SIAM J. SCI. COMPUT.
Vol. 46, No. 6, pp. A3427–A3447

FASTER RANDOMIZED PARTIAL TRACE ESTIMATION*

TYLER CHEN†, ROBERT CHEN†, KEVIN LI†, SKAI NZEUTON‡, YILU PAN§, AND YIXIN WANG†

Abstract. We develop randomized matrix-free algorithms for estimating partial traces, a generalization of the trace arising in quantum physics and chemistry. Our algorithm improves on the typicality-based approach used in [T. Chen and Y-C. Cheng, *J. Chem. Phys.,* 157 (2022), 064106] by deflating important subspaces (e.g., corresponding to the low-energy eigenstates) explicitly. This results in a significant variance reduction, leading to several order-of-magnitude speedups over the previous state of the art. We then apply our algorithm to the study of the thermodynamics of several Heisenberg spin systems, particularly the entanglement spectrum and ergotropy.

Key words. partial trace, deflation, stochastic trace, Krylov subspace method

MSC codes. 68Q25, 68R10, 68U05

DOI. 10.1137/23M1620399

1. Introduction.

The state of a quantum system is described by a *density matrix* of dimension exponential in the system size. Often we are interested in the state of a subsystem of the total system. This can be obtained by taking the *partial trace* of the total system density matrix, which yields a density matrix of dimension depending only on the size of the subsystem of interest (called the reduced system density matrix). This matrix can then be used to understand important properties of the subsystem, for instance, its entanglement with the rest of the total system [44]. Recently, a number of numerical methods have been developed to estimate partial traces [66, 13, 10].

If the total system density matrix is known explicitly, computing the partial trace is trivial. Unfortunately, owing to the exponential dependence of the total system density matrix on the system size, it is typically prohibitively expensive to obtain (or even store) the total system density matrix. Hope is not lost; in many situations, the total system density matrix has an implicit representation in terms of a (typically sparse) Hamiltonian $\mathbf{H}$ describing the configuration of the system of interest. For instance, the total system density matrix may be proportional to $\exp(-\beta\mathbf{H})$ for some parameter $\beta > 0$ or might be obtained from $\mathbf{H}$ by solving the Schrödinger equation with a given initial condition.

Mathematically, this means that the task of computing reduced density matrices can often be viewed as computing the partial trace of a *matrix function* of $\mathbf{H}$. In many situations, the Hamiltonian is sparse and admits implicit matrix-vector products (i.e., we can compute the map $\mathbf{x} \mapsto \mathbf{H}\mathbf{x}$). Thus, for moderately sized systems for which it is possible to store dense vectors, Krylov subspace methods offer an attractive

*Submitted to the journal's Numerical Algorithms for Scientific Computing section December 5, 2023; accepted for publication (in revised form) July 2, 2024; published electronically November 4, 2024.

https://doi.org/10.1137/23M1620399

Funding: This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. The first author was partially supported by NSF grants 2045590 and 2427363.

†New York University, New York, NY 10012 USA (tyler.chen@nyu.edu, https://research.chen.pw, rc4571@nyu.edu, xl3556@nyu.edu, yw5073@nyu.edu).

‡Cornell University, Ithaca, NY 10003 USA (san82@cornell.edu).

§New York University Shanghai, Pudong New District, Shanghai, China (yp2129@nyu.edu).

potential approach. Indeed, such methods are widely used for the closely related task of trace estimation [60, 31, 58, 67, 57, 56, 32, 5].

2. Background. We begin by providing some background on a natural setting in which partial traces arise, as well as on existing methods for implicit partial trace approximation.

2.1. Equilibrium reduced density matrices. Throughout, the total system is defined on a finite dimensional Hilbert space $\mathcal{H}_\mathrm{t} = \mathcal{H}_\mathrm{s} \otimes \mathcal{H}_\mathrm{b}$, where $\mathcal{H}_\mathrm{s}$ and $\mathcal{H}_\mathrm{b}$ are the Hilbert spaces for subsystem (s) and subsystem (b), respectively. We assume the total system is governed by a Hamiltonian

$$\mathbf{H} = \bar{\mathbf{H}}_\mathrm{s} + \bar{\mathbf{H}}_\mathrm{b} + \mathbf{H}_\mathrm{sb}, \tag{2.1}$$

where $\bar{\mathbf{H}}_\mathrm{s} = \mathbf{H}_\mathrm{s} \otimes \mathbf{I}_\mathrm{b}$ corresponds to the Hamiltonian of subsystem (s), $\bar{\mathbf{H}}_\mathrm{b} = \mathbf{I}_\mathrm{s} \otimes \mathbf{H}_\mathrm{b}$ corresponds to the Hamiltonian of subsystem (b), and $\mathbf{H}_\mathrm{sb}$ accounts for *nonnegligible* interactions between the two subsystems.

When the total system is in thermal equilibrium at inverse temperature β (due to weak coupling with a "superbath"), the state of the system is described by a density matrix

$$\boldsymbol{\rho}_\mathrm{t} = \boldsymbol{\rho}_\mathrm{t}(\beta) = \frac{\exp(-\beta\mathbf{H})}{Z_\mathrm{t}}, \qquad Z_\mathrm{t} = Z_\mathrm{t}(\beta) = \mathrm{tr}(\exp(-\beta\mathbf{H})); \tag{2.2}$$

see, for instance, [22, 64, 1]. The quantity $Z_\mathrm{t}(\beta)$ is called the partition function and provides insight into a number of thermodynamic properties of the system.

Often, we are interested in the state of subsystem (s) rather than the total system. If subsystem (s) did not interact with subsystem (b) (i.e., if $\mathbf{H}_\mathrm{sb} = \mathbf{0}$), then the density matrix for subsystem (s) would simply be proportional to $\exp(-\beta\mathbf{H}_\mathrm{s})$. However, when the interactions between subsystems (s) and (b) are nonnegligible, the density matrix $\boldsymbol{\rho}^*$ for subsystem (s) is instead obtained by "tracing out"[1] the effects of subsystem (b), i.e.,

$$\boldsymbol{\rho}^* = \boldsymbol{\rho}^*(\beta) = \mathrm{tr}_\mathrm{b}(\boldsymbol{\rho}_\mathrm{t}) = \frac{\mathrm{tr}_\mathrm{b}(\exp(-\beta\mathbf{H}))}{\mathrm{tr}(\exp(-\beta\mathbf{H}))}, \tag{2.3}$$

where $\mathrm{tr}_b(\,\cdot\,)$ is the *partial trace* over subsystem (b) [9, 30, 61].

2.2. Partial traces. Let d_s and d_b be the dimension of $\mathcal{H}_\mathrm{s}$ and $\mathcal{H}_\mathrm{b}$, respectively, so that $d_\mathrm{t} = d_\mathrm{s}d_\mathrm{b}$ is the dimension of $\mathcal{H}_\mathrm{t} = \mathcal{H}_\mathrm{s} \otimes \mathcal{H}_\mathrm{b}$. A general matrix $\mathbf{A} : \mathcal{H}_\mathrm{t} \to \mathcal{H}_\mathrm{t}$ can be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \cdots & \mathbf{A}_{1,d_\mathrm{s}} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \cdots & \mathbf{A}_{2,d_\mathrm{s}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{d_\mathrm{s},1} & \mathbf{A}_{d_\mathrm{s},2} & \cdots & \mathbf{A}_{d_\mathrm{s},d_\mathrm{s}} \end{bmatrix}, \tag{2.4}$$

where $\mathbf{A}_{i,j} : \mathcal{H}_\mathrm{b} \to \mathcal{H}_\mathrm{b}$ for each i, j. The partial trace of $\mathbf{A}$ over $\mathcal{H}_\mathrm{b}$ is defined as

$$\mathrm{tr}_\mathrm{b}(\mathbf{A}) := \begin{bmatrix} \mathrm{tr}(\mathbf{A}_{1,1}) & \mathrm{tr}(\mathbf{A}_{1,2}) & \cdots & \mathrm{tr}(\mathbf{A}_{1,d_\mathrm{s}}) \\ \mathrm{tr}(\mathbf{A}_{2,1}) & \mathrm{tr}(\mathbf{A}_{2,2}) & \cdots & \mathrm{tr}(\mathbf{A}_{2,d_\mathrm{s}}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{tr}(\mathbf{A}_{d_\mathrm{s},1}) & \mathrm{tr}(\mathbf{A}_{d_\mathrm{s},2}) & \cdots & \mathrm{tr}(\mathbf{A}_{d_\mathrm{s},d_\mathrm{s}}) \end{bmatrix}. \tag{2.5}$$

[1] This is analogous to "integrating out" variables from a joint probability distribution to obtain the marginal distribution for a variable of interest.

From (2.5) it is clear that the partial trace is easy to compute if we have an explicit representation of $\mathbf{A}$. However, we are most interested in the case $\mathbf{A} = \exp(-\beta\mathbf{H})$, where $\mathbf{H}$ is so large that storing and/or computing an explicit representation of $\mathbf{A}$ is intractable. As such, we will consider only methods which access $\mathbf{A}$ through matrix-vector products, which can then be approximated using Krylov subspace methods.

2.3. Stochastic trace estimation. Consider a real matrix $\mathbf{M} : \mathcal{H}_\mathrm{b} \to \mathcal{H}_\mathrm{b}$. If $\mathbf{v} \in \mathcal{H}_\mathrm{b}$ is a random vector whose entries are independent and identically distributed (i.i.d.) standard real Gaussian random variables, it is straightforward to show [39] that

$$\mathbb{E}\big[\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v}\big] = \mathrm{tr}(\mathbf{M}), \qquad \mathbb{V}\big[\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v}\big] = \frac{1}{2}\|\mathbf{M} + \mathbf{M}^\mathsf{T}\|_\mathsf{F}^2 \le 2\|\mathbf{M}\|_\mathsf{F}^2. \tag{2.6}$$

The use of estimators of this form[2] (although possibly with a different distribution) for approximating the trace of implicit matrices has been used since the late 1980s [23, 60, 29]. Theoretical tail bounds appear in the physics [52, 53, 24] and numerical analysis [3, 54, 40, 14] literature. These bounds control the probability that the estimator $\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v}$ is far from the trace in terms of properties of $\mathbf{M}$, such as $\|\mathbf{M}\|_\mathsf{F}$ and $\|\mathbf{M}\|_2$.

Standard trace estimators can be extended to partial traces [10]. In particular,

$$(\mathbf{I}_{d_\mathrm{s}} \otimes \mathbf{v})^\mathsf{T}\mathbf{A}(\mathbf{I}_{d_\mathrm{s}} \otimes \mathbf{v}) = \begin{bmatrix} \mathbf{v}^\mathsf{T}\mathbf{A}_{1,1}\mathbf{v} & \mathbf{v}^\mathsf{T}\mathbf{A}_{1,2}\mathbf{v} & \cdots & \mathbf{v}^\mathsf{T}\mathbf{A}_{1,d_\mathrm{s}}\mathbf{v} \\ \mathbf{v}^\mathsf{T}\mathbf{A}_{2,1}\mathbf{v} & \mathbf{v}^\mathsf{T}\mathbf{A}_{2,2}\mathbf{v} & \cdots & \mathbf{v}^\mathsf{T}\mathbf{A}_{2,d_\mathrm{s}}\mathbf{v} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}^\mathsf{T}\mathbf{A}_{d_\mathrm{s},1}\mathbf{v} & \mathbf{v}^\mathsf{T}\mathbf{A}_{d_\mathrm{s},2}\mathbf{v} & \cdots & \mathbf{v}^\mathsf{T}\mathbf{A}_{d_\mathrm{s},d_\mathrm{s}}\mathbf{v} \end{bmatrix} \tag{2.7}$$

provides an unbiased estimator for $\mathrm{tr}_\mathrm{b}(\mathbf{A})$ when $\mathbf{v}$ is sampled as described above. Given i.i.d. copies $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of $\mathbf{v}$, we arrive at an estimator

$$\widehat{\mathsf{tr}}_\mathrm{b}^m(\mathbf{A}) := \frac{1}{m}\sum_{i=1}^m (\mathbf{I}_{d_\mathrm{s}} \otimes \mathbf{v}_i)^\mathsf{T}\mathbf{A}(\mathbf{I}_{d_\mathrm{s}} \otimes \mathbf{v}_i). \tag{2.8}$$

This estimator was studied in [10] for approximating reduced density matrices and will serve as the backbone of the algorithms developed in this paper.

The variance of a random matrix $\mathbf{X}$ can be defined as

$$\mathbb{V}[\mathbf{X}] := \mathbb{E}\Big[\big\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\big\|_\mathsf{F}^2\Big]. \tag{2.9}$$

This is equivalent to the sum of the variances of the entries of $\mathbf{X}$, so assuming $\mathbf{v}$ has i.i.d. standard normal entries, we find that

$$\mathbb{V}\Big[\widehat{\mathsf{tr}}_\mathrm{b}^m(\mathbf{A})\Big] = \frac{1}{m}\mathbb{V}\Big[(\mathbf{I}_{d_\mathrm{s}} \otimes \mathbf{v})^\mathsf{T}\mathbf{A}(\mathbf{I}_{d_\mathrm{s}} \otimes \mathbf{v})\Big] \le \frac{1}{m}\sum_{i=1}^{d_\mathrm{s}}\sum_{j=1}^{d_\mathrm{s}} 2\|\mathbf{A}_{i,j}\|_\mathsf{F}^2 = \frac{2}{m}\|\mathbf{A}\|_\mathsf{F}^2. \tag{2.10}$$

As with all Monte Carlo estimators, which output a sample average, the estimator (2.8) often suffers from large variance with fluctuations about the mean on the order of $\|\mathbf{A}\|_\mathsf{F}/\sqrt{m}$. When computing quantities of the form $\mathrm{tr}_\mathrm{b}(\mathbf{A})/\mathrm{tr}(\mathbf{A})$, we see that these

[2] In quantum physics, such estimators are closely related to the idea of quantum typicality [59, 65], which refers to the idea that, in many cases, a random state is representative of the overall state of a system; see [25] for a review.

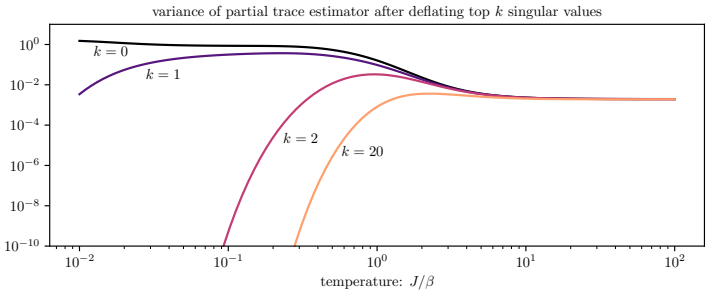

FIG. 1. *Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{d_t}$ be the singular values of $\mathbf{A}$. Then the variance of the partial trace estimator (2.7) is bounded above by $2\|\mathbf{A}\|_{\mathsf{F}}^2 = 2(\sigma_1^2 + \cdots + \sigma_{d_t}^2)$. By deflating the top k singular values, we can reduce the variance to at most $2(\sigma_{k+1}^2 + \cdots + \sigma_{d_t}^2)$ (see section 3). Here we take $\mathbf{A} = \boldsymbol{\rho}_t = \exp(-\beta\mathbf{H})/\operatorname{tr}(\exp(-\beta\mathbf{H}))$, where $\mathbf{H}$ corresponds to the solvable spin chain with $N = 10$ and $h = 0.3$ (described in subsection 5.5). We then plot the variance bound $2(\sigma_{k+1}^2 + \cdots + \sigma_{d_t}^2)$ for several values of k and a range of β. Takeaway: The $k = 0$ curve exhibits high variance at low temperature. Through the use of deflation, the variance can be reduced significantly.*

effects become particularly pronounced when $\mathbf{A}$ has a very quickly decaying spectrum. Indeed, if $\mathbf{A}$ has one dominating (positive) eigenvalue, then $\|\mathbf{A}\|_{\mathsf{F}}/\operatorname{tr}(\mathbf{A}) \approx 1$, whereas if $\mathbf{A}$ is close to the identity, then $\|\mathbf{A}\|_{\mathsf{F}}/\operatorname{tr}(\mathbf{A}) \approx 1/\sqrt{d_t}$. In the case $\mathbf{A} = \boldsymbol{\rho}_t = \exp(-\beta\mathbf{H})/\operatorname{tr}(\exp(-\beta\mathbf{H}))$, this corresponds to difficulties when β is very large (low temperature). We visualize how the variance depends on the temperature for this matrix in Figure 1.

In the zero-temperature limit $\beta \to \infty$, the trace $\operatorname{tr}(\exp(-\beta\mathbf{H}))$ is determined entirely by the smallest (most negative) eigenvalue of $\mathbf{H}$, and the partial trace $\operatorname{tr}_b(\exp(-\beta\mathbf{H}))$ by the corresponding eigenvector (assuming a one dimensional eigenspace). This means a randomized estimator such as (2.8) is not needed! Rather, one can simply apply classical techniques for obtaining extremal eigenvectors. This paper is built on the fact that at low (but nonzero) temperatures, knowledge of the eigenvectors corresponding to small eigenvalues is still very useful. In particular, we provide a deflation-based technique, which can significantly reduce the variance of (2.8) at low temperatures. Our approach is closely related to [43] and other deflation-based approaches for regular trace estimation [67, 55, 21]. The potential for variance reduction is also visualized in Figure 1.

2.4. Contributions. The primary contribution of this paper is a variance reduction technique for (2.7). This results in several order-of-magnitude speedups over the current state-of-the-art algorithm for approximating partial traces of matrix functions [10]. As such, we are able to study properties of quantum systems too large for existing methods. While similar variance reduction approaches have been used for regular trace estimation [23, 67, 38, 69, 21, 43, 40], partial traces do not satisfy a cyclic property, which makes generalizing past work (numerically) difficult. We propose some solutions to this difficulty.

Another contribution of this paper is highlighting some important problems from quantum physics which have been under-explored by the scientific computing community. We feel that there is significant potential for increased collaboration between these fields which is currently limited by a lack of cross-disciplinary knowledge transfer.

3. A variance reduced algorithm. We now describe a general technique for reducing the variance of the estimator (2.8) for an arbitrary matrix $\mathbf{A}$. By the linearity of the partial trace, for any matrix $\widetilde{\mathbf{A}}$,

$$\mathrm{tr}_{\mathrm{b}}(\mathbf{A}) = \mathrm{tr}_{\mathrm{b}}(\widetilde{\mathbf{A}}) + \mathrm{tr}_{\mathrm{b}}(\mathbf{A} - \widetilde{\mathbf{A}}). \tag{3.1}$$

If we are able to compute the partial trace of the first term exactly, we can estimate the partial trace of $\mathrm{tr}_{\mathrm{b}}(\mathbf{A})$ by applying the randomized estimator (2.7) to the residual term, that is, by

$$\mathrm{tr}_{\mathrm{b}}(\mathbf{A}) \approx \mathrm{tr}_{\mathrm{b}}(\widetilde{\mathbf{A}}) + \widehat{\mathsf{tr}}_{\mathrm{b}}^{m}(\mathbf{A} - \widetilde{\mathbf{A}}). \tag{3.2}$$

The variance of such an estimate is entirely due to the variance of $\widehat{\mathsf{tr}}_{\mathrm{b}}^{m}(\mathbf{A} - \widetilde{\mathbf{A}})$. Thus, if $\|\mathbf{A} - \widetilde{\mathbf{A}}\|_{\mathsf{F}}^2 < \|\mathbf{A}\|_{\mathsf{F}}^2$, then the variance of the estimator on the right-hand side of (3.2) is reduced compared to that of $\widehat{\mathsf{tr}}_{\mathrm{b}}^{m}(\mathbf{A})$.

Splittings similar to (3.2) have previously been used as a variance reduction technique for regular trace estimation [23, 67, 38, 69, 21, 43]. Perhaps the most widely known approach in the numerical analysis and theoretical computer science communities is the Hutch++ algorithm [40], which produces a $1 \pm \epsilon$ relative approximation to the trace of a positive semidefinite matrix using just $O(\epsilon^{-1})$ matrix-vector products with $\mathbf{A}$. Several improvements to this algorithm have been proposed [49, 18], including for the case $\mathbf{A} = f(\mathbf{H})$ [50, 11].

3.1. Partial trace of low-rank matrices. The partial trace of rank-1 matrix can be computed efficiently given a factorization as an outer product. In particular, for any $\mathbf{x} \in \mathcal{H}_{\mathrm{t}}$ we can write the outer product as

$$\mathbf{x}\mathbf{x}^{\mathsf{T}} = \begin{bmatrix} \mathbf{x}_{(1)}\mathbf{x}_{(1)}^{\mathsf{T}} & \mathbf{x}_{(1)}\mathbf{x}_{(2)}^{\mathsf{T}} & \cdots & \mathbf{x}_{(1)}\mathbf{x}_{(d_{\mathrm{s}})}^{\mathsf{T}} \\ \mathbf{x}_{(2)}\mathbf{x}_{(1)}^{\mathsf{T}} & \mathbf{x}_{(2)}\mathbf{x}_{(2)}^{\mathsf{T}} & \cdots & \mathbf{x}_{(2)}\mathbf{x}_{(d_{\mathrm{s}})}^{\mathsf{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{(n_v)}\mathbf{x}_{(1)}^{\mathsf{T}} & \mathbf{x}_{(d_{\mathrm{s}})}\mathbf{x}_{(2)}^{\mathsf{T}} & \cdots & \mathbf{x}_{(d_{\mathrm{s}})}\mathbf{x}_{(d_{\mathrm{s}})}^{\mathsf{T}} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} \mathbf{x}_{(1)} \\ \mathbf{x}_{(2)} \\ \vdots \\ \mathbf{x}_{(d_{\mathrm{s}})} \end{bmatrix}. \tag{3.3}$$

Using the fact that $\mathrm{tr}(\mathbf{x}_{(i)}\mathbf{x}_{(j)}^{\mathsf{T}}) = \mathbf{x}_{(i)}^{\mathsf{T}}\mathbf{x}_{(j)}$, we find

$$\mathrm{tr}_{\mathrm{b}}(\mathbf{x}\mathbf{x}^{\mathsf{T}}) = \begin{bmatrix} \mathbf{x}_{(1)}^{\mathsf{T}}\mathbf{x}_{(1)} & \mathbf{x}_{(2)}^{\mathsf{T}}\mathbf{x}_{(1)} & \cdots & \mathbf{x}_{(d_{\mathrm{s}})}^{\mathsf{T}}\mathbf{x}_{(1)} \\ \mathbf{x}_{(1)}^{\mathsf{T}}\mathbf{x}_{(2)} & \mathbf{x}_{(2)}^{\mathsf{T}}\mathbf{x}_{(2)} & \cdots & \mathbf{x}_{(d_{\mathrm{s}})}^{\mathsf{T}}\mathbf{x}_{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{(1)}^{\mathsf{T}}\mathbf{x}_{(d_{\mathrm{s}})} & \mathbf{x}_{(2)}^{\mathsf{T}}\mathbf{x}_{(d_{\mathrm{s}})} & \cdots & \mathbf{x}_{(d_{\mathrm{s}})}^{\mathsf{T}}\mathbf{x}_{(d_{\mathrm{s}})} \end{bmatrix}. \tag{3.4}$$

This observation and the linearity of the partial trace allow us to efficiently compute the partial trace of a generic symmetric rank-k matrix

$$\widetilde{\mathbf{A}} = \sum_{i=1}^{k} \theta_i \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \tag{3.5}$$

given access to the factors $(\theta_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, k$.

3.2. Implicit partial trace estimation. It is clear that choosing $\widetilde{\mathbf{A}}$ as a rank-k approximation to $\mathbf{A}$ will suit our needs. In particular, let $\mathbf{Q} \in \mathbb{R}^{d \times k}$ be a matrix with orthonormal columns ($\mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}_k$), and define

$$\widetilde{\mathbf{A}} := \mathbf{Q}\mathbf{Q}^{\mathsf{T}}\mathbf{A}\mathbf{Q}\mathbf{Q}^{\mathsf{T}}. \tag{3.6}$$

We can efficiently obtain a factorization of the form (3.5) using just k matrix-vector products with $\mathbf{A}$. Indeed, form $\mathbf{G} := \mathbf{Q}^{\mathsf{T}}\mathbf{A}\mathbf{Q}$, compute an eigendecomposition

Algorithm 3.1 Variance reduced partial trace estimation.

procedure PARTIAL TRACE($\mathbf{A}, \mathbf{Q}, m$)
  $\mathbf{\Theta}, \mathbf{S} = \text{EIG}(\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q})$ ▷ $k \times k$ matrix
  $\mathbf{X} = \mathbf{Q}\mathbf{S}$ ▷ $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_k]$
  for $i = 1, 2, \dots, k$ **do**
    $\mathbf{B}_{\text{defl}}^{(i)} = \theta_i \operatorname{tr}_{\text{b}}(\mathbf{x}_i \mathbf{x}_i^\mathsf{T})$ ▷ using (3.4)
  for $i = 1, 2, \dots, m$ **do**
    $\mathbf{Y} = (\mathbf{I}_{d_{\text{s}}} \otimes \mathbf{v})$, $\mathbf{v}$ is length d_{b} i.i.d. Gaussian vector
    $\mathbf{B}_{\text{rem}}^{(i)} = \mathbf{Y}^\mathsf{T}\mathbf{A}\mathbf{Y} - \mathbf{Y}^\mathsf{T}\mathbf{X}\mathbf{\Theta}\mathbf{X}^\mathsf{T}\mathbf{Y}$
  return $\widehat{\mathsf{tr}}_{\text{b}}^{m}(\mathbf{A}; \mathbf{Q}) = \sum_{i=1}^{r} \mathbf{B}_{\text{defl}}^{(i)} + \frac{1}{m}\sum_{i=1}^{m} \mathbf{B}_{\text{rem}}^{(i)}$

$\mathbf{G} =: \sum_{i=1}^{k} \theta_i \mathbf{s}_i \mathbf{s}_i^\mathsf{T}$, and set $\mathbf{x}_i := \mathbf{Q}\mathbf{s}_i$ to obtain a rank-k approximation to (3.6) of the form (3.5).

Then, rewriting (3.2), we arrive at a computationally feasible estimator,

$$\widehat{\mathsf{tr}}_{\text{b}}^{m}(\mathbf{A}; \mathbf{Q}) := \sum_{i=1}^{k} \theta_i \operatorname{tr}_{\text{b}}(\mathbf{x}_i \mathbf{x}_i^\mathsf{T}) + \widehat{\mathsf{tr}}_{\text{b}}^{m}(\mathbf{A} - \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q}\mathbf{Q}^\mathsf{T}). \tag{3.7}$$

We provide pseudocode for computing (3.7) and a corresponding error estimate in Algorithm 3.1. The total number of matrix-vector products with $\mathbf{A}$ is $k + md_{\text{s}}$; k products are used to compute $\mathbf{A}\mathbf{Q}$ in Algorithm 3.1, and in each of the m loops, d_{s} products are used to compute $\mathbf{A}\mathbf{Y}$ in Algorithm 3.1.

3.3. Choosing the projection space. Algorithm 3.1 takes as input the matrix $\mathbf{Q}$, which determines the projection space used for deflation, and the quality of the output depends strongly on $\mathbf{Q}$. For regular trace estimation, a natural approach is to obtain $\mathbf{Q}$ by sketching $\mathbf{A}$ [40]. While sketching can be used to generate $\mathbf{Q}$ for use in Algorithm 3.1, there are some practical difficulties for the case $\mathbf{A} = f(\mathbf{H})$, which is the focus of this paper. We discuss these difficulties in section 4.

Another reasonable choice is to take $\mathbf{Q}$ aligned with eigenvectors of $\mathbf{A}$. In fact, the choice of $\mathbf{Q}$ with k columns, which minimizes the variance of Algorithm 3.1 is to take $\mathbf{Q}$ as the eigenvectors corresponding to the k eigenvalues of $\mathbf{A}$ with largest magnitude (i.e. corresponding to the top k singular values of $\mathbf{A}$). In this case

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q}\mathbf{Q}^\mathsf{T}\|_{\mathsf{F}}^2 = \min_{\operatorname{rank}(\widetilde{\mathbf{A}})=k} \|\mathbf{A} - \widetilde{\mathbf{A}}\|_{\mathsf{F}}^2 = \sum_{i=k+1}^{d_{\text{t}}} \sigma_i^2, \tag{3.8}$$

where $\{\sigma_i\}$ are the singular values of $\mathbf{A}$ arranged in non-increasing order. In the case $\mathbf{A} = \exp(-\beta\mathbf{H})$, the singular values of $\mathbf{A}$ are $\exp(-\beta\lambda_i)$, where λ_i are the eigenvalues of $\mathbf{H}$. When β is large (low-temperature), several of these singular values are significantly larger than the others, and deflation is effective at decreasing the norm.

To illustrate this idea quantitatively, suppose that, for some fixed constants $c, \alpha \in (0, 1)$ and $c', c'' > 0$,

$$c'(c\alpha)^i < \sigma_i < c''\alpha^i, \qquad i = 1, 2, \dots, d_{\text{t}}. \tag{3.9}$$

Then, if $\mathbf{Q}$ contains the k eigenvectors of $\mathbf{A}$ corresponding to the largest magnitude eigenvalues, the variance reduced estimator (3.7) satisfies

$$\mathbb{V}\left[\widehat{\mathsf{tr}}_{\text{b}}^{m}(\mathbf{A}; \mathbf{Q})\right]^{1/2} < C\,\alpha^k\,\mathbb{V}\left[\widehat{\mathsf{tr}}_{\text{b}}^{m}(\mathbf{A})\right]^{1/2} \tag{3.10}$$

for some $C = C(\alpha, c, c', c'')$ that does not depend on k or the dimension d_{t}.[3] In other words, if the singular values of $\mathbf{A}$ decay exponentially, deflating the top k eigenvalues results in an exponential decrease in the magnitude of the fluctuations of the partial trace estimator (3.7) over the basic estimator (2.8) from [10].

4. Partial traces of matrix functions. The primary focus of this paper is on estimating $\mathrm{tr}_{\mathrm{b}}(\exp(-\beta\mathbf{H}))$; i.e., the partial trace of a matrix proportional to the density matrix $\boldsymbol{\rho}$ describing the state of the total system in thermal equilibrium at inverse temperature β. In this section, we describe an implementation of Algorithm 3.1 for general $f(\mathbf{H})$, with a particular focus on the case $f(x) = \exp(-\beta x)$. In the case of the standard trace $\mathrm{tr}(f(\mathbf{H}))$, this is commonly addressed using a combination of the typicality estimator $\mathbf{v}^{\mathsf{T}} f(\mathbf{H})\mathbf{v}$ and Krylov subspace methods [60, 31, 67, 63, 28, 57, 56, 32, 12].

Algorithm 3.1 requires computing quantities like

$$\mathbf{Y}^{\mathsf{T}}\mathbf{A}\mathbf{Y} - \mathbf{Y}^{\mathsf{T}}(\mathbf{Q}\mathbf{Q}^{\mathsf{T}}\mathbf{A}\mathbf{Q}\mathbf{Q}^{\mathsf{T}})\mathbf{Y}, \qquad \mathbf{Y} := (\mathbf{I}_{d_s} \otimes \mathbf{v}), \tag{4.1}$$

which can be difficult to compute accurately due to the potential for cancellation errors. In particular, each term of the difference may be much larger than the difference itself, so an accurate approximation to each term in a relative sense need not yield a good relative (or even additive) approximation to the difference. For Algorithm 3.1, where products with $\mathbf{A}$ are assumed to be exact, this is not an issue. However, efficient methods for computing products with $\mathbf{A} = \exp(-\beta\mathbf{H})$, such as time-stepping and Krylov subspace methods [17, 20, 41], result in some level of approximation error.

We will consider mainly the case where $\mathbf{Q}$ contains eigenvectors of $\mathbf{H}$ and hence of $\mathbf{A} = f(\mathbf{H})$. In subsection 4.2.3 we discuss how one may be able to avoid the cancellation errors of (4.1) for other $\mathbf{Q}$ with orthonormal columns. We also discuss some pros and cons of various choices of $\mathbf{Q}$ and potential implementation difficulties, particularly in the context of $f(x) = \exp(-\beta x)$.

Write the eigendecomposition of $\mathbf{H}$ as

$$\mathbf{H} = \begin{bmatrix}\mathbf{Q} & \widehat{\mathbf{Q}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda} & \\ & \widehat{\boldsymbol{\Lambda}}\end{bmatrix}\begin{bmatrix}\mathbf{Q}^{\mathsf{T}} \\ \widehat{\mathbf{Q}}^{\mathsf{T}}\end{bmatrix}, \tag{4.2}$$

where $\mathbf{Q}$ contains r eigenvectors and $\boldsymbol{\Lambda}$ the corresponding r eigenvalues. We will assume that $\mathbf{Q}$ and $\boldsymbol{\Lambda}$ can be computed exactly. Since there are many black-box and problem-dependent techniques for this, we do not discuss particular methods for obtaining these quantities.

Using (4.2), we see that $f(\mathbf{H})$ can be decomposed as

$$f(\mathbf{H}) = \mathbf{Q}\mathbf{Q}^{\mathsf{T}} f(\boldsymbol{\Lambda})\mathbf{Q}\mathbf{Q}^{\mathsf{T}} + \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{\mathsf{T}} f(\widehat{\boldsymbol{\Lambda}})\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{\mathsf{T}}. \tag{4.3}$$

This implies

$$\mathbf{Q}\mathbf{Q}^{\mathsf{T}} f(\mathbf{H})\mathbf{Q}\mathbf{Q}^{\mathsf{T}} = \mathbf{Q} f(\boldsymbol{\Lambda})\mathbf{Q}^{\mathsf{T}}, \tag{4.4}$$

[3] Since $n \geq 1$ and $c, \alpha \in (0,1)$, $\alpha^{2n} > 0$, and $(c\alpha)^{2n} < (c\alpha)^2 < \alpha^2$. Therefore,

$$\frac{\sum_{i=k+1}^{d_{\mathrm{t}}} \sigma_i^2}{\sum_{i=1}^{d_{\mathrm{t}}} \sigma_i^2} < \frac{c'}{c'}\frac{\sum_{i=k+1}^{n} \alpha^{2i}}{\sum_{i=1}^{d_{\mathrm{t}}} (c\alpha)^{2i}} = \frac{c'}{c'}\frac{(\alpha^{2k} - \alpha^{2d_{\mathrm{t}}})(1-(c\alpha)^2)}{c^2(1-\alpha^2)(1-(c\alpha)^{2d_{\mathrm{t}}})} < \frac{c'}{c'}\frac{\alpha^{2k}}{c^2(1-\alpha^2)^2}. \tag{3.11}$$

Rearranging and taking a square root gives (3.10).

which can be easily computed given $\mathbf{Q}$ and $\mathbf{\Lambda}$. Thus, using (4.3) and the fact that the orthogonal projector $\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^\mathsf{T}$ can equivalently be written $(\mathbf{I}-\mathbf{Q}\mathbf{Q}^\mathsf{T})$, we obtain the following alternate expression of the difference (4.1):

$$\mathbf{Y}^\mathsf{T} f(\mathbf{H})\mathbf{Y} - \mathbf{Y}^\mathsf{T}\mathbf{Q}\mathbf{Q}^\mathsf{T} f(\mathbf{H})\mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{Y} = \mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z}, \qquad \mathbf{Z} := (\mathbf{I}-\mathbf{Q}\mathbf{Q}^\mathsf{T})\mathbf{Y}. \tag{4.5}$$

This allows us to avoid cancellation errors by ensuring that our approximation to $\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z}$ respects the fact that $\mathbf{Z}$ is orthogonal to $\mathbf{Q}$.

4.1. The Lanczos algorithm with deflation. In order to approximate the quantity

$$\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z}, \qquad \mathbf{Z} := (\mathbf{I}-\mathbf{Q}\mathbf{Q}^\mathsf{T}), \mathbf{Y} \tag{4.6}$$

we make use of the block-Lanczos algorithm with explicit deflation. The block-Lanczos method implicitly constructs an orthonormal basis $\mathbf{V} = [\mathbf{V}_0, \ldots, \mathbf{V}_{t-1}]$ for the block-Krylov subspace

$$\operatorname{span}\{\mathbf{Z}, \mathbf{H}\mathbf{Z}, \ldots, \mathbf{H}^{t-1}\mathbf{Z}\} \tag{4.7}$$

such that for all $j = 0, 1, \ldots, t-1$,

$$\operatorname{span}\{\mathbf{V}_0, \ldots, \mathbf{V}_j\} = \operatorname{span}\{\mathbf{Z}, \mathbf{H}\mathbf{Z}, \ldots, \mathbf{H}^j\mathbf{Z}\}. \tag{4.8}$$

In addition, the algorithm outputs a symmetric block-tridiagonal matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{M}_0 & \mathbf{R}_1^\mathsf{T} & & \\ \mathbf{R}_1 & \ddots & \ddots & \\ & \ddots & \ddots & \mathbf{R}_{t-2}^\mathsf{T} \\ & & \mathbf{R}_{t-2} & \mathbf{M}_{t-1} \end{bmatrix} \tag{4.9}$$

satisfying $\mathbf{T} = \mathbf{V}^\mathsf{T}\mathbf{H}\mathbf{V}$. Here we assume that the block-Krylov subspace does not become degenerate.

Since the input $\mathbf{Z}$ is orthogonal to the eigenvectors $\mathbf{Q}$, in exact arithmetic $\mathbf{V}$ will be entirely orthogonal to $\mathbf{Q}$ as well. However, in finite precision arithmetic this cannot be guaranteed, and rounding errors might introduce small components in the directions of the $\mathbf{Q}$. These errors can grow rapidly. Thus, the block-Lanczos algorithm should be implemented to explicitly maintain orthogonality against $\mathbf{Q}$. Such an implementation is given in Algorithm 4.1.

In exact arithmetic, the Lanczos approximation to $\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z}$ is given by

$$\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z} \approx \mathbf{R}_0^\mathsf{T}\mathbf{E}_1^\mathsf{T} f(\mathbf{T})\mathbf{E}_1\mathbf{R}_0, \tag{4.10}$$

where $\mathbf{E}_1 = \mathbf{e}_1 \otimes \mathbf{I}$ and $\mathbf{R}_0$ is the R factor in the QR factorization of $\mathbf{Z}$. This approximation is a block-Gauss quadrature approximation and is exact if f is polynomial of degree at most $2t-1$ [26, section 6.6], [19, Theorem 2.7]. From this, we can obtain a simple bound for the convergence.

Let $\mathsf{Error} = \|\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z} - \mathbf{R}_0^\mathsf{T}\mathbf{E}_1^\mathsf{T} f(\mathbf{T})\mathbf{E}_1\mathbf{R}_0\|$, and suppose p is a polynomial of degree at most $2t-1$. Then, with $\mathbf{P}_\mathbf{Q} = \mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathsf{T}$,

Algorithm 4.1 Block-Lanczos algorithm with deflation.

procedure BLOCK-LANCZOS-DEFL($\mathbf{H}, \mathbf{Z}, \mathbf{Q}, t$)
  $\mathbf{V}_0, \mathbf{R}_0 = \text{QR}(\mathbf{Z} - \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{Z})$,
  for $j = 0, 1, \ldots, t-1$ **do**
    $\mathbf{X} = \mathbf{H}\mathbf{V}_j - \mathbf{V}_{j-1}\mathbf{R}_{j-1}^\mathsf{T}$ ▷ if $j = 0$, $\mathbf{X} = \mathbf{H}\mathbf{V}_0$
    $\mathbf{M}_j = \mathbf{V}_j^\mathsf{T}\mathbf{X}$
    $\mathbf{X} = \mathbf{X} - \mathbf{V}_j\mathbf{M}_j$
    $\mathbf{X} = \mathbf{X} - \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{X}$ ▷ Explicit deflation
    optionally, reorthogonalize $\mathbf{X}$ against $\mathbf{V}_0, \ldots, \mathbf{V}_{j-1}$
    $\mathbf{V}_{j+1}, \mathbf{R}_j = \text{QR}(\mathbf{X})$
  return $\mathbf{T}, \mathbf{R}_0$ ▷ $\mathbf{T}$ as defined in (4.9)

$$\mathsf{Error} = \|\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z} - \mathbf{Z}^\mathsf{T} p(\mathbf{H})\mathbf{Z} + \mathbf{R}_0^\mathsf{T}\mathbf{E}_1^\mathsf{T} p(\mathbf{T})\mathbf{E}_1\mathbf{R}_0 - \mathbf{R}_0^\mathsf{T}\mathbf{E}_1^\mathsf{T} f(\mathbf{T})\mathbf{E}_1\mathbf{R}_0\| \tag{4.11}$$

$$\leq \|\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z} - \mathbf{Z}^\mathsf{T} p(\mathbf{H})\mathbf{Z}\| + \|\mathbf{R}_0^\mathsf{T}\mathbf{E}_1^\mathsf{T} p(\mathbf{T})\mathbf{E}_1\mathbf{R}_0 - \mathbf{R}_0^\mathsf{T}\mathbf{E}_1^\mathsf{T} f(\mathbf{T})\mathbf{E}_1\mathbf{R}_0\| \tag{4.12}$$

$$\leq \|\mathbf{Z}\|^2 \|\mathbf{P}_\mathbf{Q}(f(\mathbf{H}) - p(\mathbf{H}))\mathbf{P}_\mathbf{Q}\| + \|\mathbf{R}_0\|^2 \|f(\mathbf{T}) - p(\mathbf{T})\|. \tag{4.13}$$

Note that, with $\hat{\Lambda}$ denoting the diagonal entries of $\hat{\mathbf{\Lambda}}$ defined in (4.2),

$$\|\mathbf{P}_\mathbf{Q}(f(\mathbf{H}) - p(\mathbf{H}))\mathbf{P}_\mathbf{Q}\| = \max_{x \in \hat{\Lambda}} |f(x) - p(x)|. \tag{4.14}$$

Likewise, since $\mathbf{T} = \mathbf{V}^\mathsf{T}\mathbf{H}\mathbf{V} = \mathbf{V}^\mathsf{T}\mathbf{P}_\mathbf{Q}\mathbf{H}\mathbf{P}_\mathbf{Q}\mathbf{V}$, the eigenvalues of $\mathbf{T}$ are contained in the convex closure $\text{conv}(\hat{\Lambda})$ of $\hat{\Lambda}$. Thus,

$$\|f(\mathbf{T}) - p(\mathbf{T})\| = \max_{x \in \text{spec}(\mathbf{T})} |f(x) - p(x)| \leq \max_{x \in \text{conv}(\hat{\Lambda})} |f(x) - p(x)|. \tag{4.15}$$

Then, using the facts that $\|\mathbf{Z}\| = \|\mathbf{R}_0\|$ and that p was arbitrary, we obtain the bound

$$\|\mathbf{Z}^\mathsf{T} f(\mathbf{H})\mathbf{Z} - \mathbf{R}_0^\mathsf{T}\mathbf{E}_1^\mathsf{T} f(\mathbf{T})\mathbf{E}_1\mathbf{R}_0\| \leq 2\|\mathbf{Z}\| \min_{\deg(p) < 2t} \max_{x \in \text{conv}(\hat{\Lambda})} |f(x) - p(x)|. \tag{4.16}$$

Without deflation, note that $\hat{\Lambda} = \Lambda$, the set of eigenvalues of $\mathbf{H}$. However, even when r is small, $\text{conv}(\hat{\Lambda})$ can be much smaller than $\text{conv}(\Lambda)$. In such cases, deflation helps not only with variance reduction of the partial trace estimator but also with the matrix function approximation. This will also be an important consideration when we discuss the use of other projection spaces in subsection 4.2.3.

4.1.1. A note on finite precision arithmetic. In exact arithmetic, the reorthogonalization step of Algorithm 4.1 is unnecessary as $\mathbf{X}$ is already orthogonal to $\mathbf{V}_0, \ldots, \mathbf{V}_{j-1}$. However, in finite precision arithmetic, failure to orthogonalize against these vectors at each iteration can lead to a drastic loss of orthogonality in $\mathbf{V}$. In practice the formal expression (4.5) still converges in all examples we have observed. This has been rigorously justified for the standard Lanczos method for approximating quadratic forms of matrix functions ($d_\text{s} = 1$) without deflation [34]. The analysis in [34] is based on a careful analysis of the Lanczos algorithm in finite precision arithmetic [47, 48]. However, to the best of our knowledge there is no similar analysis of the block-Lanczos algorithm or the Lanczos algorithm with deflation.

Algorithm 4.2 Variance reduced partial trace estimation for matrix functions.

procedure PARTIAL TRACE-FUNC($\mathbf{H}, f, k, m, t$)
  Compute eigenvectors/values $\mathbf{Q}, \mathbf{\Lambda}$ ▷ $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_k]$, $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_k)$
  for $i = 1, 2, \dots, k$ **do**
    $\mathbf{B}_{\text{defl}}^{(i)} = f(\lambda_i) \operatorname{tr}_{\text{b}}(\mathbf{q}_i \mathbf{q}_i^{\mathsf{T}})$ ▷ using (3.4)
  for $i = 1, 2, \dots, m$ **do**
    $\mathbf{Y} = (\mathbf{I}_{d_{\text{s}}} \otimes \mathbf{v})$, $\mathbf{v}$ is length d_{b} i.i.d. Gaussian vector
    $\mathbf{Z} = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathsf{T}})\mathbf{Y}$ ▷ Deflation
    $\mathbf{T}, \mathbf{R}_0 =$BLOCK-LANCZOS-DEFL$(\mathbf{H}, \mathbf{Z}, \mathbf{Q}, t)$
    $\mathbf{B}_{\text{rem}}^{(i)} = \mathbf{R}_0^{\mathsf{T}} \mathbf{E}_1^{\mathsf{T}} f(\mathbf{T}) \mathbf{E}_1 \mathbf{R}_0$ ▷ $\mathbf{E}_1 = \mathbf{e}_1 \otimes \mathbf{I}$
  return $\widehat{\operatorname{tr}}_{\text{b}}^{m}(\mathbf{A}; \mathbf{Q}) = \sum_{i=1}^{k} \mathbf{B}_{\text{defl}}^{(i)} + \frac{1}{m} \sum_{i=1}^{m} \mathbf{B}_{\text{rem}}^{(i)}$

4.2. Algorithm. We now have all the tools required to implement a version of Algorithm 3.1 in the case $\mathbf{A} = f(\mathbf{H})$. The resulting algorithm is summarized in Algorithm 4.2. In the context of regular trace estimation ($d_{\text{s}} = 1$), similar approaches have been used successfully [21, 67, 43]. In particular, [43] studies the task of computing the partition function $Z(\beta)$ for a range of β, a task closely related to our main application of focus.

4.2.1. Computational costs. Algorithm 4.2 requires computing the k eigenvalues/vectors of $\mathbf{H}$, the cost of which is context dependent. The remaining number of matrix-vector products with $\mathbf{H}$ is mtd_s: in each of the m loops, d_s products are required to compute $\mathbf{H}\mathbf{V}_j$ in each of the t iterations of the block-Lanczos algorithm. The parameters m and t, respectively, control the statistical variance of the partial trace estimator and the accuracy with which products with $f(\mathbf{H})$ are approximated. We note that the matrix-vector products for each of the m samples can be computed in parallel.

4.2.2. Limitations and extensions. The Lanczos-based method described in this section requires the storage of roughly d_{s} dense vectors of length d_{t}, as well as repeated matrix-vector products with the total system Hamiltonian $\mathbf{H}_{\text{t}}$. Since d_{t} depends exponentially on the system size, this approach is only viable for moderately sized systems far from the thermodynamic limit.

The partial trace estimator (2.7) uses $\mathbf{v} \in \mathcal{H}_{\text{b}}$ drawn from the uniform distribution on the hypersphere of radius $\sqrt{d_{\text{b}}}$. However, the analogous estimator is still unbiased so long as $\mathbb{E}[\mathbf{v}\mathbf{v}^{\mathsf{T}}] = \mathbf{I}_{\text{b}}$. This opens the possibility of using an appropriate distribution on tensor network states [45, 46]. While tensor network versions of the Lanczos algorithm have been studied [16], a more common imaginary time evolution approach [45, 51, 15, 35] is likely suitable for approximating the action of $\exp(-\beta \mathbf{H}_{\text{t}})$ for sufficiently low temperatures, at least as long as the total system has sufficiently local interactions.

4.2.3. Using arbitrary projection matrices. We hope to obtain a matrix $\mathbf{Q}$ which reduces the variance of the partial trace estimator nearly as much as when using the exact top eigenspace more efficiently than computing the top eigenspace exactly. For any matrix $\mathbf{Q}$ with orthonormal columns, it can be verified that

$$\mathbf{A} - \mathbf{Q}\mathbf{Q}^{\mathsf{T}}\mathbf{A}\mathbf{Q}\mathbf{Q}^{\mathsf{T}} = \frac{1}{2}\Big[(\mathbf{I} + \mathbf{Q}\mathbf{Q}^{\mathsf{T}})\mathbf{A}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathsf{T}}) + (\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathsf{T}})\mathbf{A}(\mathbf{I} + \mathbf{Q}\mathbf{Q}^{\mathsf{T}})\Big]. \tag{4.17}$$

Introduce matrices

$$\mathbf{Z} := (\mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathsf{T})\mathbf{Y}, \qquad \mathbf{W} := (\mathbf{I} + \mathbf{Q}\mathbf{Q}^\mathsf{T})\mathbf{Y}. \tag{4.18}$$

Then, in place of (4.1), we can use the mathematically equivalent expression

$$\frac{1}{2}\Big[\mathbf{W}^\mathsf{T}\mathbf{A}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{A}\mathbf{W}\Big]. \tag{4.19}$$

We expect this expression to be less prone to rounding errors so long as $\mathbf{W}^\mathsf{T}\mathbf{A}\mathbf{Z}$ is far from skew-symmetric.

A common way to efficiently obtain an orthonormal matrix $\mathbf{Q}$ for which the error $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q}\mathbf{Q}^\mathsf{T}\|_\mathsf{F}^2$ is small is to take $\mathbf{Q} = \operatorname{orth}(\mathbf{A}\mathbf{\Omega})$, where $\mathbf{\Omega}$ is a $d_\mathrm{t} \times k$ random matrix with standard normal entries. The resulting approximation is commonly called the randomized SVD and requires k matrix-vector products with $\mathbf{A}$. When $\mathbf{A}$ has quickly decaying eigenvalues, the resulting $\mathbf{Q}$ results in an approximation nearly as good as the exact top eigenspace. For matrices with more slowly decaying eigenvalues, there are more complicated algorithms [27, 62].

When $\mathbf{A} = \exp(-\beta\mathbf{H}_\mathrm{t})/\operatorname{tr}(\exp(-\beta\mathbf{H}_\mathrm{t}))$, there are a number of challenges to using an approximate top subspace rather than the true top eigenspace. First, computing matrix-vector products with $\mathbf{A}$ requires the use of some sort of iterative method using products with $\mathbf{H}_\mathrm{t}$. While there are tools for this [17, 20, 41], these tools are not as mature as eigensolvers. Second, we would like to use a single matrix $\mathbf{Q}$ for all values of β in some range of interest. The best $\mathbf{Q}$ will be obtained by applying the randomized SVD to the matrix corresponding to the largest value of β; in fact, as $\beta \to \infty$, this will result in obtaining exactly the top subspace. Finally, and perhaps most subtly, when we exactly deflate the top eigenspace of $\mathbf{A}$, as noted in (4.16), the convergence of the iterative method used to compute products with $\mathbf{A}$ is accelerated. This acceleration is often significant, especially when β is large. This means the cost savings of using an approximate top subspace must be compared with the additional overhead of subsequent products with $\mathbf{A}$.

In Figure 2 we plot the quantity

$$2\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q}\mathbf{Q}^\mathsf{T}\|_\mathsf{F}^2, \qquad \mathbf{A} = \exp(-\beta\mathbf{H}_\mathrm{t})/\operatorname{tr}(\exp(-\beta\mathbf{H}_\mathrm{t})), \tag{4.20}$$

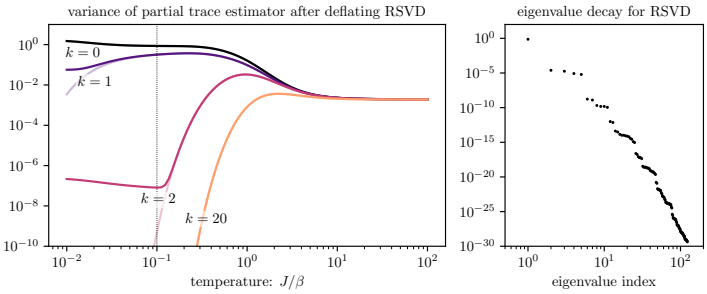

FIG. 2. *Using the same example as Figure* 1*, we plot the approximate variance of the partial trace estimator for* $\mathbf{A} = \exp(-\beta\mathbf{H}_\mathrm{t})/\operatorname{tr}(\exp(-\beta\mathbf{H}_\mathrm{t}))$ *when* $\mathbf{Q} = \operatorname{orth}(\exp(-\beta_0\mathbf{H}_\mathrm{t})\mathbf{\Omega})$ *is an approximate top subspace. We use* $J/\beta_0 = 0.1$ *(dotted vertical line) and sample* $\mathbf{\Omega} \in \mathbb{R}^{d_\mathrm{t} \times k}$ *with independent standard normal entries. The greyed out curves are those of Figure* 1 *and correspond to the optimal rank-*k *subspace. While the approximate top subspace does reduce the variance, the variance does not go to zero in the zero-temperature limit* $\beta \to \infty$*. The right plot shows the normalized singular values of* $\exp(-\beta_0\mathbf{H}_\mathrm{t})$*, which decay rapidly.*

where $\mathbf{Q}$ is obtained by applying the randomized SVD to $\exp(-\beta_0 \mathbf{H}_{\mathrm{t}})$ for some fixed value β_0. This is compared to Figure 1, where we plot the analogous quantity when the projection space is taken as the top subspace. For $\beta < \beta_0$, the variance reduction is essentially the same as if the exact top eigenspace were used. For $\beta > \beta_0$ there is still some variance reduction; it does not result in a zero-variance approximation in the zero-temperature limit $\beta \to \infty$.

A deeper exploration of the trade-offs between the cost to compute $\mathbf{Q}$, the quality of the variance reduction, and the costs of computing $f(\mathbf{H})\mathbf{Q}$ is beyond the scope of the present paper but is an important topic for future work.

5. Numerical experiments. Our experiments focus on Heisenberg spin systems in an isotropic magnetic field oriented in the positive z-direction:

$$\mathbf{H} := \sum_{i,j=1}^{N} \left[J^{\mathrm{x}}_{i,j} \boldsymbol{\sigma}^{\mathrm{x}}_i \boldsymbol{\sigma}^{\mathrm{x}}_j + J^{\mathrm{y}}_{i,j} \boldsymbol{\sigma}^{\mathrm{y}}_i \boldsymbol{\sigma}^{\mathrm{y}}_j + J^{\mathrm{z}}_{i,j} \boldsymbol{\sigma}^{\mathrm{z}}_i \boldsymbol{\sigma}^{\mathrm{z}}_j \right] + \frac{h}{2} \sum_{i=1}^{N} \boldsymbol{\sigma}^{\mathrm{z}}_i. \tag{5.1}$$

Here $\boldsymbol{\sigma}^{\mathrm{x/y/z}}_i$ is defined by

$$\boldsymbol{\sigma}^{\mathrm{x/y/z}}_i = \underbrace{\mathbf{I} \otimes \cdots \otimes \mathbf{I}}_{i-1 \text{ terms}} \otimes \boldsymbol{\sigma}^{\mathrm{x/y/z}} \otimes \underbrace{\mathbf{I} \otimes \cdots \otimes \mathbf{I}}_{N-i \text{ terms}}, \tag{5.2}$$

where $\boldsymbol{\sigma}^{\mathrm{x/y/z}}$ are the Pauli spin-$\frac{1}{2}$ matrices

$$\boldsymbol{\sigma}^{\mathrm{x}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \boldsymbol{\sigma}^{\mathrm{y}} = \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix}, \qquad \boldsymbol{\sigma}^{\mathrm{z}} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{5.3}$$

We remark that while $\boldsymbol{\sigma}^{\mathrm{y}}$ is Hermitian, $\boldsymbol{\sigma}^{\mathrm{y}}_i \boldsymbol{\sigma}^{\mathrm{y}}_j$ (and thus $\mathbf{H}$) is *real symmetric.*

5.1. Experimental setup. Our experiments are implemented in Python using double precision arithmetic. We use SciPy's sparse library to represent Hamiltonians and SciPy's `sparse.linalg.eigsh` to compute the top eigenvectors. The latter is a wrapper for ARPACK's `dsaupd`, which is an implementation of the implicitly restarted Lanczos method, and computes the eigenvectors to machine precision [36].

We set the number of Lanczos iterations to t so that products involving $\exp(-\beta\mathbf{H})$ are computed to a relative error of roughly 10^{-10}. The statistical noise from the random samples is much larger, so computing the matrix functions to any additional accuracy does not impact the results in any noticeable way. Reorthogonalization is not used, but we do explicitly orthogonalize against the deflated subspace.

Code used to generate the data plotted in the figures is available at `https://github.com/tchen-research/faster_partial_trace`.

5.2. Quantities of interest.

5.2.1. Entanglement spectrum and von Neumann entropy. The von Neumann entropy is an information theoretic measure of the entropy of a quantum system and can be viewed as a measure of how far a quantum state is from being pure. Thus, the von Neumann entropy of subsystem (s) provides information about the entanglement between subsystems (s) and (b).

The von Neumann entropy of subsystem (s) is defined by the formula

$$S = S(\beta, h) := -\operatorname{tr}\left(\boldsymbol{\rho}^* \ln(\boldsymbol{\rho}^*)\right), \tag{5.4}$$

where $\boldsymbol{\rho}^*$ is the reduced density matrix defined in (2.3).

The set of eigenvalues of $-\ln(\boldsymbol{\rho}^*)$ is sometimes referred to as the entanglement spectrum and provides a more complete picture of the entanglement of subsystems (s) and (b) than the von Neumann entropy [37]. Up to a scaling factor $1/\beta$, $-\ln(\boldsymbol{\rho}^*)$ is the same as the Hamiltonian of the mean force and is an important quantity in equilibrium thermodynamics [61].

5.2.2. Ergotropy of quantum batteries. Quantum batteries use quantum systems to store energy and offer the potential for faster charging and higher efficiency than classical batteries [8]. We consider the setup of [6] in which the battery consists of the spins in subsystem (s) and charges the spins in subsystem (b). Once the total system (battery+charger) is in thermal equilibrium, the battery is instantaneously disconnected from the charger and is therefore in state $\boldsymbol{\rho}^*$ with internal energy $\operatorname{tr}(\mathbf{H}_{\mathrm{s}}\boldsymbol{\rho}^*)$. We can extract energy from the battery by evolution with a unitary $\mathbf{U}$, which brings us to state $\mathbf{U}\boldsymbol{\rho}^*\mathbf{U}^{\mathsf{T}}$ with internal energy $\operatorname{tr}(\mathbf{H}_{\mathrm{s}}\mathbf{U}\boldsymbol{\rho}^*\mathbf{U}^{\mathsf{T}})$. The ergotropy [2, 7] $\mathcal{E} = \mathcal{E}(\beta, h)$ is defined as the total possible energy which could be extracted from the battery:

$$\mathcal{E} = \mathcal{E}(\beta, h) := \max_{\mathbf{U}^\dagger\mathbf{U}=\mathbf{I}} \left(\operatorname{tr}\left(\mathbf{H}_{\mathrm{s}}\boldsymbol{\rho}^*\right) - \operatorname{tr}\left(\mathbf{H}_{\mathrm{s}}\mathbf{U}\boldsymbol{\rho}^*\mathbf{U}^{\mathsf{T}}\right) \right). \tag{5.5}$$

The unitary $\mathbf{U}$ minimizing $\operatorname{tr}\left(\mathbf{H}_{\mathrm{s}}\mathbf{U}\boldsymbol{\rho}^*\mathbf{U}^{\mathsf{T}}\right)$ can be obtained explicitly [2]. Specifically, if $\mathbf{H}_{\mathrm{s}}$ and $\boldsymbol{\rho}^*$ are diagonalized (with eigenvalues in nonincreasing order) as $\mathbf{H}_{\mathrm{s}} = \mathbf{Q}_{\mathrm{s}}\boldsymbol{\Lambda}_{\mathrm{s}}\mathbf{Q}_{\mathrm{s}}^{\mathsf{T}}$ and $\boldsymbol{\rho}^* = \mathbf{Q}_\rho\boldsymbol{\Lambda}_\rho\mathbf{Q}_\rho^{\mathsf{T}}$, then

$$\mathbf{U} = \mathbf{Q}_{\mathrm{s}}\mathbf{P}\mathbf{Q}_\rho^{\mathsf{T}}, \tag{5.6}$$

where $\mathbf{P}$ is the reversal permutation matrix (identity with columns reversed).

5.3. Kagome-strip chain. In this experiment, we consider Kagome-strip chain systems [4, 68, 42] as show in Figure 3. We take subsystem (s) to be the five spins indicated in Figure 3.

We choose several values of J_2 and fix $J_1 = J_3 = J$. For each choice of J_2, we consider a range of h and β for each system. To determine the values of h at which to run our algorithm, we use bisection on the von Neumann entropy of the ground state to determine intervals where the von Neumann entropy appears constant. We then run our algorithm at values of h corresponding to Chebyshev nodes shifted and scaled to each interval. Throughout, we use $k = 25$ and $m = 5$. The results are illustrated in Figure 4.

We also consider the entanglement spectrum at a fixed value of β. This is illustrated in Figure 5.

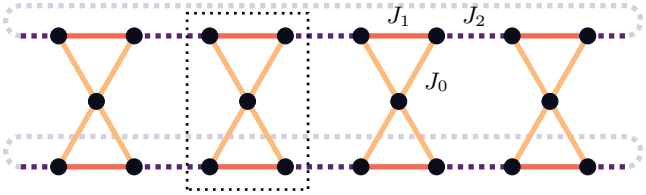

FIG. 3. *Kagome-strip chain with $N = 20$ sites and periodic boundary conditions. Subsystem (s) is encircled.*

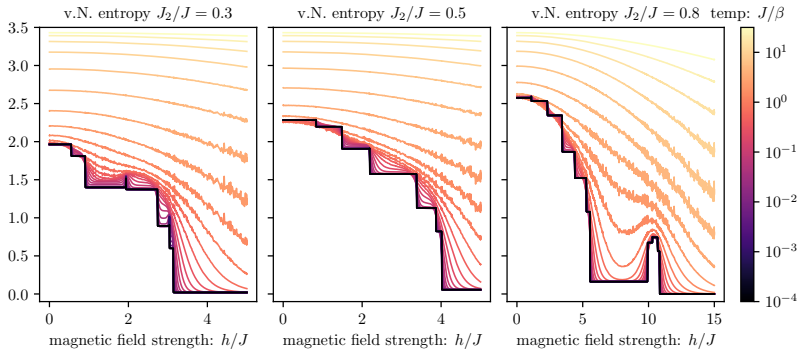

FIG. 4. *Von Neumann entropy for Kagome-strip chain at varying values of J_2, h, and β.*

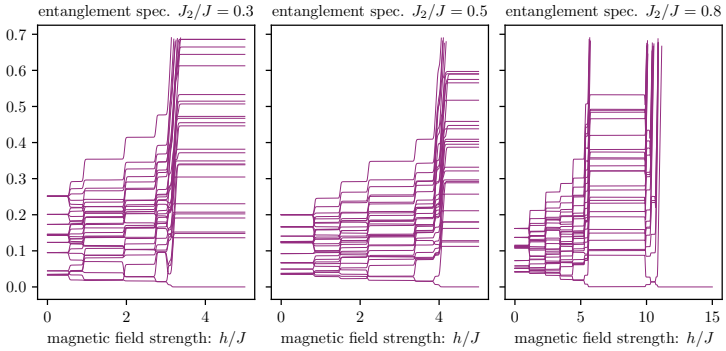

FIG. 5. *Entanglement spectrum for Kagome-strip chain at varying values of J_2 and h at fixed temperature $J/\beta = 2 \times 10^{-2}$.*

5.4. Long range spin chain. We now consider the XX spin chain with long-range power-law interactions

$$J^{\mathrm{x}}_{i,j} = J^{\mathrm{y}}_{i,j} = |i-j|^{-\alpha}, \qquad J^{\mathrm{z}}_{i,j} = 0. \tag{5.7}$$

Here we take subsystem (s) to be the first two spins and subsystem (b) to be the remaining spins. We remark that in the case $\alpha = \infty$, this system is exactly solvable, and we use this to verify the accuracy of our algorithm in subsection 5.5.

Within this framework, we set $N = 16$ and vary the parameters α, h, and β. We use the same bisection-based approach to determine suitable values of h at which to run our algorithm. For each value of α and h, we run our algorithm with $k = 25$ and $m = 5$ and again use enough Lanczos iterations to accurately compute the matrix functions for each value of β.

In Figure 6 we visualize the von Neumann entropy of subsystem (s) for several values of α. We observe that at zero temperature, the von Neumann entropy appears piecewise constant. In the solvable model, the steps in the zero-temperature von Neumann entropy correspond to values of h/J for which a fermionic eigenmode vanishes [9, eq. (69)]. The inset panel of Figure 6 shows a plot zoomed in to the right edge of the top "plateau" and illustrates that the von Neumann entropy of the ground state changes continuously in this region.

In Figure 7 we show the ergotropy of subsystem (s) for several values of α. We use the same values of h as used for the von Neumann entropy. While the sharp

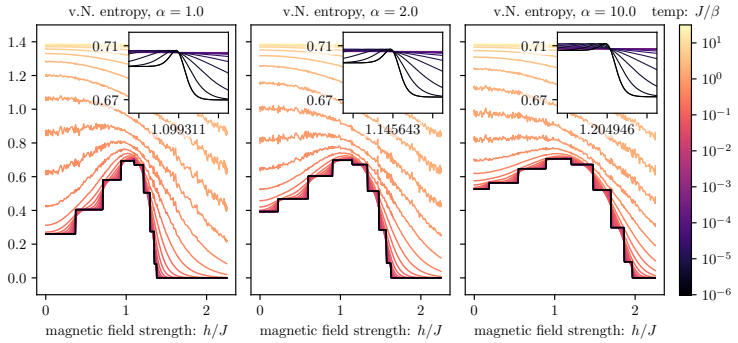

FIG. 6. *Von Neumann entropy for long-range spin chain with* $N = 16$ *and the system taken as the first two spins at varying values of* α*,* h*, and* β*. Inset figure shows the transition to the right of the top "plateau."*

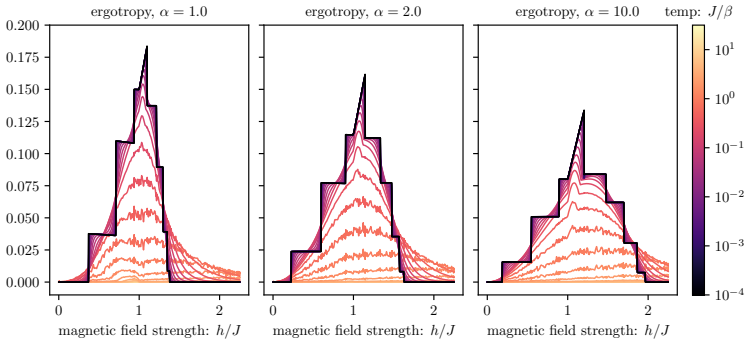

FIG. 7. *Ergotropy for long-range spin chain with* $N = 16$ *and the system taken as the first two spins at varying values of* α*,* h*, and* β*. Observe the nonpiecewise continuous behavior to the right of* 1*, despite the von Neumann entropy appearing constant in this region.*

increases appear to happen in the same places as in the von Neumann entropy, the regions between jumps appear linear rather than constant. In addition, there is an apparent discontinuity in the derivative of the $\beta = \infty$ curve at $h/J = 1$, which does not appear in the von Neumann entropy.

5.5. Validation on the solvable XX spin chain. In the special case where

$$J^{\mathrm{x}}_{i,j} = J^{\mathrm{y}}_{i,j} = \begin{cases} J & |i-j| = 1 \\ 0 & |i-j| \neq 1 \end{cases}, \qquad J^{\mathrm{z}}_{i,j} = 0, \tag{5.8}$$

the system (5.1) is exactly solvable via the "Bethe ansatz" [33]. That is, it can be diagonalized analytically. In addition, expressions for the partial trace of the first two spins have been obtained [9]; see also [10, Appendix C]. This allows us to test our algorithm against a known solution for problem sizes where exact diagonalization is intractable.

5.5.1. Variance study. We begin by studying the variability of the output of our algorithm. We use $N = 18$ and $h/J = 0.3$ and run the algorithm at varying values of k and m and compute the eigenvalues of ρ^* for a range of β. For each choice of k and m, we repeatedly and independently run our algorithm 10 times. This gives some indication of the variance in the algorithm. In all cases, we use enough Lanczos

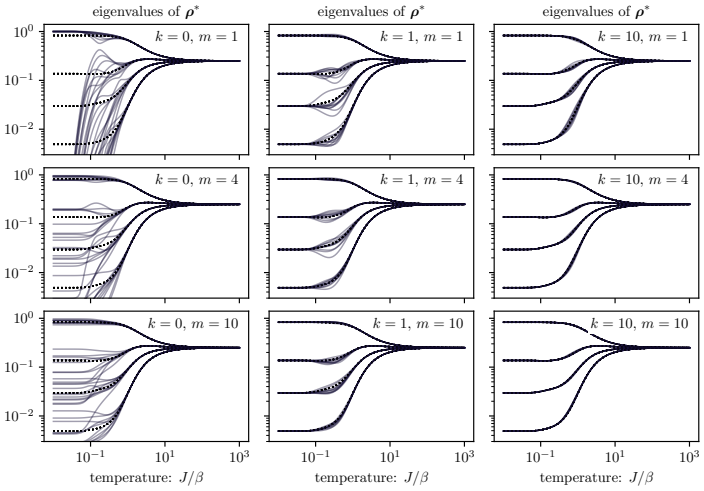

FIG. 8. *Comparison of $k = 0$ (equivalent to [10]), $k = 1$, and $k = 10$ for several values of m when our algorithm is run on the solvable model with $N = 16$ and the system taken as the first two spins. Lines correspond to repeated runs of our algorithm. Takeaway: As k and m increase, the variance decreases. However, while the variance decreases linearly with m, it may decrease more quickly with k.*

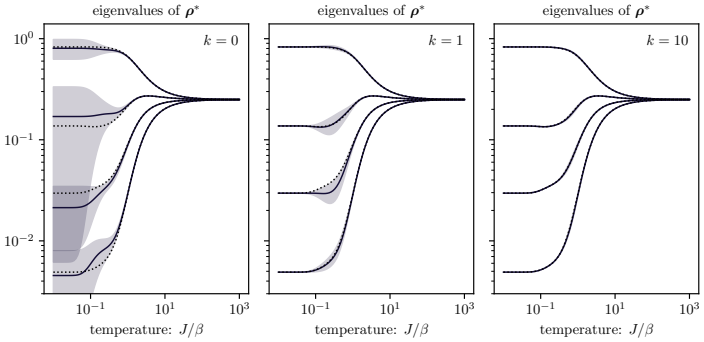

FIG. 9. *Leave-one-out estimator for standard error obtained from a single run of the algorithm (same setup as Figure 8) with $m = 10$. Takeaway: Using just the information from a single run of the algorithm, we can get reasonable estimates for the variability of the output.*

iterations to accurately compute the matrix functions for each value of β. Figure 8 shows the results of this experiment. Note that the $k = 0$ plots in Figure 8 correspond to the algorithm from [10], which does not use deflation; see [10, Figure 1].

5.5.2. Jackknife variance estimates. Generating plots like Figure 8 is not practical, as it requires multiple runs of the algorithm for each parameter setting. However, since our estimator (3.7) involves averaging m i.i.d. samples of an unbiased random matrix, we can use a jackknife (leave-one-out) estimator for the variance. We visualize the error estimated by the jackknife method for a single run of the algorithm in Figure 9. Here, the estimated standard error seems to align well with the true error of the algorithm.

5.5.3. Von Neumann entropy. In this experiment, we set $N = 16$ and $N_s = 2$ and vary the magnetic field strength h/J. Figure 10 shows the exact von Neumann

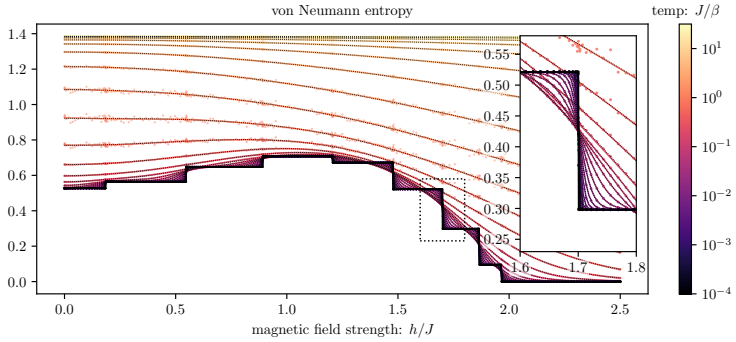

FIG. 10. *Approximation of von Neumann entropy on the solvable model with $N = 16$ and the system taken as the first two spins. Smoothing is used to reduce noise at moderate temperatures (raw data shown as dots). We use $k = 25$ eigenvectors for deflation and $m = 5$ copies of the stochastic trace estimator. Compared to* [10, *Figure* 3] *($k = 0, m = 400$), our approach not only is computationally cheaper but also results in a much more accurate approximation. Takeaway: Deflation allows the low-temperature parts of the curve to be obtained very accurately. While the algorithm's output has some noise at intermediate temperature, this is easily overcome by smoothing, which naturally averages the randomness of nearby points.*

entropy as a function of h/J for a range of β as well as the values computed by our algorithm with the parameters $k = 25$ and $m = 5$. In the right panel of Figure 10, we show a cropped version of the left panel. In this plot, the low-variance behavior of the algorithm at low but nonzero temperatures is clearly visible. This is in sharp contrast to [10], in which high variance is observed at low temperatures, even with $m = 400$.

For high temperatures $\beta/J < 5$, we simply fit a degree 10 polynomial with least squares. While the raw data deviates considerably from the true von Neumann entropy, the least squares fit seems extremely accurate. This suggests that the bias of the algorithm's output is fairly small. For low temperature $\beta/J \in (5, 500)$, we use cubic splines. Here the algorithm's output has little noise, and the smoothing is mainly to interpolate the data to values of h/J, which we did not run the algorithm on. Finally, for very low temperature $\beta/J > 500$, we do not do any smoothing.

6. Conclusion and outlook. We have incorporated projection as a means of variance reduction in typicality estimators for the partial traces. Since the partial trace does not satisfy the cyclic property, there is a potential for cancellation if the partial trace estimator cannot be computed exactly. To avoid this, we use deflation (projection of top eigenspace) in combination with the block-Lanczos algorithm and explicitly orthogonalize against the projected subspace. Our approach significantly reduces the runtime required to obtain accurate estimates, often by several orders of magnitude.

In the future, we would like to take further advantage of the structure of the systems we are dealing with. For example, in many situations, one would like to compute a quantity of interest over a range of parameter values (e.g., magnetic field strength, coupling strength, etc.). Presently, we apply our algorithm independently at each parameter setting. This leaves open the potential for a faster algorithm which can take advantage of the fact that many of the quantities may not change significantly.

REFERENCES

[1] R. Alicki and R. Kosloff, *Introduction to quantum thermodynamics: History and prospects*, in Thermodynamics in the Quantum Regime, Fundamental Theories of Physics 195, F. Binder et al., eds., Springer, 2018, pp. 1–33, https://doi.org/10.1007/978-3-319-99046-0_1.

[2] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen, *Maximal work extraction from finite quantum systems*, Europhys. Lett. (EPL), 67 (2004), pp. 565–571, https://doi.org/10.1209/epl/i2004-10101-2.

[3] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. ACM, 58 (2011), 8, https://doi.org/10.1145/1944345.1944349.

[4] P. Azaria, C. Hooley, P. Lecheminant, C. Lhuillier, and A. M. Tsvelik, *Kagomé lattice antiferromagnet stripped to its basics*, Phys. Rev. Lett., 81 (1998), pp. 1694–1697, https://doi.org/10.1103/physrevlett.81.1694.

[5] R. Baer, D. Neuhauser, and E. Rabani, *Stochastic vector techniques in ground-state electronic structure*, Annu. Rev. Phys. Chem., 73 (2022), pp. 255–272, https://doi.org/10.1146/annurev-physchem-090519-045916.

[6] F. Barra, K. V. Hovhannisyan, and A. Imparato, *Quantum batteries at the verge of a phase transition*, New J. Phys., 24 (2022), 015003, https://doi.org/10.1088/1367-2630/ac43ed.

[7] F. Campaioli, S. Gherardini, J. Q. Quach, M. Polini, and G. M. Andolina, *Colloquium: Quantum batteries*, Rev. Mod. Phys., 96 (2024), 03100.

[8] F. Campaioli, F. A. Pollock, and S. Vinjanampathy, *Quantum batteries*, in Thermodynamics in the Quantum Regime, Fundamental Theories of Physics 195, F. Binder et al., eds., Springer, 2018, pp. 207–225, https://doi.org/10.1007/978-3-319-99046-0_8.

[9] M. Campisi, D. Zueco, and P. Talkner, *Thermodynamic anomalies in open quantum systems: Strong coupling effects in the isotropic XY model*, Chem. Phys., 375 (2010), pp. 187–194, https://doi.org/10.1016/j.chemphys.2010.04.026.

[10] T. Chen and Y.-C. Cheng, *Numerical computation of the equilibrium-reduced density matrix for strongly coupled open quantum systems*, J. Chem. Phys., 157 (2022), 064106, https://doi.org/10.1063/5.0099761.

[11] T. Chen and E. Hallman, *Krylov-aware stochastic trace estimation*, SIAM J. Matrix Anal. Appl., 44 (2023), pp. 1218–1244, https://doi.org/10.1137/22m1494257, https://arxiv.org/abs/2205.01736.

[12] T. Chen, T. Trogdon, and S. Ubaru, *Randomized Matrix-Free Quadrature for Spectrum and Spectral Sum Approximation*, preprint, https://arxiv.org/abs/2204.01941, 2022.

[13] Y.-F. Chiu, A. Strathearn, and J. Keeling, *Numerical evaluation and robustness of the quantum mean-force Gibbs state*, Phys. Rev. A, 106 (2022), 012204, https://doi.org/10.1103/physreva.106.012204.

[14] A. Cortinovis and D. Kressner, *On randomized trace estimates for indefinite matrices with an application to determinants*, Found. Comput. Math., 22 (2021), pp. 875–903, https://doi.org/10.1007/s10208-021-09525-9.

[15] P. Czarnik, M. M. Rams, and J. Dziarmaga, *Variational tensor network renormalization in imaginary time: Benchmark results in the Hubbard model at finite temperature*, Phys. Rev. B, 94 (2016), 235142, https://doi.org/10.1103/physrevb.94.235142.

[16] P. E. Dargel, A. Wöllert, A. Honecker, I. P. McCulloch, U. Schollwöck, and T. Pruschke, *Lanczos algorithm with matrix product states for dynamical correlation functions*, Phys. Rev. B, 85 (2012), 205119, https://doi.org/10.1103/physrevb.85.205119.

[17] V. Druskin and L. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, USSR Comput. Math. Math. Phys., 29 (1989), pp. 112–121, https://doi.org/10.1016/s0041-5553(89)80020-5.

[18] E. N. Epperly, J. A. Tropp, and R. J. Webber, *Xtrace: Making the most of every sample in stochastic trace estimation*, SIAM J. Matrix Anal. Appl., 45 (2024), pp. 1–23, https://doi.org/10.1137/23m1548323.

[19] A. Frommer, K. Lund, and D. B. Szyld, *Block Krylov subspace methods for functions of matrices II: Modified block FOM*, SIAM J. Matrix Anal. Appl., 41 (2020), pp. 804–837, https://doi.org/10.1137/19m1255847.

[20] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 1236–1264, https://doi.org/10.1137/0913071.

[21] A. S. Gambhir, A. Stathopoulos, and K. Orginos, *Deflation as a method of variance reduction for estimating the trace of a matrix inverse*, SIAM J. Sci. Comput., 39 (2017), pp. A532–A558, https://doi.org/10.1137/16m1066361.

[22] J. Gemmer, M. Michel, and G. Mahler, *Quantum Thermodynamics: Emergence of Thermodynamic Behavior within Composite Quantum Systems*, 2nd ed., Lecture Notes in Phys. 784, Springer, 2009.

[23] D. Girard, *Un algorithme rapide pour le calcul de la trace de l'inverse d'une grande matrice*, Rapport de recherche RR 665-M, TIM3, Informatique et Mathématiques Appliquées de Grenoble, 1987.

[24] C. Gogolin, *Pure State Quantum Statistical Mechanics*, Master's thesis, Julius-Maximilians-Universität Würzburg, 2010, https://arxiv.org/abs/1003.5058.

[25] S. Goldstein, J. L. Lebowitz, C. Mastrodonato, R. Tumulka, and N. Zanghì, *Normal typicality and von Neumann's quantum ergodic theorem*, Proc. Roy. Soc. A Math. Phys. Eng. Sci., 466 (2010), pp. 3203–3224, https://doi.org/10.1098/rspa.2009.0635.

[26] G. Golub and G. Meurant, *Matrices, Moments and Quadrature with Applications*, Princeton Ser. Appl. Math., Princeton University Press, 2009.

[27] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288, https://doi.org/10.1137/090771806.

[28] I. Han, D. Malioutov, H. Avron, and J. Shin, *Approximating spectral sums of large-scale matrices using stochastic Chebyshev approximations*, SIAM J. Sci. Comput., 39 (2017), pp. A1558–A1585, https://doi.org/10.1137/16m1078148.

[29] M. Hutchinson, *A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines*, Commun. Statist. Simul. Comput., 18 (1989), pp. 1059–1076, https://doi.org/10.1080/03610918908812806.

[30] G.-L. Ingold, P. Hänggi, and P. Talkner, *Specific heat anomalies of open quantum systems*, Phys. Rev. E, 79 (2009), 061105, https://doi.org/10.1103/physreve.79.061105.

[31] J. Jaklič and P. Prelovšek, *Lanczos method for the calculation of finite-temperature quantities in correlated systems*, Phys. Rev. B, 49 (1994), pp. 5065–5068, https://doi.org/10.1103/physrevb.49.5065.

[32] F. Jin, D. Willsch, M. Willsch, H. Lagemann, K. Michielsen, and H. De Raedt, *Random state technology*, J. Phys. Soc. Jpn., 90 (2021), 012001, https://doi.org/10.7566/jpsj.90.012001.

[33] M. Karabach, G. Müller, H. Gould, and J. Tobochnik, *Introduction to the Bethe ansatz* , Comput. Phys., 11 (1997), pp. 36–43, https://doi.org/10.1063/1.4822511.

[34] L. A. Knizhnerman, *The simple Lanczos procedure: Estimates of the error of the Gauss quadrature formula and their applications*, Comput. Math. Math. Phys., 36 (1996), pp. 1481–1492.

[35] A. Kshetrimayum, M. Rizzi, J. Eisert, and R. Orús, *Tensor network annealing algorithm for two-dimensional thermal states*, Phys. Rev. Lett., 122 (2019), 070502, https://doi.org/10.1103/physrevlett.122.070502.

[36] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environments Tools 6, SIAM, 1998, https://doi.org/10.1137/1.9780898719628.

[37] H. Li and F. D. M. Haldane, *Entanglement spectrum as a generalization of entanglement entropy: Identification of topological order in non-Abelian fractional quantum Hall effect states*, Phys. Rev. Lett., 101 (2008), 010504, https://doi.org/10.1103/physrevlett.101.010504.

[38] L. Lin, *Randomized estimation of spectral densities of large matrices made accurate*, Numer. Math., 136 (2017), pp. 183–213, https://doi.org/10.1007/s00211-016-0837-7.

[39] P.-G. Martinsson and J. A. Tropp, *Randomized numerical linear algebra: Foundations and algorithms*, Acta Numer., 29 (2020), pp. 403–572, https://doi.org/10.1017/s0962492920000021.

[40] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff, *Hutch++: Optimal stochastic trace estimation*, in Proceedings of the 2021 Symposium on Simplicity in Algorithms (SOSA), SIAM, 2021, pp. 142–155, https://doi.org/10.1137/1.9781611976496.16.

[41] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49, https://doi.org/10.1137/s00361445024180.

[42] K. Morita, T. Sugimoto, S. Sota, and T. Tohyama, *Magnetization plateaus in the spin-antiferromagnetic Heisenberg model on a kagome-strip chain*, Phys. Rev. B, 97 (2018), 014412, https://doi.org/10.1103/physrevb.97.014412.

[43] K. Morita and T. Tohyama, *Finite-temperature properties of the Kitaev-Heisenberg models on kagome and triangular lattices studied by improved finite-temperature Lanczos methods*, Phys. Rev. Res., 2 (2020), 013205, https://doi.org/10.1103/physrevresearch.2.013205.

[44] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 10th anniversary ed., Cambridge University Press, 2010.
[45] R. Orús, *A practical introduction to tensor networks: Matrix product states and projected entangled pair states*, Ann. Phys., 349 (2014), pp. 117–158, https://doi.org/10.1016/j.aop.2014.06.013.
[46] R. Orús, *Tensor networks for complex quantum systems*, Nature Rev. Phys., 1 (2019), pp. 538–550, https://doi.org/10.1038/s42254-019-0086-7.
[47] C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, IMA J. Appl. Math., 18 (1976), pp. 341–349, https://doi.org/10.1093/imamat/18.3.341.
[48] C. C. Paige, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258, https://doi.org/10.1016/0024-3795(80)90167-6.
[49] D. Persson, A. Cortinovis, and D. Kressner, *Improved variants of the Hutch++ algorithm for trace estimation*, SIAM J. Matrix Anal. Appl., 43 (2022), pp. 1162–1185, https://doi.org/10.1137/21m1447623.
[50] D. Persson and D. Kressner, *Randomized low-rank approximation of monotone matrix functions*, SIAM J. Matrix Anal. Appl., 44 (2023), pp. 894–918, https://doi.org/10.1137/22m1523923.
[51] H. N. Phien, I. P. McCulloch, and G. Vidal, *Fast convergence of imaginary time evolution tensor network algorithms by recycling the environment*, Phys. Rev. B, 91 (2015), 115137, https://doi.org/10.1103/physrevb.91.115137.
[52] S. Popescu, A. J. Short, and A. Winter, *Entanglement and the foundations of statistical mechanics*, Nat. Phys., 2 (2006), pp. 754–758, https://doi.org/10.1038/nphys444.
[53] P. Reimann, *Typicality for generalized microcanonical ensembles*, Phys. Rev. Lett., 99 (2007), 160404, https://doi.org/10.1103/physrevlett.99.160404.
[54] F. Roosta-Khorasani and U. Ascher, *Improved bounds on sample size for implicit matrix trace estimators*, Found. Comput. Math., 15 (2015), pp. 1187–1212, https://doi.org/10.1007/s10208-014-9220-1.
[55] A. K. Saibaba, A. Alexanderian, and I. C. Ipsen, *Randomized matrix-free trace and log-determinant estimators*, Numer. Math., 137 (2017), pp. 353–395.
[56] H. Schlüter, F. Gayk, H.-J. Schmidt, A. Honecker, and J. Schnack, *Accuracy of the typicality approach using Chebyshev polynomials*, Z. Naturforschung A, 76 (2021), pp. 823–834, https://doi.org/10.1515/zna-2021-0116.
[57] J. Schnack, J. Richter, and R. Steinigeweg, *Accuracy of the finite-temperature Lanczos method compared to simple typicality-based estimates*, Phys. Rev. Res., 2 (2020), 013186, https://doi.org/10.1103/physrevresearch.2.013186.
[58] R. Schnalle and J. Schnack, *Calculating the energy spectra of magnetic molecules: Application of real- and spin-space symmetries*, Internat. Rev. Phys. Chem., 29 (2010), pp. 403–452, https://doi.org/10.1080/0144235x.2010.485755.
[59] E. Schrödinger, *Energieaustausch nach der Wellenmechanik*, Ann. Phys., 388 (1927), pp. 956–968, https://doi.org/10.1002/andp.19273881504.
[60] J. Skilling, *The eigenvalues of mega-dimensional matrices*, in Maximum Entropy and Bayesian Methods, Fundamental Theories of Physics 36, J. Skilling, ed., Springer, 1989, pp. 455–466, https://doi.org/10.1007/978-94-015-7860-8_48.
[61] P. Talkner and P. Hänggi, *Colloquium: Statistical mechanics and thermodynamics at strong coupling: Quantum and classical*, Rev. Modern Phys., 92 (2020), 041002, https://doi.org/10.1103/revmodphys.92.041002.
[62] J. A. Tropp and R. J. Webber, *Randomized Algorithms for Low-Rank Matrix Approximation: Design, Analysis, and Applications*, preprint, https://arxiv.org/abs/2306.12418, 2023.
[63] S. Ubaru, J. Chen, and Y. Saad, *Fast estimation of* $\mathrm{tr}(f(A))$ *via stochastic Lanczos quadrature*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1075–1099, https://doi.org/10.1137/16m1104974.
[64] S. Vinjanampathy and J. Anders, *Quantum thermodynamics*, Contemp. Phys., 57 (2016), pp. 545–579, https://doi.org/10.1080/00107514.2016.1201896.
[65] J. von Neumann, *Beweis des Ergodensatzes und des H-Theorems in der neuen Mechanik*, Z. Phys., 57 (1929), pp. 30–70, https://doi.org/10.1007/bf01339852, English translation available at https://arxiv.org/abs/1003.2133.
[66] H. Wang, S. Ashhab, and F. Nori, *Quantum algorithm for simulating the dynamics of an open quantum system*, Phys. Rev. A, 83 (2011), 062317, https://doi.org/10.1103/physreva.83.062317.
[67] A. Weisse, G. Wellein, A. Alvermann, and H. Fehske, *The kernel polynomial method*, Rev. Modern Phys., 78 (2006), pp. 275–306, https://doi.org/10.1103/revmodphys.78.275.

[68] S. R. White and R. R. P. Singh, *Comment on "Kagomé lattice antiferromagnet stripped to its basics,"* Phys. Rev. Lett., 85 (2000), p. 3330, https://doi.org/10.1103/physrevlett.85.3330.
[69] L. Wu, J. Laeuchli, V. Kalantzis, A. Stathopoulos, and E. Gallopoulos, *Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse*, J. Comput. Phys., 326 (2016), pp. 828–844.