Randomized trace estimation

Tyler Chen

https://chen.pw/slides.pdf

Trace estimation and the matvec query model

Setting: Given a $n \times n$ matrix $\mathbf{A}$

Goal: Estimate $\mathrm{tr}(\mathbf{A}) = A_{1,1} + A_{2,2} + \cdots + A_{n,n}$

Constraint: Can only access $\mathbf{A}$ by matrix-vector product queries

Why a matrix vector product query model?

Pros:

- In many linear-algebra algorithms, matrix-vector products dominate the cost of computation
- We can hope to prove query complexity low-bounds to understand the hardness of linear algebra problems

Cons:

- Ignores arithmetic costs
- Matvecs may not be true core primitive

The basic algorithm

We can get the exact trace with n matvec queries.

The basic algorithm

We can get the exact trace with n matvec queries.

1. Multiply $\mathbf{A}$ with $\mathbf{e}_i = [0, \dots, 0, 1, 0, \dots, 0]$ to read off the i-th column.
2. extract the i-th entry of $\mathbf{A}\mathbf{e}_i$, and add to running sum
3. Repeat.

The basic algorithm

We can get the exact trace with n matvec queries.

1. Multiply $\mathbf{A}$ with $\mathbf{e}_i = [0, \dots, 0, 1, 0, \dots, 0]$ to read off the i-th column.
2. extract the i-th entry of $\mathbf{A}\mathbf{e}_i$, and add to running sum
3. Repeat.

Can we get the trace approximately with far fewer queries?

A randomized estimator

Let $\mathbf{v}$ be a random vector where $v_i \sim \text{unif}(-1, +1)$ iid. Consider the estimator,

$$\mathbf{v}^\top \mathbf{A} \mathbf{v} = \sum_{i,j} v_i v_j A_{i,j}.$$

A randomized estimator

Let $\mathbf{v}$ be a random vector where $v_i \sim \text{unif}(-1, +1)$ iid. Consider the estimator,

$$\mathbf{v}^\top \mathbf{A} \mathbf{v} = \sum_{i,j} v_i v_j A_{i,j}.$$

What is the expectation?

$$\mathbb{E}[\mathbf{v}^\top \mathbf{A} \mathbf{v}] = \sum_{i} \mathbb{E}[v_i^2] A_{i,i} + \sum_{i \neq j} \mathbb{E}[v_i v_j] A_{i,j}$$

A randomized estimator

Let $\mathbf{v}$ be a random vector where $v_i \sim \text{unif}(-1, +1)$ iid. Consider the estimator,

$$\mathbf{v}^\top \mathbf{A} \mathbf{v} = \sum_{i,j} v_i v_j A_{i,j}.$$

What is the expectation?

$$\mathbb{E}[\mathbf{v}^\top \mathbf{A} \mathbf{v}] = \sum_{i} \mathbb{E}[v_i^2] A_{i,i} + \sum_{i \neq j} \mathbb{E}[v_i v_j] A_{i,j} = \sum_{i} 1 A_{i,i} + \sum_{i \neq j} 0 A_{i,j} = \text{tr}(\mathbf{A}).$$

A randomized estimator

Let $\mathbf{v}$ be a random vector where $v_i \sim \mathrm{unif}(-1, +1)$ iid. Consider the estimator,

$$\mathbf{v}^\top \mathbf{A} \mathbf{v} = \sum_{i,j} v_i v_j A_{i,j}.$$

What is the expectation?

$$\mathbb{E}[\mathbf{v}^\top \mathbf{A} \mathbf{v}] = \sum_{i} \mathbb{E}[v_i^2] A_{i,i} + \sum_{i \neq j} \mathbb{E}[v_i v_j] A_{i,j} = \sum_{i} 1 A_{i,i} + \sum_{i \neq j} 0 A_{i,j} = \mathrm{tr}(\mathbf{A}).$$

What about the variance?

$$\mathbb{V}[\mathbf{v}^\top \mathbf{A} \mathbf{v}] = \sum_{i,j,k,\ell} \mathbb{E}[\mathrm{blah}(i, j, k, \ell)] = 2\|\mathbf{A} - \mathrm{diag}(\mathbf{A})\|_\mathsf{F}^2 = 2 \sum_{i \neq j} A_{i,j}^2.$$

Other distributions?

Some simple distributions:

- iid signs: $2\|\mathbf{A} - \operatorname{diag}(\mathbf{A})\|_\mathsf{F}^2$
  - For vectors with real iid entries, this is the minimum variance distribution[1]
- iid Gaussians: $2\|\mathbf{A}\|_\mathsf{F}^2$
- real sphere: $\frac{2n}{n+2}\left(\|\mathbf{A}\|_\mathsf{F}^2 - \frac{1}{n}|\operatorname{tr}(\mathbf{A})|^2\right)$
  - This is the minimax distribution over all $n \times n$ (symmetric) matrices

Great overview: Epperly 2023

[1]Hutchinson 1989.

[2]Wimmer, Wu, and Zhang 2014.

Other distributions?

Some simple distributions:

- iid signs: $2\|\mathbf{A} - \operatorname{diag}(\mathbf{A})\|_\mathsf{F}^2$
  - For vectors with real iid entries, this is the minimum variance distribution[1]
- iid Gaussians: $2\|\mathbf{A}\|_\mathsf{F}^2$
- real sphere: $\frac{2n}{n+2}\left(\|\mathbf{A}\|_\mathsf{F}^2 - \frac{1}{n}|\operatorname{tr}(\mathbf{A})|^2\right)$
  - This is the minimax distribution over all $n \times n$ (symmetric) matrices

Great overview: Epperly 2023

By averaging m iid copies, we can get accuracy $O(\|\mathbf{A}\|_\mathsf{F}/\sqrt{m})$. Note that this is independent of n (if the norm is constant)!

Lower bounds show that even m adaptive quadratic form queries can't do better than $O(1/\sqrt{m})$ queries[2]

[1]Hutchinson 1989.

[2]Wimmer, Wu, and Zhang 2014.

Variance reduction

Idea: Given a matrix $\tilde{\mathbf{A}}$ of known trace, decompose

$$\operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\tilde{\mathbf{A}}) + \operatorname{tr}(\mathbf{A} - \tilde{\mathbf{A}}).$$

If $\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathsf{F}}^2 \ll \|\mathbf{A}\|_{\mathsf{F}}^2$, the variance can be improved greatly by applying the random estimator to the remainder $\mathbf{A} - \tilde{\mathbf{A}}$.

Variance reduction

Idea: Given a matrix $\tilde{\mathbf{A}}$ of known trace, decompose

$$\operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\tilde{\mathbf{A}}) + \operatorname{tr}(\mathbf{A} - \tilde{\mathbf{A}}).$$

If $\|\mathbf{A} - \tilde{\mathbf{A}}\|_F^2 \ll \|\mathbf{A}\|_F^2$, the variance can be improved greatly by applying the random estimator to the remainder $\mathbf{A} - \tilde{\mathbf{A}}$.

Question: How do we determine $\tilde{\mathbf{A}}$ (in the matvec query model)?

- Sketching!

Hutch++ (Meyer, Musco, Musco, and Woodruff 2021)

We will construct $\mathbf{Q}$ approximating the top subspace of $\mathbf{A}$ and set $\tilde{\mathbf{A}} = \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}$. We can get a variance reduced estimator:

$$\text{Hutch++} = \operatorname{tr}(\tilde{\mathbf{A}}) + \frac{1}{m}\sum_{i=1}^{m} \mathbf{v}_i^\mathsf{T}(\mathbf{A} - \tilde{\mathbf{A}})\mathbf{v}_i$$

1. Form $\mathbf{Y} = \mathbf{A}\mathbf{G}$, $\mathbf{G}$ $n \times m$ Gaussian — m matvecs
2. Form $\mathbf{Q} = \operatorname{orth}(\mathbf{Y})$
3. Form $\tilde{\mathbf{A}} = \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}$ — m matvecs
4. Compute $\operatorname{tr}(\tilde{\mathbf{A}}) = \operatorname{tr}(\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q})$
5. Approximate $\operatorname{tr}(\mathbf{A} - \tilde{\mathbf{A}}) = \operatorname{tr}((\mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathsf{T})\mathbf{A})$ by $\frac{1}{m}\sum_{i=1}^{m} \mathbf{v}_i^\mathsf{T}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathsf{T})\mathbf{A}\mathbf{v}_i$ — m matvecs

Hutch++: Analysis

The entire variance of the estimator comes from step 5. Suppose $\mathbf{A}$ is positive definite. Then:

Fact: $\|\mathbf{A} - [\mathbf{A}]_k\|_\mathsf{F}^2 \leq \frac{1}{4k}\operatorname{tr}(\mathbf{A})^2$, where $[\mathbf{A}]_k$ is the optimal rank-k approximation

Fact: $\mathbb{E}\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}\|_\mathsf{F}^2 \lesssim 2\|\mathbf{A} - [\mathbf{A}]_{m/2}\|_\mathsf{F}^2$

[3] Meyer, Musco, Musco, and Woodruff 2021.

Hutch++: Analysis

The entire variance of the estimator comes from step 5. Suppose $\mathbf{A}$ is positive definite. Then:

Fact: $\|\mathbf{A} - [\mathbf{A}]_k\|_\mathsf{F}^2 \leq \frac{1}{4k}\operatorname{tr}(\mathbf{A})^2$, where $[\mathbf{A}]_k$ is the optimal rank-k approximation

Fact: $\mathbb{E}\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|_\mathsf{F}^2 \lesssim 2\|\mathbf{A} - [\mathbf{A}]_{m/2}\|_\mathsf{F}^2$

Together:

$$\mathbb{V}[\text{Hutch++}] \approx \frac{2}{m}\mathbb{E}\|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{A}\|_\mathsf{F}^2 \leq \frac{1}{m^2}\operatorname{tr}(\mathbf{A}).$$

Using $O(m)$ vectors, we get a $O(1/m)$ relative approximation. This is a quadratic improvement and nearly optimal[3] in matvec query models!

[3] Meyer, Musco, Musco, and Woodruff 2021.

Related ideas

Similar deflation ideas suggested in physics[4] and numerical analysis[5]

Subsequent improvmenets:

- Persson, Cortinovis, and Kressner 2022: automatic allocation of matvecs to low-rank approximation and stochastic trace estimation
- Epperly, Tropp, and Webber 2023: exchangability princple and cheap downdating– use all matvecs for both

[4]Girard 1987; Lin 2016; Morita and Tohyama 2020.

[5]Wu, Laeuchli, Kalantzis, Stathopoulos, and Gallopoulos 2016; Gambhir, Stathopoulos, and Orginos 2017.

Structured trace estimation

What if $\mathbf{A}$ has additional structure?

- If $\mathbf{A}$ is nearly diagonal, using $[\pm 1, \pm 1, \pm 1, \ldots]$ works really well.
- If $\mathbf{A}$ is nearly tridiagonal, we can use $[\pm 1, 0, \pm 1, 0, \ldots]$ and $[0, \pm 1, 0, \pm 1, \ldots]$

More generally, can try to recover $\mathbf{A}$ from matvec queries.[6]

[6] Halikias and Townsend 2022.

What about matrix functions?

An $n \times n$ symmetric matrix $\mathbf{H}$ has real eigenvalues and orthonormal eigenvectors:

$$\mathbf{H} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top.$$

The matrix function $f(\mathbf{H})$ is defined as

$$f(\mathbf{H}) := \sum_{i=1}^{n} f(\lambda_i) \mathbf{u}_i \mathbf{u}_i^\top.$$

In many applications of trace estimation, $\mathbf{A} = f(\mathbf{H})$.

Approximating products with $f(\mathbf{H})$: Krylov subspace methods

If $f(x)$ is a degree k polynomial, then we can exactly commpute $f(\mathbf{H})\mathbf{v}$ using k matvecs with $\mathbf{H}$.

More generally we can approximate $f(\mathbf{H})\mathbf{v}$ from the information in the Krylov subspace

$$\mathcal{K}_{k+1}(\mathbf{H}, \mathbf{v}) = \{p(\mathbf{H})\mathbf{v} : \deg(p) \leq k\} = \operatorname{span}\{\mathbf{v}, \mathbf{H}\mathbf{v}, \dots, \mathbf{H}^k\mathbf{v}\}.$$

One well-known approach is by using the information generated by the Lanczos algorithm.[7]

[7] Druskin and Knizhnerman 1992; Saad 1992.

Hutch++ for matrix functions?

1. Form $\mathbf{Y} = \mathbf{AG}$, $\mathbf{G}$ $n \times m$ Gaussian
2. Form $\mathbf{Q} = \text{orth}(\mathbf{Y})$
3. Form $\tilde{\mathbf{A}} = \mathbf{Q}\mathbf{Q}^\top\mathbf{A}$
4. Compute $\text{tr}(\tilde{\mathbf{A}}) = \text{tr}(\mathbf{Q}^\top\mathbf{A}\mathbf{Q})$
5. Approximate $\text{tr}(\mathbf{A} - \tilde{\mathbf{A}}) = \text{tr}((\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{A})$ by $\frac{1}{m}\sum_{i=1}^{m} \mathbf{v}_i^\top(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{A}\mathbf{v}_i$

Thought: in step 1, approximate $\mathbf{AG}$ from $\text{span}\{\mathbf{G}, \mathbf{HG}, \ldots, \mathbf{H}^q\mathbf{G}\}$.

Hutch++ for matrix functions?

1. Form $\mathbf{Y} = \mathbf{A}\mathbf{G}$, $\mathbf{G}$ $n \times m$ Gaussian
2. Form $\mathbf{Q} = \text{orth}(\mathbf{Y})$
3. Form $\tilde{\mathbf{A}} = \mathbf{Q}\mathbf{Q}^\mathsf{T}\mathbf{A}$
4. Compute $\text{tr}(\tilde{\mathbf{A}}) = \text{tr}(\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q})$
5. Approximate $\text{tr}(\mathbf{A} - \tilde{\mathbf{A}}) = \text{tr}((\mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathsf{T})\mathbf{A})$ by $\frac{1}{m}\sum_{i=1}^{m} \mathbf{v}_i^\mathsf{T}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathsf{T})\mathbf{A}\mathbf{v}_i$

Thought: in step 1, approximate $\mathbf{A}\mathbf{G}$ from $\text{span}\{\mathbf{G}, \mathbf{H}\mathbf{G}, \ldots, \mathbf{H}^q\mathbf{G}\}$.

Observation: If we take $\mathbf{Q} = \text{span}\{\mathbf{G}, \mathbf{H}\mathbf{G}, \ldots, \mathbf{H}^q\mathbf{G}\}$ the projection stage will be better (for the same number of matvecs with $\mathbf{H}$)

Worry: In step 3, we will approximate $\mathbf{A}\mathbf{Q}$ from $\text{span}\{\mathbf{Q}, \mathbf{H}\mathbf{Q}, \ldots, \mathbf{H}^t\mathbf{Q}\}$. If $\mathbf{Q}$ has a lot of columns, this will be more expensive.

Krylov-aware trace estimation

Suppose $\mathbf{Q} = \text{span}\{\mathbf{G}, \mathbf{AG}, \ldots, \mathbf{A}^{q-1}\mathbf{G}\}$

Observation: We can build our approximation to $\mathbf{AQ}$ by continuing the block Krylov susbpace with $\mathbf{G}$.

$$\begin{aligned}
\text{span}\{\mathbf{Q}, \mathbf{HQ}, \ldots, \mathbf{H}^t\mathbf{Q}\} &= \text{span}\{\mathbf{G}, \mathbf{HG}, \ldots, \mathbf{H}^q\mathbf{G}, \\
&\qquad \mathbf{HG}, \mathbf{H}^2\mathbf{G}, \ldots, \mathbf{H}^{q+1}\mathbf{G}, \\
&\qquad\qquad \ddots \\
&\qquad\qquad \mathbf{H}^t\mathbf{G}, \mathbf{H}^t\mathbf{G}, \ldots, \mathbf{H}^{t+q}\mathbf{G}\} \\
&= \text{span}\{\mathbf{G}, \mathbf{HG}, \ldots, \mathbf{H}^{t+q}\mathbf{G}\}.
\end{aligned}$$

Krylov aware stochastic trace estimation[8]

This "Krylov aware" idea is simple, but provides many benefits.

- use a (much) larger projection space "for free"
- algorithm is now agnostic to f
  - we can easily compute approximations to $\operatorname{tr}(f(\mathbf{H}))$ for multiple f without additional matrix products with $\mathbf{H}$.
  - in particular, the approximation we get is a quadrature approximation for Ψ

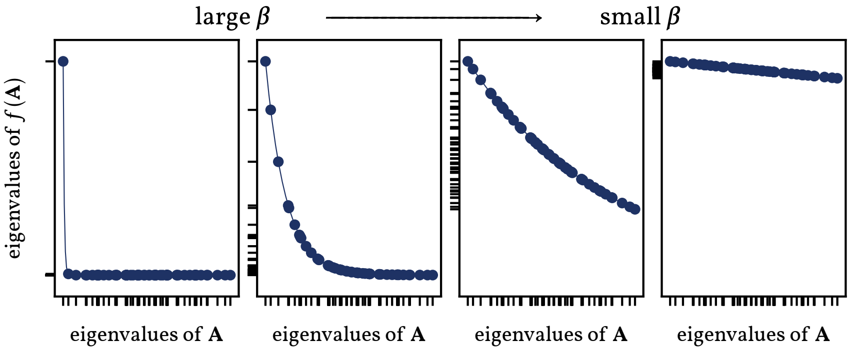
Note also that this illusrates that we should't just naievely use matvec query algorithms with Krylov subspace methods to compute matvecs with $\mathbf{A} = f(\mathbf{H})$.

[8]Chen and Hallman 2022.

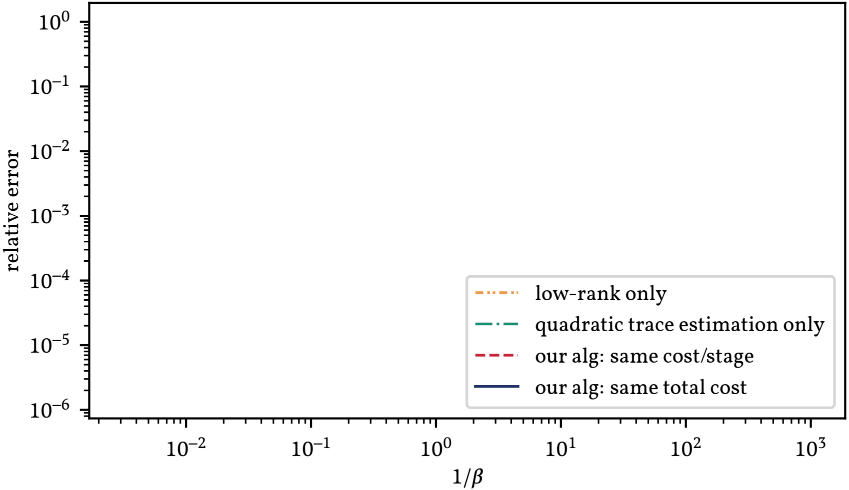
Example: equilibrium thermodynamics of quantum spin systems

In quantum physics, we often wish to compute $\operatorname{tr}(f(\mathbf{H})) = \operatorname{tr}(\exp(-\beta\mathbf{H}))$ for all $\beta > 0$.

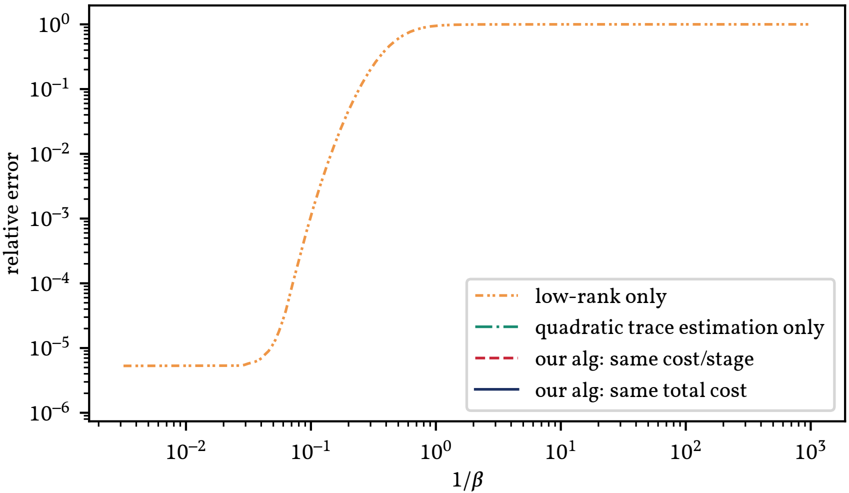
- if $\beta = \infty$ (zero temperature), then we only need ground state(s)
- if $\beta = 0$ (high temperature), then quadratic trace estimation works very well
- for intermediate beta, we might expect low-rank approaches to work well



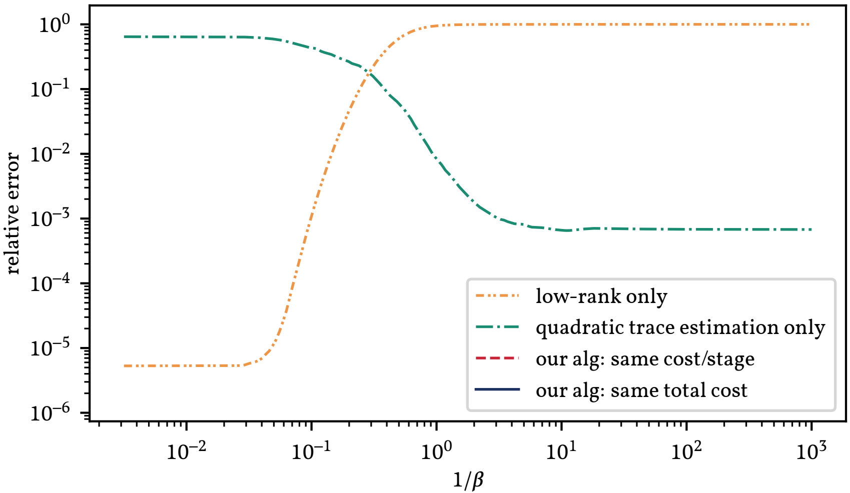
Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta \mathbf{A}))$



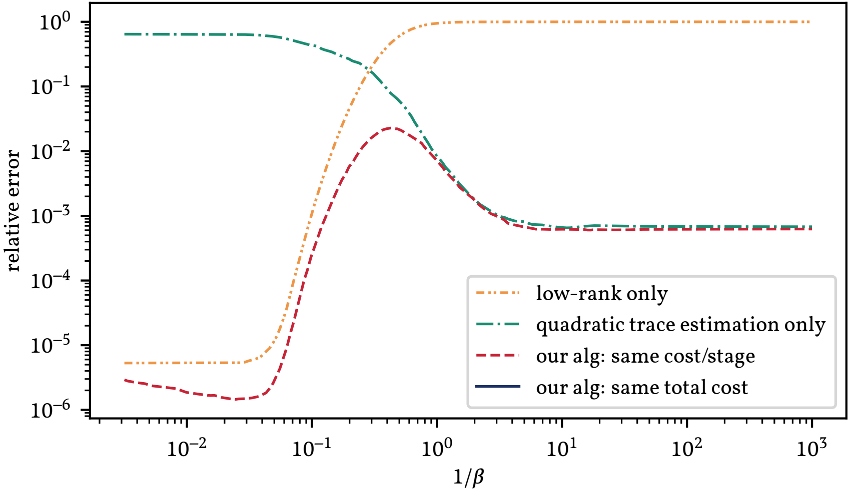
Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta\mathbf{A}))$



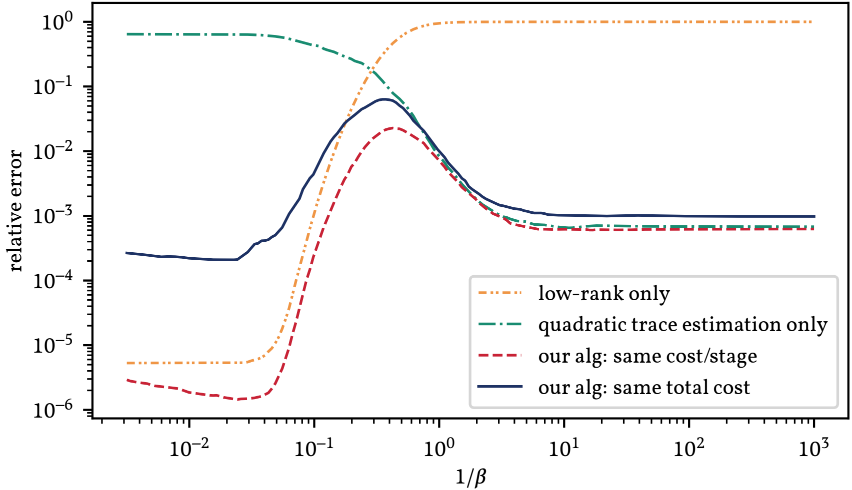
Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta\mathbf{A}))$



Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta\mathbf{A}))$



Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta \mathbf{A}))$



Variants

We also have a number of modifications to make this idea more practical:

- Using the information in the space $\operatorname{span}\{\mathbf{G}, \mathbf{AG}, \ldots, \mathbf{A}^{q+t}\mathbf{G}\}$ we can approximate

  $$\|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\mathsf{T} f(\mathbf{A})\|$$

  in order to determine a good value of q; see also[9]
- If memory or reorthogonalization costs are an issue, we can use restarting, and pick $\mathbf{Q} \subset \operatorname{span}\{\mathbf{G}, \mathbf{AG}, \ldots, \mathbf{A}^{q+1}\mathbf{G}\}$
  - e.g. $\mathbf{Q} = \mathbf{A}^{q+1}\mathbf{G}$

[9] Persson, Cortinovis, and Kressner 2022.

Conclusion

There is a lot of work on trace estimation in the matvec query model

Lots of potential for lower bounds

Instead of applying matvecs with $\mathbf{A} = f(\mathbf{H})$ with a black-box Krylov method, we should look into the box

References I

Chen, Tyler and Eric Hallman (2022). *Krylov-aware stochastic trace estimation*.

Druskin, Vladimir and Leonid Knizhnerman (July 1992). “Error Bounds in the Simple Lanczos Procedure for Computing Functions of Symmetric Matrices and Eigenvalues”. In: *Comput. Math. Math. Phys.* 31.7, pp. 20–30.

Epperly, Ethan (2023). *Stochastic trace estimation*.

Epperly, Ethan N. , Joel A. Tropp, and Robert J. Webber (2023). *XTrace: Making the most of every sample in stochastic trace estimation*.

Gambhir, Arjun Singh, Andreas Stathopoulos, and Kostas Orginos (Jan. 2017). “Deflation as a Method of Variance Reduction for Estimating the Trace of a Matrix Inverse”. In: *SIAM Journal on Scientific Computing* 39.2, A532–A558.

Girard, Didier (1987). *Un algorithme simple et rapide pour la validation croisée généralisée sur des problèmes de grande taille*.

Halikias, Diana and Alex Townsend (2022). *Structured matrix recovery from matrix-vector products*.

Hutchinson, M.F. (Jan. 1989). “A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines”. In: *Communications in Statistics - Simulation and Computation* 18.3, pp. 1059–1076.

Lin, Lin (Aug. 2016). “Randomized estimation of spectral densities of large matrices made accurate”. In: *Numerische Mathematik* 136.1, pp. 183–213.

Meyer, Raphael A. et al. (Jan. 2021). “Hutch++: Optimal Stochastic Trace Estimation”. In: *Symposium on Simplicity in Algorithms (SOSA)*. Society for Industrial and Applied Mathematics, pp. 142–155.

References II

Morita, Katsuhiro and Takami Tohyama (Feb. 2020). "Finite-temperature properties of the Kitaev-Heisenberg models on kagome and triangular lattices studied by improved finite-temperature Lanczos methods". In: *Physical Review Research* 2.1.

Persson, David, Alice Cortinovis, and Daniel Kressner (July 2022). "Improved Variants of the Hutch++ Algorithm for Trace Estimation". In: *SIAM Journal on Matrix Analysis and Applications* 43.3, pp. 1162–1185.

Saad, Yousef (1992). "Analysis of Some Krylov Subspace Approximations to the Matrix Exponential Operator". In: *SIAM Journal on Numerical Analysis* 29.1, pp. 209–228.

Wimmer, Karl, Yi Wu, and Peng Zhang (2014). "Optimal Query Complexity for Estimating the Trace of a Matrix". In: *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*. Ed. by Javier Esparza et al. Vol. 8572. Lecture Notes in Computer Science. Springer, pp. 1051–1062.

Wu, Lingfei et al. (2016). "Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse". In: *Journal of Computational Physics* 326, pp. 828–844.