF) *in time* $O(n\ell^2\eta^{-1} \log(1/\varepsilon))$.

The proof of Theorem 8.6 relied on the following lemmas.

Lemma 8.7. *The covariance matrix $\Sigma(\mu) = Q(\mu) - q(\mu)q(\mu)^T \in \mathbb{R}^{n\times n}$ is symmetric and diagonally dominant, and satisfies*

$$0 \preceq \Sigma(\mu) \preceq 2D(\mu) \tag{176}$$

where $D(\mu)$ is the diagonal approximation to $\Sigma(\mu)$ with entries $D(\mu)_{ii} = \Sigma(\mu)_{ii}$ and $D(\mu)_{ij} = 0$ for $i \neq j$.

Proof. The operator $\Sigma(\mu)$ is positive semidefinite because it is a covariance matrix: for any $v \in \mathbb{R}^n$ we have $\langle v, \Sigma v\rangle = \mathbb{E}[\langle v, q - x\rangle^2] \geq 0$. For the upper bound, note that $\sum_j Q(\mu)_{ij} = \ell q(\mu)_i$ and $\sum_j q(\mu)_j = \ell$, so $\sum_j \Sigma(\mu)_{ij} = 0$. Along with the fact that $\Sigma(\mu)_{ii} > 0$ and $\Sigma(\mu)_{ij} < 0$ for $i \neq j$, this implies that $\Sigma(\mu) - D(\mu) \preceq D(\mu)$. □

Lemma 8.8. *Given $\mu \in \mathbb{R}^n$ with $q = q(\mu)$, assume that $q_j \in [\eta, 1 - \eta]$ for each $j \in [n]$, Then* $\max_i \mu_i - \min_i \mu_i \leq 2\log(1/\eta)$.

In particular, we can always rescale μ (like in Lemma 8.4), so this implies that there exists an optimal μ^* with $\|\mu^*\|_\infty \leq 2\log(1/\eta)$.

Proof. Consider the ratio of probabilities (under the distribution p_μ)

$$\Delta_{ij} = \frac{\Pr[x_i = 1, x_j = 0]}{\Pr[x_i = 0, x_j = 1]} = \frac{q_i - Q_{ij}}{q_j - Q_{ij}} \tag{177}$$

using that $\Pr[x_i = 1, x_j = 0] = \Pr[x_i = 1] - \Pr[x_i = 1, x_j = 1]$ and similar for i and j reversed. On the other hand, for any x with $x_i = 1$ and $x_j = 0$, we have $p_\mu(x - e_i + e_j) = p_\mu(x) \cdot e^{-\mu_i+\mu_j}$. Summing over all such strings we find

$$\Delta_{ij} = e^{-\mu_i+\mu_j}. \tag{178}$$

Next, suppose $\mu_j \geq \mu_i$, so $q_j \leq q_i$. Recall that that we have negative correlation, so $Q_{ij} \leq q_i q_j$. Then we get

$$e^{\mu_j - \mu_i} = \Delta_{ij} = \frac{q_i - Q_{ij}}{q_j - Q_{ij}} \leq \frac{q_i - q_i q_j}{q_j - q_i q_j} = \frac{q_i(1-q_j)}{q_j(1-q_i)} \leq \frac{(1-\eta)^2}{\eta^2} \tag{179}$$

where we use that $t \mapsto \frac{q_i - t}{q_j - t}$ is an increasing function in t for $t < q_j$, and our assumption on q. We conclude that $\mu_j - \mu_i \leq 2\log(1/\eta)$. □

Finally, we have the following statement pertaining to the second-order robustness of the objective function g_q. The special case where $\ell = n$ appears in [CKV20].

Lemma 8.9. *Let $\mu, h \in \mathbb{R}^n$ and $\Sigma(\mu)$ be the Hessian defined in Eq. (160). It holds that*

$$\exp(-4\ell\|h\|_\infty)\Sigma(\mu) \preceq \Sigma(\mu + h) \preceq \exp(4\ell\|h\|_\infty)\Sigma(\mu). \tag{180}$$

Proof. Let $P = p_\mu$ and $P' = p_{\mu+h}$, so $P'(x) = P(x)e^{-\langle h, x\rangle} / \mathbb{E}_{y\sim P}[e^{-\langle h, y\rangle}]$ and define $\kappa \coloneqq \max_{x \in H_\ell} \frac{P'(x)}{P(x)}$. Note that we have

$$\kappa \leq e^{\max_{x,y \in H_\ell} \langle h, x-y\rangle} \leq e^{2\ell\|h\|_\infty}. \tag{181}$$

Also, for any function $F\colon \{0,1\}^n \to \mathbb{R}$ we have

$$\begin{aligned}
\operatorname*{Var}_{x\sim P'}[F(x)] &= \frac{1}{2} \mathop{\mathbb{E}}_{x\sim P'} \mathop{\mathbb{E}}_{y\sim P'} (F(x) - F(y))^2 && (182)\\
&\leq \kappa^2 \mathop{\mathbb{E}}_{x\sim P} \mathop{\mathbb{E}}_{y\sim P} (F(x) - F(y))^2 && (183)\\
&= \kappa^2 \operatorname*{Var}_{x\sim P}[F(x)]. && (184)
\end{aligned}$$

If $F(x) = \langle v, x\rangle$ for some $v \in \mathbb{R}^n$ then $\operatorname*{Var}_{x\sim P}[F(x)] = v^T\Sigma(\mu)v$ and $\operatorname*{Var}_{x\sim P'}[F(x)] = v^T\Sigma(\mu+h)v$. Since this holds for arbitrary $v \in \mathbb{R}^n$ we have $\Sigma(\mu+h) \preceq e^{4\ell\|h\|_\infty}\Sigma(\mu)$ by Eqs. (181) and (184), as required. The reverse inequality follows by applying the above with $\mu + h$ in place of μ and $-h$ in place of h, since $\|-h\|_\infty = \|h\|_\infty$. □

8.4 Heuristics and conjecture for faster inference

A natural approach that has been proposed [Che00] is to simply use Newton's method, which has the iterates

$$\mu^{(t+1)} = \mu^{(t)} - \alpha\Sigma(\mu^{(t)})^{-1}(q - q(\mu^{(t)})). \tag{185}$$

The parameter α can be set to 1 or found by line search. Finally Σ is technically noninvertible, since $\Sigma\mathbf{1} = 0$. So we need to take the inverse in the subspace orthogonal to $\mathbf{1}$. This is essentially the same as the BN method, with the ℓ_∞ box, and is a good heuristic algorithm even though it currently lacks a provable convergence guarantee.

There is a yet faster method that avoids the need to compute all of Q or to solve a linear system of equations. We can approximate the covariance matrix $\Sigma(\mu)$ with its diagonal $D(\mu) = \operatorname{diag}(q_i(\mu)(1 - q_i(\mu)))$. This means we only need to compute $q(\mu)$ at each step instead of $Q(\mu)$, so the time per iteration becomes $O(n\ell)$. In practice, this method appears to converge quickly to the optimal solution, and we use it in our numerics.

Finally we state a plausible conjecture which would imply a provable bound on this heuristic method. In Lemma 8.7 we proved that $0 \preceq \Sigma(\mu) \preceq 2D(\mu)$. We believe that the lower bound on Σ can be significantly improved. We conjecture that $D(\mu) \preceq \Sigma(\mu)$ on the subspace orthogonal to $\mathbf{1}$. Formally,

Conjecture 8.10. $\lambda_2(D(\mu)^{-1/2}\Sigma(\mu)D(\mu)^{-1/2}) \geq 1$ *where* λ_2 *denotes the second smallest eigenvalue.*

This conjecture was independently proposed by [AHM26]. We have run numerical tests and not found any counterexamples to this conjecture. In the difficult case where some q_i is very close to 0 or 1, λ_2 is only slightly larger than 1, suggesting that 1 is the exact numerical value of the minimum of $\lambda_2(D(\mu)^{-1/2}\Sigma(\mu)D(\mu)^{-1/2})$ over all μ. However, for algorithmic purposes it would be enough to prove that $\lambda_2(D(\mu)^{-1/2}\Sigma(\mu)D(\mu)^{-1/2}) \geq c$ for some universal constant $c > 0$.

If Conjecture 8.10 holds then we can replace $\Sigma(\mu)$ with $D(\mu)$ in the BN algorithm and will need only $O(1)$ more iterations. Now each iteration runs in time $O(n\ell)$ instead of $O(n^2)$, so the total runtime becomes $O(n\ell^2 \log(\eta/\varepsilon))$.

Note for arXiv v3: As noted in the introduction, we are posting this paper now to comply with the QIP 2027 requirement that submissions be available on arxiv.org. However, we expect to further update the paper in the near future. Recently [arXiv:2607.13990v2], a proof of Conjecture 8.10. With a little work, this implies that a Newton-type algorithm can find μ from q in time $\tilde{O}(n\ell)$. We will expand on this point in a future version. Even without this change, BC Newton is proven to run faster than SVD in Theorem 8.6.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator (Award No. DE-SCL0000121). This work was partially funded by the Wellcome Leap as part of the Quantum for Bio (Q4Bio) Program. AWH was supported by the Simons Foundation (MP-SIP-00001553, AWH), the NSF (PHY-2325080) and a Fulbright Scholar Grant. FW was supported by the European Union (ERC, ASC-Q, 101040624). We thank Garnet Chan, Jakob Günther, Allen Liu, John Martyn, Ankur Moitra, Norbert Schuch, and Ewin Tang for helpful conversations.

We used Claude Opus 5.0 to generate proof ideas for some of the algorithms in Section 8.3, and unsuccessfully used ChatGPT to attempt to prove Conjecture 8.10.

A Background material

For the convenience of the reader, we collect a few standard definitions and facts from convex analysis, nonlinear optimization and matrix analysis used in the proofs in the main text.

A.1 Minimax and zero-sum games

A minimax inequality of the form

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y) \leq \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) \tag{186}$$

holds for arbitrary choices of sets $\mathcal{X}, \mathcal{Y}$ and a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Here and throughout when we take maxima and minima of functions, it is implied that these values are attained. Such an

inequality admits a game-theoretic interpretation as follows. Consider a zero-sum, two-player game where the payoff function for the x-player is $-f(x,y)$ and that for the y-player is $+f(x,y)$. Then the right-hand side of Eq. (186) represents the worst-case payoff for the y-player when they "go second" in the game, i.e., the strategy for the y-player may depend on x. Through this lens, the minimax inequality amounts to the statement that going second in the game is a privilege. We encounter a specific example of such a two-player game in Section 4.

A minimax theorem claims that the minimax inequality holds with equality under certain conditions. One example which suffices for our purposes is Sion's Minimax Theorem, provided below. We refer the interested reader to [Sim95] for a history of other minimax theorems.

Theorem A.1 (Implied by Thm. 3 in [Fan52])**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact, convex subsets of a Hilbert space and $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a real-valued function such that:*

1. *$f(\cdot, y)$ is continuous and convex on $\mathcal{X}$ for every fixed $y \in \mathcal{Y}$, and*
2. *$f(x, \cdot)$ is continuous and concave on $\mathcal{Y}$ for every fixed $x \in \mathcal{X}$.*

Then

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y). \tag{187}$$

A related concept is the existence of a saddle point, also referred as an equilbrium point in the context of two-player games.

Definition A.2 (Saddle point)**.** Given arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$ and a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, a point $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ is a saddle point if

$$f(x^\star, y) \le f(x^\star, y^\star) \le f(x, y^\star) \tag{188}$$

for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

The following lemma states that the existence of a saddle point is equivalent to a minimax equality where both the minimum and the maximum are attained.

Lemma A.3 (Well-known)**.** *A saddle point $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ for the function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ exists if and only if*

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) =: f^\star. \tag{189}$$

Furthermore, for such a saddle point we have $f(x^\star, y^\star) = f^\star$.

Proof. Suppose Eq. (187) holds. Then we may choose

$$x^\star \in \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y), \qquad y^\star \in \arg\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y)$$

in which case

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y) = \min_{x\in\mathcal{X}} f(x,y^\star) \le f(x^\star,y^\star) \le \max_{y\in\mathcal{Y}} f(x^\star,y) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y)$$

and, by the minimax equality, the two inequalities hold with equality. Hence, the relations in Eq. (188) are satisfied.

On the other hand, suppose there is a saddle point $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$. Then

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) \le \max_{y\in\mathcal{Y}} f(x^\star,y) \le f(x^\star,y^\star) \le \min_{x\in\mathcal{X}} f(x,y^\star) \le \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y) \tag{190}$$

where the second and third inequalities use the relations in Eq. (188). But this is an inequality in the opposite direction of the minimax inequality, which always holds. □

A.2 Convex analysis

We now list some standard facts from convex analysis. See [Roc70, HUL01] for an introduction to these concepts. Throughout, we use the extended reals and let $\operatorname{dom} f$ denote the set of all values x for which $f(x)$ is finite.

Definition A.4 (Subgradient and subdifferential)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function. A vector $g \in \mathbb{R}^d$ is a *subgradient* of f at $x \in \mathbb{R}^d$ if

$$f(y) - f(x) \;\geq\; \langle g,\, y - x\rangle \qquad \text{for all } y \in \mathbb{R}^d. \tag{191}$$

The set of all such subgradients is the *subdifferential* of f at x, denoted $\partial f(x)$.

Some rules familiar from calculus are applicable in many cases.

Lemma A.5 (Implied by Theorem 23.8 in [Roc70])**.** *Let $f, g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex functions and suppose that the relative interiors of $\operatorname{dom} f$ and $\operatorname{dom} g$ have a point in common. Then for any $x \in \mathbb{R}^d$ it holds that $\partial(f + g)(x) = \partial f(x) + \partial g(x)$.*

Definition A.6 (Normal cone)**.** For $\Omega \subseteq \mathbb{R}^d$ and $x \in \Omega$, the *normal cone* of Ω at x is

$$\mathcal{N}_\Omega(x) = \{g \in \mathbb{R}^d :\; \langle g,\, y - x\rangle \leq 0 \text{ for all } y \in \Omega\}. \tag{192}$$

The above definition allows us to state the following first-order optimality condition, which is valid even when the objective function is not differentiable.

Lemma A.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function and $\Omega \subseteq \mathbb{R}^d$ be a convex set with nonempty relative interior. A point $x \in \Omega$ minimizes f over Ω if and only if*

$$0 \in \partial f(x) + \mathcal{N}_\Omega(x). \tag{193}$$

Proof. It is straightforward to verify that for any convex function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ it holds that x minimizes h if and only if $0 \in \partial h(x)$. This is sometimes referred to as *Fermat's Rule*. Now, the set of optimal solutions for $\min\{f(x) : x \in \Omega\}$ is identical to that for the optimization $\min\{f(x) + \delta_\Omega(x) : x \in \mathbb{R}^d\}$ where $\delta_\Omega : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ satisfies $\delta_\Omega(x) = 0$ if $x \in \Omega$ and $\delta_\Omega(x) = +\infty$ otherwise. Furthermore, making use of Definition A.6 we see that $\partial \delta_\Omega(x) = \mathcal{N}_\Omega(x)$ for any $x \in \Omega$. Putting $h = f + \delta_\Omega$ and applying Lemma A.5 completes the proof. □

If f is differentiable and the feasible set is specified using equality constraints only then the condition above gives back the usual method of Lagrange multipliers.

Given a convex set $\mathcal{X}$ in a finite-dimensional real inner-product space and $f : \mathcal{X} \to \mathbb{R}$, the *Fenchel conjugate* $f^* : \mathcal{X} \to \mathbb{R}$ is given by

$$f^*(y) = \sup_{x \in \mathcal{X}} \big\{\langle x, y\rangle - f(x)\big\}. \tag{194}$$

When f is a squared norm we have the following lemma. We omit the proof, which is a simple application of the Cauchy-Schwarz inequality.

Lemma A.8. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and set $f(x) = \frac{1}{2}\|x\|^2$. Then*

$$f^*(y) = \tfrac{1}{2}\|y\|_*^2, \tag{195}$$

where $\|\cdot\|_*$ is the dual norm, defined as $\|y\|_* = \sup_{x \in \mathcal{X}} \{\langle x, y\rangle : \|x\| \leq 1\}$.

Corollary A.9. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ with dual $\|\cdot\|_*$. For any $x, y \in \mathbb{R}^d$, the following are equivalent:*

1. *$y \in \partial(\frac{1}{2}\|\cdot\|^2)(x)$, and x optimizes $\max_{x'}\{\langle x', y\rangle - \frac{1}{2}\|x'\|^2\}$,*
2. *$x \in \partial(\frac{1}{2}\|\cdot\|_*^2)(y)$, and y optimizes $\max_{y'}\{\langle x, y'\rangle - \frac{1}{2}\|y'\|_*^2\}$, and*
3. *$\|x\|^2 + \|y\|_*^2 = 2\langle x, y\rangle$.*

Proof. To see that x optimizes $\max_{x'}\{\langle x', y\rangle - \frac{1}{2}\|x'\|^2\}$ whenever $y \in \partial(\frac{1}{2}\|\cdot\|^2)(x)$, or the analogous statement for y, use the definition of the subgradient. The equivalence between the first and third statements then follows from Lemma A.8, and similarly for the equivalence between the second and third statements. □

A.3 Spectral stability

The following is a well-known consequence of Weyl's Inequality concerning the spectral stability of Hermitian matrices. We let $\lambda_1(A), \lambda_2(A), \ldots$ denote the eigenvalues of a Hermitian matrix A sorted in nondecreasing order.

Fact A.10 (Spectral stability)**.** *Let $k, d \in \mathbb{Z}_{>0}$ be such that $1 \le k \le d$ and $A, B \in \mathbb{C}^{d\times d}$ be Hermitian matrices. It holds that*

$$\lambda_k(A+B) - \lambda_k(A) \in [\lambda_1(B), \lambda_d(B)].$$

B Proof of Lemma 2.11

We make use of the following optimization fact, the proof of which we omit.

Lemma B.1. *Let $a, b, w_1, w_2 \in \mathbb{R}_{>0}$ such that $b > a$ and $w_1, w_2 \in [a, b]$. The unique optimal solution to the optimization problem*

$$\begin{aligned} \max_{x\in\mathbb{R}^2}. \quad & w_1x_1 + w_2x_2 - \frac{1}{2}(x_1^2 + x_2^2) \\ \text{s.t.} \quad & b \ge x_1 \ge x_2 \ge a \end{aligned} \tag{196}$$

is given by $x_1 = x_2 = \frac{1}{2}(w_1 + w_2)$ if $w_1 < w_2$ and $x = (w_1, w_2)^\top$ otherwise.

We now prove Lemma 2.11. The sufficiency of the condition in Eq. (25) is explicitly shown in the proof of Proposition 2.1 in [AFS12], so we restrict ourselves to the argument for necessity. We may assume without loss of generality that v is nonnegative and sorted in nonincreasing order. We will first take x to be sorted as well before justifying this assumption at the end of the proof. Clearly, an optimal solution x which is sorted such that $x_1 \ge x_2 \ge \cdots \ge x_{\ell-1}$ exists.

First, similar to the proof of Proposition 2.1 in [AFS12], we give necessary conditions on the form of such an optimal solution. Consider the optimization

$$\begin{aligned} \max. \quad & f(x) \coloneqq \sum_{i=1}^{\ell-1} v_i x_i - \frac{1}{2}\sum_{i=1}^{k} x_i^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^d \\ & x_1 \ge x_2 \ge \cdots \ge x_{\ell-1} > 0. \end{aligned} \tag{197}$$

By the definition of ℓ, a sorted solution x to the maximization in Eq. (24) is optimal only if its first $\ell - 1$ entries are optimal for the problem above. Additionally, an optimal solution must have $x_k = x_{k+1} = \cdots = x_{\ell-1}$. We can rewrite the objective function for such an optimal solution as

$$f(x) = \sum_{i=1}^{k-2} v_i x_i - \frac{1}{2}\sum_{i=1}^{k-2} x_i^2 + v_{k-1}x_{k-1} + s_k^v x_k - \frac{1}{2}x_{k-1}^2 - \frac{1}{2}x_k^2. \tag{198}$$

If $s_k^v < v_{k-1}$ then the optimal solution must satisfy $x_i = v_i$ for $i = 1, \ldots, k-1$ and $x_k = s_k^v$. (The optimization just becomes the Fenchel conjugate of half the 2-norm squared in this case.) Otherwise, by Lemma B.1 we must have $x_{k-1} = x_k$ and therefore

$$f(x) = \sum_{i=1}^{k-3} v_i x_i - \frac{1}{2}\sum_{i=1}^{k-3} x_i^2 + v_{k-2}x_{k-2} + s_{k-1}^v x_{k-1} - \frac{1}{2}x_{k-2}^2 - x_{k-1}^2 \tag{199}$$

at the optimum. Continuing in this manner, we arrive at precisely the condition in the statement of the lemma.

Now, for *any* optimal solution x (not necessarily sorted), we must have $x_j \geq x_{j+1}$ for all $j \in [\ell-2]$ such that $v_j \neq v_{j+1}$; otherwise, the objective function is less than that obtained by permuting the indices j and $j+1$ in the vector x. On the other hand, if $v_j = v_{j+1}$ for some $j \in [\ell-2]$, then the objective function is unchanged by permuting the indices j and $j+1$ in the vector x. Therefore, if there exists an optimal solution with $x_j < x_{j+1}$ for some $j \in [\ell-2]$, this contradicts the above necessary condition on the form of the sorted optimal solution. In particular, we cannot have simultaneously that $x_j \neq x_{j+1}$ and $v_j = v_{j+1}$ for entries j and $j+1$ of x which match v. For the case where $j = k - r - 1$, we cannot have that the condition in Eq. (23) is satisfied when $v_j = v_{j+1}$.

References

[ABW15] Amir Abboud, Arturs Backurs, and Virginia Vassilevska Williams. If the current clique algorithms are optimal, so is Valiant's parser. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 98–117, 2015, arXiv:1504.01431.

[AFS12] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. *Advances in Neural Information Processing Systems*, 25, 2012, arXiv:1204.5043.

[AHM26] Nima Anari, Alireza Haqi, and Eric Ma. On rounding on the hypersimplex, 2026, arXiv:2606.00996.

[AKT22] Seiseki Akibue, Go Kato, and Seiichiro Tani. Quadratic improvement on accuracy of approximating pure quantum states and unitary gates by probabilistic implementation, 2022, arXiv:2111.05531.

[AKT24a] Seiseki Akibue, Go Kato, and Seiichiro Tani. Probabilistic state synthesis based on optimal convex approximation. *npj Quantum Information*, 10(1):3, 2024, arXiv:2303.10860.

[AKT24b] Seiseki Akibue, Go Kato, and Seiichiro Tani. Probabilistic unitary synthesis with optimal accuracy. *ACM Transactions on Quantum Computing*, 5(3):1–27, August 2024, arXiv:2301.06307.

[BBPS96] C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher. Concentrating partial entanglement by local operations. 53:2046–2052, 1996, arXiv:quant-ph/9511030.

[Bha97] Rajendra Bhatia. *Matrix Analysis*. Springer New York, 1997.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 11 2015, arXiv:1405.4980.

[Cam17] Earl Campbell. Shorter gate sequences for quantum computing by mixing unitaries. *Physical Review A*, 95(4), April 2017, arXiv:1612.02689.

[Cam19] Earl Campbell. Random compiler for fast hamiltonian simulation. *Physical Review Letters*, 123(7), August 2019, arXiv:1811.08017.

[CDL94] Xiang-Hui Chen, Arthur P. Dempster, and Jun S. Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3):457–469, 09 1994.

[CG19] Eric Chitambar and Gilad Gour. Quantum resource theories. *Reviews of modern physics*, 91(2):025001, 2019, arXiv:1806.06107.

[CGJ+16] Jianxin Chen, Shane Grogan, Nathaniel Johnston, Chi-Kwong Li, and Sarah Plosker. Quantifying the coherence of pure quantum states. *Phys. Rev. A*, 94:042313, Oct 2016, arXiv:1601.06269.

[Che00] Sean X Chen. General properties and estimation of conditional bernoulli models. *Journal of Multivariate Analysis*, 74(1):69–87, 2000.

[CHKT21] Chi-Fang Chen, Hsin-Yuan Huang, Richard Kueng, and Joel A Tropp. Concentration for random product formulas. *PRX Quantum*, 2(4):040305, 2021, arXiv:2008.11751.

[CK87] David G Cantor and Erich Kaltofen. Fast multiplication of polynomials over arbitrary rings. *Acta Inf*, 28, 1987.

[CKV20] L. Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1349–1359. PMLR, 13–18 Jul 2020, arXiv:1906.02164.

[CMTV17] Michael B Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained newton's method and interior point methods. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 902–913. IEEE, 2017, arXiv:1704.02310.

[COS19] Andrew M. Childs, Aaron Ostrander, and Yuan Su. Faster quantum simulation by randomization. *Quantum*, 3:182, September 2019, arXiv:1805.08385.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, October 2012, arXiv:1012.0621.

[CRS10] Dragoš M Cvetković, Peter Rowlinson, and Slobodan Simić. *An introduction to the theory of graph spectra*, volume 75. Cambridge university press Cambridge, 2010.

[DS40] W Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936.

[Fan52] Ky Fan. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proceedings of the National Academy of Sciences of the United States of America*, 38(2):121–126, 1952.

[Fer15] Andrew J Ferris. Unbiased Monte Carlo for the age of tensor networks, 2015, arXiv:1507.00767.

[Gha08] Sevag Gharibian. Strong NP-hardness of the quantum separability problem, 2008, arXiv:0810.4507.

[GKZ25] David Gosset, Robin Kothari, and Chenyi Zhang. Multi-qubit Toffoli with exponentially fewer T gates, 2025, arXiv:2510.07223.

[GLS12] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

[Gur03] Leonid Gurvits. Classical deterministic complexity of Edmonds' problem and quantum entanglement. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 10–19, 2003, arXiv:quant-ph/0303055.

[GWS+25] Jakob Günther, Freek Witteveen, Alexander Schmidhuber, Marek Miller, Matthias Christandl, and Aram Harrow. Phase estimation with partially randomized time evolution, 2025, arXiv:2503.05647.

[Has16] M. B. Hastings. Turning gate synthesis errors into incoherent errors, 2016, arXiv:1612.01011.

[HJJ20] Jeremy Heng, Pierre E. Jacob, and Nianqiao Ju. A simple Markov chain for independent Bernoulli variables conditioned on their sum, 2020, arXiv:2012.03103.

[HL13] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013, arXiv:0911.1393.

[HUL01] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Berlin Heidelberg, 2001.

[HW23] Matthew Hagan and Nathan Wiebe. Composite quantum simulations. *Quantum*, 7:1181, November 2023, arXiv:2206.06409.

[JLP+18] Nathaniel Johnston, Chi-Kwong Li, Sarah Plosker, Yiu-Tung Poon, and Bartosz Regula. Evaluating the robustness of k-coherence and k-entanglement. *Phys. Rev. A*, 98(2):022328, 2018, arXiv:1806.00653.

[JS68] Kumar Jogdeo and Stephen M Samuels. Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. *The Annals of Mathematical Statistics*, 39(4):1191–1195, 1968.

[KMB24] Bálint Koczor, John JL Morton, and Simon C Benjamin. Probabilistic interpolation of quantum rotation angles. *Physical Review Letters*, 132(13):130602, 2024, arXiv:2305.19881.

[KT24] Anastasia Kireeva and Joel A. Tropp. Randomized matrix computations: themes and variations, 2024, arXiv:2402.17873.

[LBT19] Zi-Wen Liu, Kaifeng Bu, and Ryuji Takagi. One-shot operational quantum resource theory. *Phys. Rev. Lett.*, 123:020401, Jul 2019, arXiv:1904.05840.

[Low21] Guang Hao Low. Halving the cost of quantum multiplexed rotations, 2021, arXiv:2110.13439.

[MF85] Govind S. Mudholkar and Marshall Freimer. A structure theorem for the polars of unitarily invariant norms. *Proceedings of the American Mathematical Society*, 95(3):331–337, 1985.

[Mir60] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[MR25] John M. Martyn and Patrick Rall. Halving the cost of quantum algorithms with randomization. *npj Quantum Information*, 11(1), March 2025, arXiv:2409.03744.

[NC12] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2012.

[OWC20] Yingkai Ouyang, David R. White, and Earl T. Campbell. Compilation by stochastic Hamiltonian sparsification. *Quantum*, 4:235, February 2020, arXiv:1910.06255.

[Reg17] Bartosz Regula. Convex geometry of quantum resource quantification. *Journal of Physics A: Mathematical and Theoretical*, 51(4):045303, Dec 2017, arXiv:1707.06298.

[Roc70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[SAP17] Alexander Streltsov, Gerardo Adesso, and Martin B. Plenio. Colloquium: Quantum coherence as a resource. *Reviews of Modern Physics*, 89(4), October 2017, arXiv:1609.02439.

[Sch11] Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96–192, 2011. January 2011 Special Issue.

[Sim95] Stephen Simons. Minimax theorems and their proofs. In Ding-Zhu Du and Panos M. Pardalos, editors, *Minimax and Applications*, pages 1–23. Springer US, Boston, MA, 1995.

[SV14] Mohit Singh and Nisheeth K. Vishnoi. Entropy, optimization and counting. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, page 50–59, New York, NY, USA, 2014. Association for Computing Machinery, arXiv:1304.8108.

[SV19] Damian Straszak and Nisheeth K. Vishnoi. Maximum entropy distributions: Bit complexity and stability. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2861–2891. PMLR, 25–28 Jun 2019, arXiv:1711.02036.

[TH00] Barbara M. Terhal and Paweł Horodecki. Schmidt number for density matrices. *Phys. Rev. A*, 61:040301, Mar 2000, arXiv:quant-ph/9911117.

[Til06] Yves Tillé. *Sampling Algorithms*. Springer New York, NY, 2006.

[Tod24] Synge Todo. Markov Chain Monte Carlo in tensor network representation, 2024, arXiv:2412.02974.

[VT99] Guifre Vidal and Rolf Tarrach. Robustness of entanglement. *Phys. Rev. A*, 59:141–155, 1999, arXiv:quant-ph/9806094.

[WZYZ25] Yue Wang, Xiao-Ming Zhang, Xiao Yuan, and Qi Zhao. Faster state preparation with randomization, 2025, arXiv:2510.12247.