Randomized matrix-free quadrature

Tyler Chen

September 30, 2022

`chen.pw/slides.pdf`

What is a matrix function?

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

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

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

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

Trace estimation

We will look at some algorithms for estimating

$$\operatorname{tr}(f(\mathbf{A})) = f(\lambda_1) + \cdots + f(\lambda_n).$$

These algorithms use:

- Stochastic trace estimation
- Krylov subspace methods

A number of widely used algorithms[1] fall into this class, but it's still an active area of algorithm development[2]

[1] Skilling 1989; Silver and Röder 1994; Silver, Roeder, Voter, and Kress 1996; Weiße, Wellein, Alvermann, and Fehske 2006; Bai, Fahey, and Golub 1996.

[2] Lin 2016; Gambhir, Stathopoulos, and Orginos 2017; Saibaba, Alexanderian, and Ipsen 2017; Morita and Tohyama 2020; Meyer, Musco, Musco, and Woodruff 2021; Li and Zhu 2021; Chen and Hallman 2022; Persson and Kressner 2022.

Direct methods

Can compute $f(\mathbf{A})$ via eigendecomposition of $\mathbf{A}$. However,

- this is slow: n^3 computation
- intractable storage costs: even if $\mathbf{A}$ is sparse, $f(\mathbf{A})$ typically is not
  - for $n = 2^{20}$, a $n \times n$ matrix of 64bit numbers requires 8.8 terrabytes

However, matrix products with $\mathbf{A}$ might be tractable.

Integral representation

We can write the trace as an integral

$$\operatorname{tr}(f(\mathbf{A})) = n \int f \mathrm{d}\Phi,$$

where the cumulative empirical spectral measure (CESM) Φ is

$$\Phi(x) := \sum_{i=1}^{n} \frac{1}{n} \mathbb{1}(\lambda_i \leq x) = \operatorname{tr}(\mathbb{1}(\mathbf{A} \leq x)).$$

Integral representation

We can write the trace as an integral

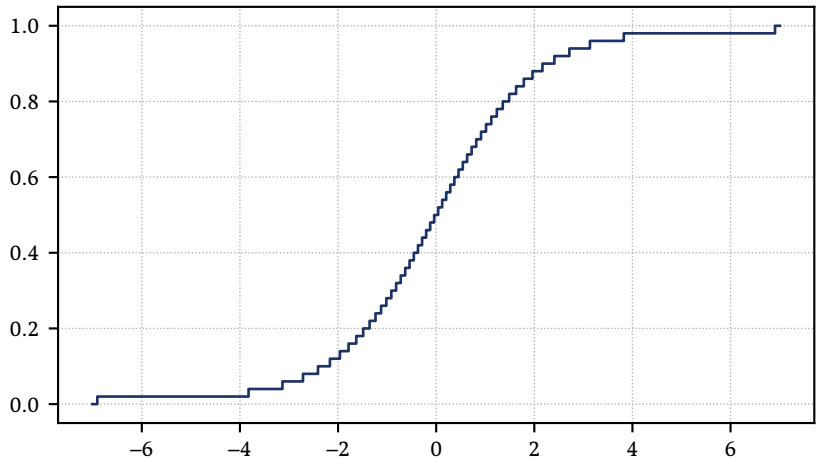
$$\operatorname{tr}(f(\mathbf{A})) = n \int f \mathrm{d}\Phi,$$

where the cumulative empirical spectral measure (CESM) Φ is

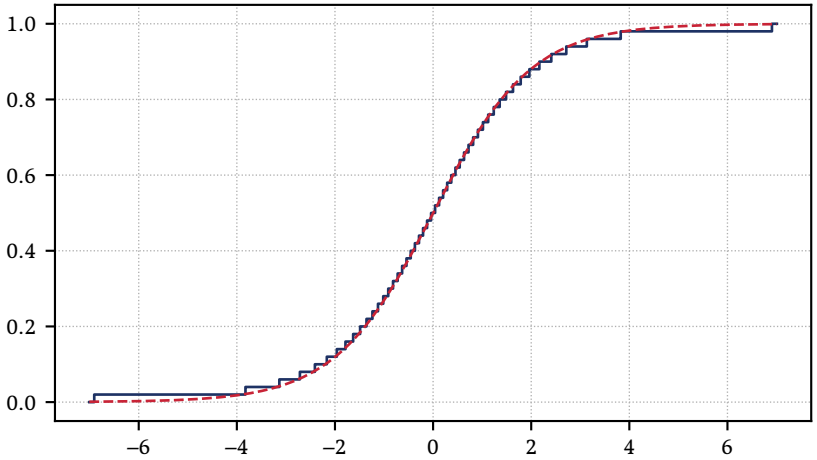
$$\Phi(x) := \sum_{i=1}^{n} \frac{1}{n} \mathbb{1}(\lambda_i \leq x) = \operatorname{tr}(\mathbb{1}(\mathbf{A} \leq x)).$$

trace approximation $\leftrightarrow$ CESM approximation

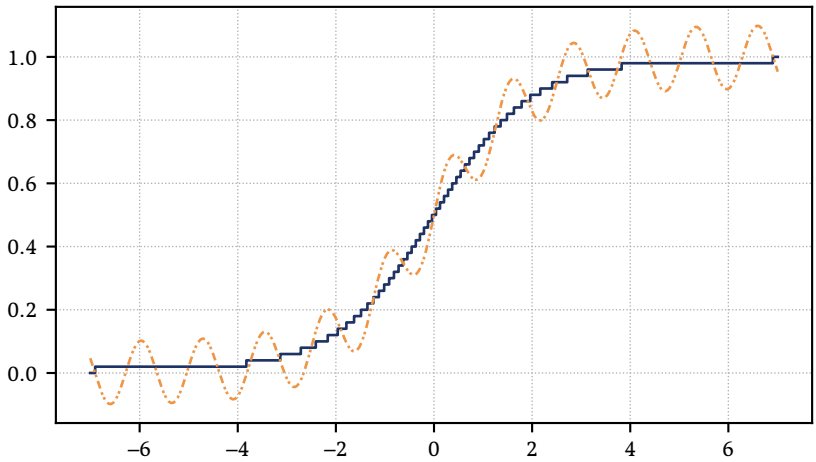
Approximating the CESM



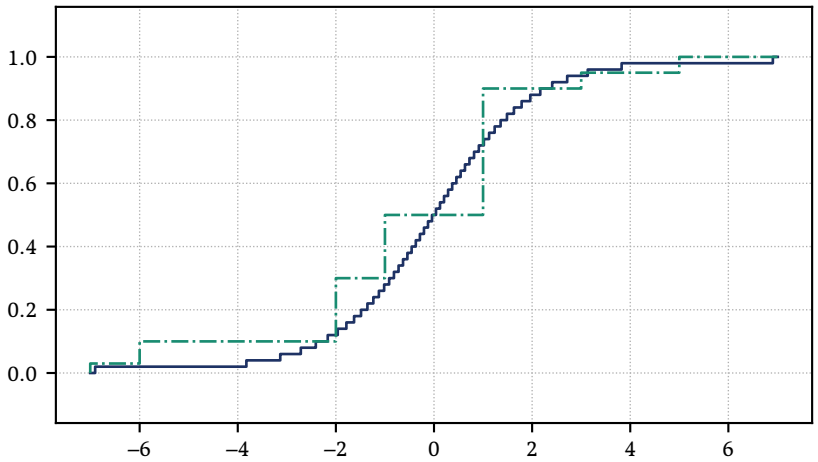
Approximating the CESM



Approximating the CESM



Approximating the CESM



Global approximation to CESM

Goal: Get a coarse approximation to Φ such that integrals are nearly preserved

Idea: Compute quadrature rule for Φ

- $\int x^k \mathrm{d}\Phi = \mathrm{tr}(\mathbf{A}^k)$

Global approximation to CESM

Goal: Get a coarse approximation to Φ such that integrals are nearly preserved

Idea: Compute quadrature rule for Φ

- $\int x^k d\Phi = \mathrm{tr}(\mathbf{A}^k)$

Problem: We can't compute integrals against Φ without computing traces

Idea: stochastic trace estimation

Stochastic trace estimation

It's well known that when $\mathbf{v}$ is such that $\mathbb{E}[\mathbf{v}\mathbf{v}^\mathsf{T}] = n^{-1}\mathbf{I}$, then

$$n\mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v}] = \operatorname{tr}(\mathbf{B}).$$

We call an estimator $\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v}$ a quadratic trace estimator.

[3] Alben, Blume, Krakauer, and Schwartz 1975.
[4] Schrödinger 1927; Neumann 1929.

Stochastic trace estimation

It's well known that when $\mathbf{v}$ is such that $\mathbb{E}[\mathbf{v}\mathbf{v}^\mathsf{T}] = n^{-1}\mathbf{I}$, then

$$n\mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v}] = \operatorname{tr}(\mathbf{B}).$$

We call an estimator $\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v}$ a quadratic trace estimator.

- Often attributed to Hutchinson 1989

[3] Alben, Blume, Krakauer, and Schwartz 1975.
[4] Schrödinger 1927; Neumann 1929.

Stochastic trace estimation

It's well known that when $\mathbf{v}$ is such that $\mathbb{E}[\mathbf{v}\mathbf{v}^\mathsf{T}] = n^{-1}\mathbf{I}$, then

$$n\mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v}] = \operatorname{tr}(\mathbf{B}).$$

We call an estimator $\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v}$ a quadratic trace estimator.

- Often attributed to Hutchinson 1989
- Earlier work by Girard 1987 and Skilling 1989
- Use of random states as algorithmic tool since at least 1970s[3]
- Morally equivalent to quantum typicality from late 1920s[4]

[3] Alben, Blume, Krakauer, and Schwartz 1975.
[4] Schrödinger 1927; Neumann 1929.

The weighted CESM

We can write quadratic forms as an integral

$$\mathbf{v}^\top f(\mathbf{A})\mathbf{v} = \int f \mathrm{d}\Psi$$

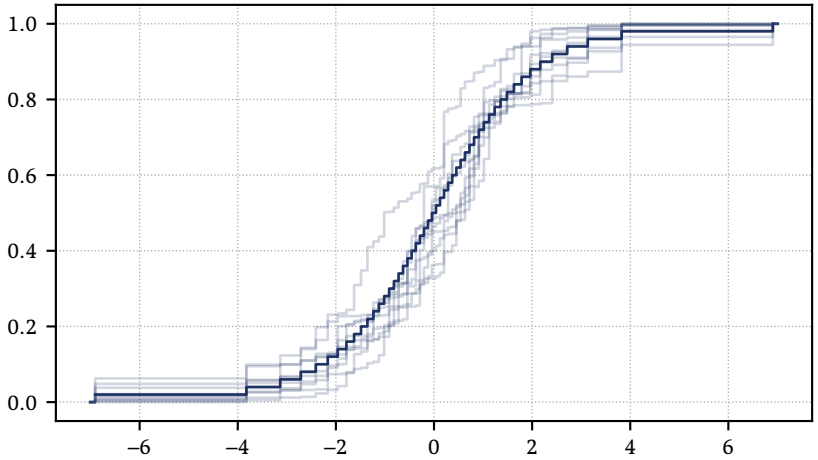
where

$$\Psi(x) := \sum_{i=1}^{n} |\mathbf{v}^\top \mathbf{u}_i|^2 \mathbb{1}(\lambda_i \leq x).$$

If $\mathbf{v}$ is an isotropic random vector,

$$\mathbb{E}[\Psi(x)] = \sum_{i=1}^{n} \mathbb{E}[|\mathbf{v}^\top \mathbf{u}_i|^2] \mathbb{1}(\lambda_i \leq x) = \sum_{i=1}^{n} \frac{1}{n} \mathbb{1}(\lambda_i \leq x) = \Phi(x).$$

The weighted CESM



Polynomial quadrature

Let $[f]_s^{\circ p}$ be a degree s polynomial approximation to f and $[\Psi]_s^{\circ q}$ the induced quadrature approximation to Ψ defined by

$$\int f \mathrm{d}[\Psi]_s^{\circ q} = \int [f]_s^{\circ p} \mathrm{d}\Psi.$$

Polynomial quadrature

Let $[f]_s^{\circ \mathrm{p}}$ be a degree s polynomial approximation to f and $[\Psi]_s^{\circ \mathrm{q}}$ the induced quadrature approximation to Ψ defined by

$$\int f \mathrm{d}[\Psi]_s^{\circ \mathrm{q}} = \int [f]_s^{\circ \mathrm{p}} \mathrm{d}\Psi.$$

Given a distribution function μ, let $\{p_i\}_{i=0}^{\infty}$ be the (normalized) orthogonal polynomials. Define moments,

$$m_i := \int p_i \mathrm{d}\Psi = \mathbf{v}^{\mathsf{T}} p_i(\mathbf{A})\mathbf{v}.$$

Polynomial quadrature

Let $[f]_s^{\circ \mathrm{p}}$ be a degree s polynomial approximation to f and $[\Psi]_s^{\circ \mathrm{q}}$ the induced quadrature approximation to Ψ defined by

$$\int f \mathrm{d}[\Psi]_s^{\circ \mathrm{q}} = \int [f]_s^{\circ \mathrm{p}} \mathrm{d}\Psi.$$

Given a distribution function μ, let $\{p_i\}_{i=0}^{\infty}$ be the (normalized) orthogonal polynomials. Define moments,

$$m_i := \int p_i \mathrm{d}\Psi = \mathbf{v}^{\mathsf{T}} p_i(\mathbf{A})\mathbf{v}.$$

If $[f]_s^{\circ \mathrm{p}} = c_0 p_0 + \cdots + c_s p_s$, then

$$\int f \mathrm{d}[\Psi]_s^{\circ \mathrm{q}} = \int \sum_{i=0}^{s} c_i p_i \mathrm{d}\Psi = \sum_{i=0}^{s} c_i \int p_i \mathrm{d}\Psi = \sum_{i=0}^{s} c_i m_i.$$

Algorithms

So, we get an algorithm:

$$\operatorname{tr}(f(\mathbf{A})) \approx \int f \mathrm{d}\langle [\Psi_i]_s^{\circ \mathrm{q}} \rangle = \frac{1}{n_{\mathrm{v}}} \sum_{i=1}^{n} \int f \mathrm{d}[\Psi_i]_s^{\circ \mathrm{q}}.$$

Different choices of $[f]_s^{\circ \mathrm{p}}$ correspond to different algorithms:

- Kernel Polynomial Method: damped Chebyshev approximation
- Stochastic Lanczos Quadrature: Interpolation at zeros of orthogonal polynomials of Ψ

Requires construction of Krylov subspaces

$$\operatorname{span}\{\mathbf{v}_i, \mathbf{A}\mathbf{v}_i, \dots, \mathbf{A}^k \mathbf{v}_i\}.$$

A unified framework[5]

Traditionally, algorithms like KPM would be implemented using an explicit Chebyshev recurrence

[5]Chen, Trogdon, and Ubaru 2022.

A unified framework[5]

Traditionally, algorithms like KPM would be implemented using an explicit Chebyshev recurrence

A key observation: we could instead use the output of Lanczos

- allows a posteriori choice of hyperparameters
  - more stable implementations
  - can cheaply try different quadrature approximations
- allows simultaneous theoretical analysis of algorithms
- allows tradeoffs between algorithms to be more clearly understood

[5] Chen, Trogdon, and Ubaru 2022.

Finite precision arithmetic

It's well know that the Lanczos algorithm is unstable, so people are afraid of using it without reorthogonalization[6] (expensive)

[6] Jaklič and Prelovšek 1994; Silver, Roeder, Voter, and Kress 1996; Aichhorn, Daghofer, Evertz, and Linden 2003; Weiße, Wellein, Alvermann, and Fehske 2006; Ubaru, Chen, and Saad 2017; Granziol, Wan, and Garipov 2019.

[7] Greenbaum 1989; Druskin and Knizhnerman 1992; Knizhnerman 1996.

Finite precision arithmetic

It's well know that the Lanczos algorithm is unstable, so people are afraid of using it without reorthogonalization[6] (expensive)

However, for certain tasks, the algorithm is actually backwards stable in a certain (useful sense)[7]even without reorthogonalization

- for these problems, "instability" is actually just an ill-conditioned task
- i.e. computing certain quantities is inherently hard, but we don't really need to compute them accurately to get what we want

[6]Jaklič and Prelovšek 1994; Silver, Roeder, Voter, and Kress 1996; Aichhorn, Daghofer, Evertz, and Linden 2003; Weiße, Wellein, Alvermann, and Fehske 2006; Ubaru, Chen, and Saad 2017; Granziol, Wan, and Garipov 2019.

[7]Greenbaum 1989; Druskin and Knizhnerman 1992; Knizhnerman 1996.

Finite precision arithmetic

It's well know that the Lanczos algorithm is unstable, so people are afraid of using it without reorthogonalization[6] (expensive)

However, for certain tasks, the algorithm is actually backwards stable in a certain (useful sense)[7]even without reorthogonalization

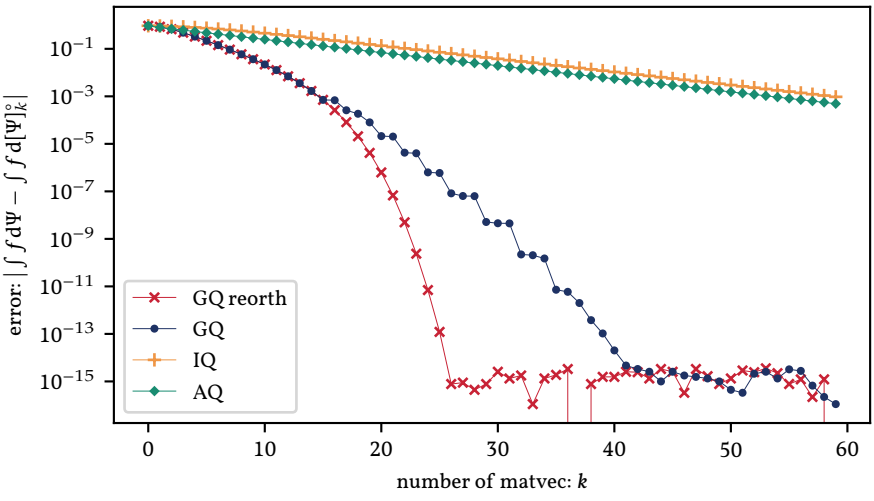
- for these problems, "instability" is actually just an ill-conditioned task
- i.e. computing certain quantities is inherently hard, but we don't really need to compute them accurately to get what we want

Moreover, direct implementations of alternate algorithms like KPM are stable given good hyper-parameter choice, but exponentially unstable otherwise.

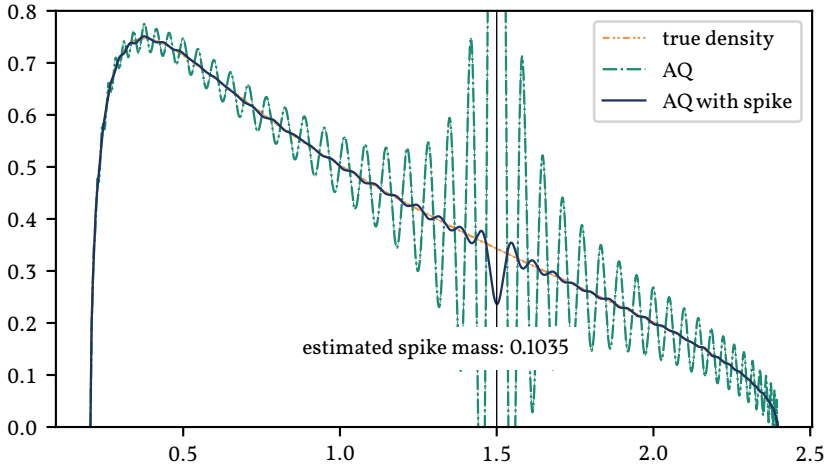
[6]Jaklič and Prelovšek 1994; Silver, Roeder, Voter, and Kress 1996; Aichhorn, Daghofer, Evertz, and Linden 2003; Weiße, Wellein, Alvermann, and Fehske 2006; Ubaru, Chen, and Saad 2017; Granziol, Wan, and Garipov 2019.

[7]Greenbaum 1989; Druskin and Knizhnerman 1992; Knizhnerman 1996.

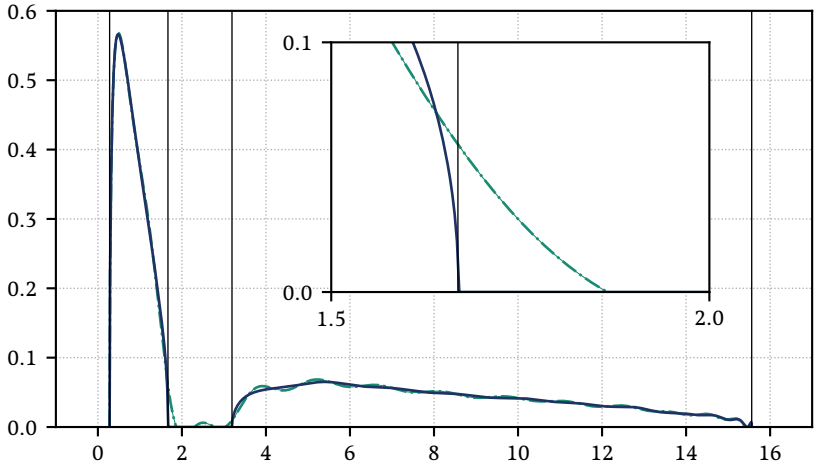
Example: Runge function



Example: random matrtix



Example: random matrtix



A simple analysis of the variance (for Gaussians) implies

$$|\mathbf{B} - \langle \mathbf{v}_\ell^\mathsf{T} \mathbf{B} \mathbf{v}_\ell \rangle| \sim \|\mathbf{B}\|_\mathsf{F} (n_\mathrm{v})^{-1/2}.$$

[11]Avron and Toledo 2011; Roosta-Khorasani and Ascher 2014 [12]Meyer, Musco, Musco, and Woodruff 2021; Cortinovis and Kressner 2021; Persson, Cortinovis, and Kressner 2022

[10]Reimann 2007 [11]Popescu, Short, and Winter 2006; Gogolin 2010 [12]Chen, Trogdon, and Ubaru 2021

Beyond quadratic trace estimation

A simple analysis of the variance (for Gaussians) implies

$$|\mathbf{B} - \langle \mathbf{v}_\ell^\top \mathbf{B} \mathbf{v}_\ell \rangle| \sim \|\mathbf{B}\|_{\mathsf{F}} (n_{\mathrm{v}})^{-1/2}.$$

More refined concentration inequalities

$$\mathbb{P}\big[\, |\mathbf{B} - \langle \mathbf{v}_\ell^\top \mathbf{B} \mathbf{v}_\ell \rangle| > \epsilon \,\big] \leq \delta.$$

– Applied Math/CS (iid entries): early analyses[8], refined analyses[9]

[11]Avron and Toledo 2011; Roosta-Khorasani and Ascher 2014 [12]Meyer, Musco, Musco, and Woodruff 2021; Cortinovis and Kressner 2021; Persson, Cortinovis, and Kressner 2022

[10]Reimann 2007 [11]Popescu, Short, and Winter 2006; Gogolin 2010 [12]Chen, Trogdon, and Ubaru 2021

Beyond quadratic trace estimation

A simple analysis of the variance (for Gaussians) implies

$$|\mathbf{B} - \langle \mathbf{v}_\ell^\top \mathbf{B} \mathbf{v}_\ell \rangle| \sim \|\mathbf{B}\|_{\mathsf{F}} (n_{\mathrm{v}})^{-1/2}.$$

More refined concentration inequalities

$$\mathbb{P}\big[\, |\mathbf{B} - \langle \mathbf{v}_\ell^\top \mathbf{B} \mathbf{v}_\ell \rangle| > \epsilon \,\big] \leq \delta.$$

- Applied Math/CS (iid entries): early analyses[8], refined analyses[9]
- Physics (uniform from hypersphere): Chebyshev[10], sub-Gaussian[11](via Lèvy's Lemma), refined bounds for practical dimensions[12]

[11]Avron and Toledo 2011; Roosta-Khorasani and Ascher 2014 [12]Meyer, Musco, Musco, and Woodruff 2021; Cortinovis and Kressner 2021; Persson, Cortinovis, and Kressner 2022

[10]Reimann 2007 [11]Popescu, Short, and Winter 2006; Gogolin 2010 [12]Chen, Trogdon, and Ubaru 2021

Low rank approximation

We can always decompose

$$\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\widehat{\mathbf{B}}) + \operatorname{tr}(\widetilde{\mathbf{B}}), \qquad \text{where} \qquad \widetilde{\mathbf{B}} := \mathbf{B} - \widehat{\mathbf{B}}.$$

[13] Lin 2016; Gambhir, Stathopoulos, and Orginos 2017; Saibaba, Alexanderian, and Ipsen 2017; Morita and Tohyama 2020; Li and Zhu 2021; Meyer, Musco, Musco, and Woodruff 2021; Chen and Hallman 2022; Persson and Kressner 2022.

Low rank approximation

We can always decompose

$$\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\widehat{\mathbf{B}}) + \operatorname{tr}(\widetilde{\mathbf{B}}), \qquad \text{where} \qquad \widetilde{\mathbf{B}} := \mathbf{B} - \widehat{\mathbf{B}}.$$

So, let's output

$$\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\widehat{\mathbf{B}}) + \langle \boldsymbol{\psi}_\ell^\mathsf{T} \widetilde{\mathbf{B}} \boldsymbol{\psi}_\ell \rangle = \operatorname{tr}(\widehat{\mathbf{B}}) + \frac{1}{m} \operatorname{tr}(\boldsymbol{\Psi}^\mathsf{T} \widetilde{\mathbf{B}} \boldsymbol{\Psi})$$

[13] Lin 2016; Gambhir, Stathopoulos, and Orginos 2017; Saibaba, Alexanderian, and Ipsen 2017; Morita and Tohyama 2020; Li and Zhu 2021; Meyer, Musco, Musco, and Woodruff 2021; Chen and Hallman 2022; Persson and Kressner 2022.

Low rank approximation

We can always decompose

$$\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\widehat{\mathbf{B}}) + \operatorname{tr}(\widetilde{\mathbf{B}}), \qquad \text{where} \qquad \widetilde{\mathbf{B}} := \mathbf{B} - \widehat{\mathbf{B}}.$$

So, let's output

$$\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\widehat{\mathbf{B}}) + \langle \boldsymbol{\psi}_\ell^\mathsf{T} \widetilde{\mathbf{B}} \boldsymbol{\psi}_\ell \rangle = \operatorname{tr}(\widehat{\mathbf{B}}) + \frac{1}{m} \operatorname{tr}(\boldsymbol{\Psi}^\mathsf{T} \widetilde{\mathbf{B}} \boldsymbol{\Psi})$$

This is beneficial if:

- $\operatorname{tr}(\widehat{\mathbf{B}})$ can be computed efficiently, and
- the variance of $\operatorname{tr}(\boldsymbol{\Psi}^\mathsf{T} \widetilde{\mathbf{B}} \boldsymbol{\Psi})$ is reduced compared to $\operatorname{tr}(\boldsymbol{\Psi}^\mathsf{T} \mathbf{B} \boldsymbol{\Psi})$

A lot of recent work[13] uses this idea to varying extent

[13] Lin 2016; Gambhir, Stathopoulos, and Orginos 2017; Saibaba, Alexanderian, and Ipsen 2017; Morita and Tohyama 2020; Li and Zhu 2021; Meyer, Musco, Musco, and Woodruff 2021; Chen and Hallman 2022; Persson and Kressner 2022.

Hutch++[14]

Take $\widehat{\mathbf{B}}$ as low rank approximation $\widehat{\mathbf{B}} = \mathbf{Q}\mathbf{Q}^\top\mathbf{B}\mathbf{Q}\mathbf{Q}^\top$

- Compute $\mathbf{Q}$ by sketching: $\mathbf{Q} = \textsc{orth}(\mathbf{B}\mathbf{\Omega})$ where $\mathbf{\Omega}$ is $n \times b$ random matrix

Simplify a bit:

- $\operatorname{tr}(\widehat{\mathbf{B}}) = \operatorname{tr}(\mathbf{Q}\mathbf{Q}^\top\mathbf{B}\mathbf{Q}\mathbf{Q}^\top) = \operatorname{tr}(\mathbf{Q}^\top\mathbf{B}\mathbf{Q})$
- $\operatorname{tr}(\widetilde{\mathbf{B}}) = \operatorname{tr}((\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{B}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top))$
- $\operatorname{tr}(\mathbf{\Psi}^\top\widetilde{\mathbf{B}}\mathbf{\Psi}) = \operatorname{tr}(\mathbf{Y}^\top\mathbf{B}\mathbf{Y})$, where $\mathbf{Y} = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{\Psi}$.

Number of matvecs with $\mathbf{B}$ is: $2b + m$, and if we set $b = m$, can get scaling

$$\text{accuracy} \sim (\#\ \text{matvecs})^{-1}$$

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

What about matrix functions? (i.e. $\mathbf{B} = f(\mathbf{A})$)

Suppose we use q iterations of Lanczos to approximate $f(\mathbf{A})\mathbf{\Omega}$.

What about matrix functions? (i.e. $\mathbf{B} = f(\mathbf{A})$)

Suppose we use q iterations of Lanczos to approximate $f(\mathbf{A})\mathbf{\Omega}$. Then, at least implicitly, we construct

$$\text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^q\mathbf{\Omega}\}$$

Naive Hutch++ would take $\mathbf{Q} \in \mathbb{R}^{n \times b}$ as a basis for our approximation to $f(\mathbf{A})\mathbf{\Omega}$.

What about matrix functions? (i.e. $\mathbf{B} = f(\mathbf{A})$)

Suppose we use q iterations of Lanczos to approximate $f(\mathbf{A})\mathbf{\Omega}$. Then, at least implicitly, we construct

$$\text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^q\mathbf{\Omega}\}$$

Naive Hutch++ would take $\mathbf{Q} \in \mathbb{R}^{n \times b}$ as a basis for our approximation to $f(\mathbf{A})\mathbf{\Omega}$.

Instead, take $\mathbf{Q} \in \mathbb{R}^{n \times (q+1)b}$ as basis for the whole Krylov subspace

- Same number of matvecs with $\mathbf{A}$, but (much) larger search space..

Efficient approximation of $f(\mathbf{A})\mathbf{Q}$

Regardless of our choice of $\mathbf{Q}$, the next step is approximating $f(\mathbf{A})\mathbf{Q}$.

Suppose we use n iterations of Lanczos. Then we need to construct

$$\text{span}\{\mathbf{Q}, \mathbf{AQ}, \dots, \mathbf{A}^n\mathbf{Q}\}$$

If $\mathbf{Q}$ has $(q+1)b$ instead of b columns, this ostensibly requires $n(q+1)b$ matvecs with $\mathbf{A}$ instead of nb required by a naive implementation.

Efficient approximation of $f(\mathbf{A})\mathbf{Q}$

Recall, $\mathbf{Q}$ is a basis for

$$\text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^q\mathbf{\Omega}\}.$$

Thus, the columns of $\mathbf{A}^j\mathbf{Q}$ span

$$\text{span}\{\mathbf{A\Omega}, \mathbf{A}^2\mathbf{\Omega}, \dots, \mathbf{A}^{q+1}\mathbf{\Omega}\}.$$

Efficient approximation of $f(\mathbf{A})\mathbf{Q}$

Recall, $\mathbf{Q}$ is a basis for

$$\text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^q\mathbf{\Omega}\}.$$

Thus, the columns of $\mathbf{AQ}$ span

$$\text{span}\{\mathbf{A}^j\mathbf{\Omega}, \mathbf{A}^{j+1}\mathbf{\Omega}, \dots, \mathbf{A}^{q+j}\mathbf{\Omega}\}.$$

Efficient approximation of $f(\mathbf{A})\mathbf{Q}$

Recall, $\mathbf{Q}$ is a basis for

$$\text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^q\mathbf{\Omega}\}.$$

Thus, the columns of $\mathbf{A}^j\mathbf{Q}$ span

$$\text{span}\{\mathbf{A\Omega}, \mathbf{A}^2\mathbf{\Omega}, \dots, \mathbf{A}^{q+1}\mathbf{\Omega}\}.$$

So actually,

$$\text{span}\{\mathbf{Q}, \mathbf{AQ}, \dots, \mathbf{A}^n\mathbf{Q}\} = \text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^{q+n}\mathbf{\Omega}\}.$$

Efficient approximation of $f(\mathbf{A})\mathbf{Q}$

Recall, $\mathbf{Q}$ is a basis for

$$\text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^q\mathbf{\Omega}\}.$$

Thus, the columns of $\mathbf{A}^j\mathbf{Q}$ span

$$\text{span}\{\mathbf{A\Omega}, \mathbf{A}^2\mathbf{\Omega}, \dots, \mathbf{A}^{q+1}\mathbf{\Omega}\}.$$

So actually,

$$\text{span}\{\mathbf{Q}, \mathbf{AQ}, \dots, \mathbf{A}^n\mathbf{Q}\} = \text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^{q+n}\mathbf{\Omega}\}.$$

In other words, to approximate $f(\mathbf{A})\mathbf{Q}$ we only need nb matrix-vector products!

Krylov aware stochastic trace estimation[15]

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

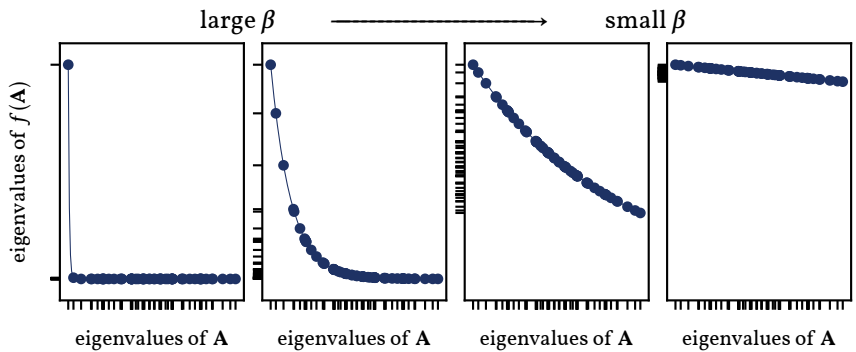
- use a (much) larger projection space "for free"
- algorithm is now agnostic to f
  - we can easily compute approximations to $\operatorname{tr}(f(\mathbf{A}))$ for multiple f without additional matrix products with $\mathbf{A}$.
  - in particular, the approximation we get is a quadrature approximation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

[15] Chen and Hallman 2022.

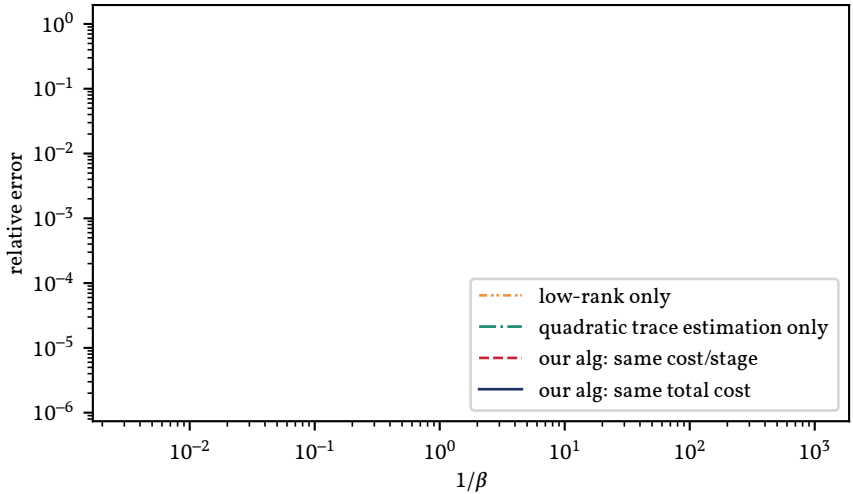
Example: equilibrium thermodynamics of quantum spin systems

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

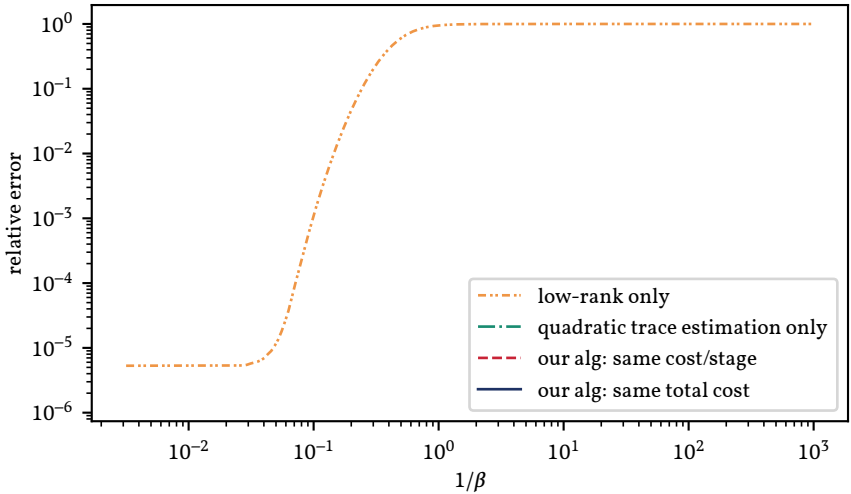
- if $\beta = \infty$ (zero temperature), then we only need ground state(s)
- if $\beta = 0$ (high temperature), then quadratic trace estimation works very well
- for intermediate beta, we might expect low-rank approaches to work well



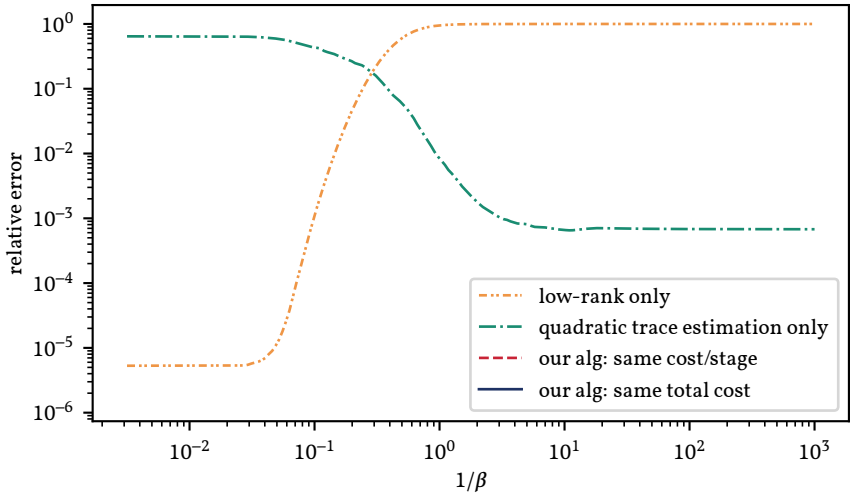
Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta \mathbf{A}))$



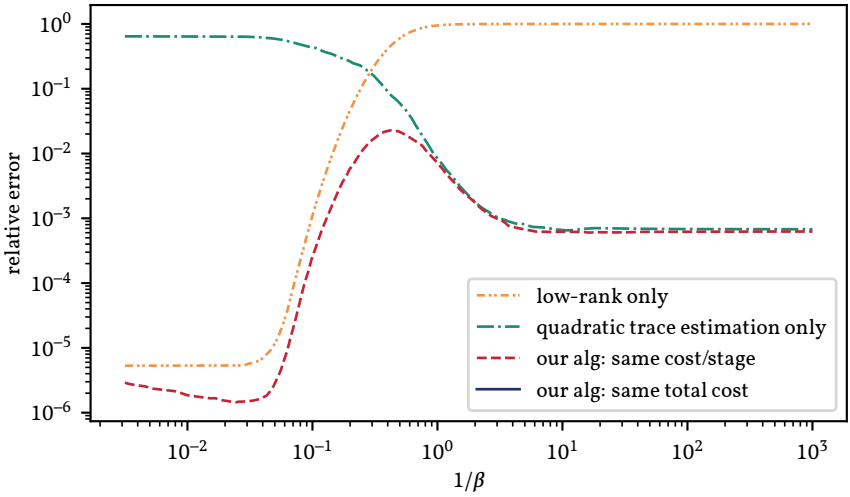
Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta \mathbf{A}))$



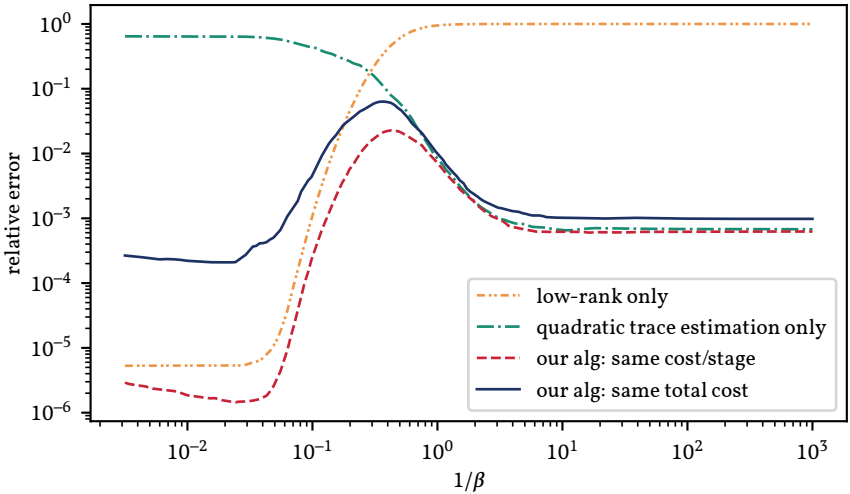
Example: quantum spin systems; $\mathrm{tr}(\exp(-\beta \mathbf{A}))$



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 $\text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^{q+n}\mathbf{\Omega}$ we can approximate

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

  in order to determine a good value of q.
- If memory or reorthogonalization costs are an issue, we can use restarting, and pick $\mathbf{Q} \subset \text{span}\{\mathbf{\Omega}, \mathbf{A\Omega}, \dots, \mathbf{A}^{q+1}\mathbf{\Omega}\}$
  - e.g. $\mathbf{Q} = \mathbf{A}^{q+1}\mathbf{\Omega}$

Future work

- $\mathrm{tr}(\exp(-\beta(\mathbf{A} + h\mathbf{B})))$ for all $\beta > 0$, $h \in [-h_0, h_0]$.
- generalize low-rank algorithms to partial traces
- better understanding of stability
- lower bounds in matrix-vector query models
- better relationships between physics, applied math, and CS

References I

Aichhorn, Markus et al. (Apr. 2003). "Low-temperature Lanczos method for strongly correlated systems". In: *Physical Review B* 67.16.

Alben, R. et al. (Nov. 1975). "Exact results for a three-dimensional alloy with site diagonal disorder: comparison with the coherent potential approximation". In: *Physical Review B* 12.10, pp. 4090–4094.

Avron, Haim and Sivan Toledo (Apr. 2011). "Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix". In: *Journal of the ACM* 58.2, pp. 1–34.

Bai, Zhaojun, Gark Fahey, and Gene Golub (Nov. 1996). "Some large-scale matrix computation problems". In: *Journal of Computational and Applied Mathematics* 74.1-2, pp. 71–89.

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

Chen, Tyler, Thomas Trogdon, and Shashanka Ubaru (18–24 Jul 2021). "Analysis of stochastic Lanczos quadrature for spectrum approximation". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 1728–1739.

— (2022). *Randomized matrix-free quadrature for spectrum and spectral sum approximation*.

Cortinovis, Alice and Daniel Kressner (July 2021). "On Randomized Trace Estimates for Indefinite Matrices with an Application to Determinants". In: *Foundations of Computational Mathematics*.

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.

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*.

Gogolin, Christian (2010). "Pure State Quantum Statistical Mechanics". MA thesis. Julius-Maximilians-Universität Würzburg.

Granziol, Diego, Xingchen Wan, and Timur Garipov (2019). *Deep Curvature Suite*.

References II

Greenbaum, Anne (1989). "Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences". In: *Linear Algebra and its Applications* 113, pp. 7–63.

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.

Jaklič, J. and P. Prelovšek (Feb. 1994). "Lanczos method for the calculation of finite-temperature quantities in correlated systems". In: *Physical Review B* 49.7, pp. 5065–5068.

Knizhnerman, L. A. (Jan. 1996). "The Simple Lanczos Procedure: Estimates of the Error of the Gauss Quadrature Formula and Their Applications". In: *Comput. Math. Math. Phys.* 36.11, pp. 1481–1492.

Li, Hanyu and Yuanyang Zhu (2021). "Randomized block Krylov subspace methods for trace and log-determinant estimators". In: *BIT Numerical Mathematics*, pp. 1–29.

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.

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.

Musco, Cameron, Christopher Musco, and Aaron Sidford (2018). "Stability of the Lanczos Method for Matrix Function Approximation". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '18. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, pp. 1605–1624.

Neumann, J. von (Jan. 1929). "Beweis des Ergodensatzes und des*H*-Theorems in der neuen Mechanik". In: *Zeitschrift für Physik* 57.1-2. English translation `https://arxiv.org/abs/1003.2133`, pp. 30–70.

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.

Persson, David and Daniel Kressner (2022). *Randomized low-rank approximation of monotone matrix functions*.

References III

Popescu, Sandu, Anthony J. Short, and Andreas Winter (Oct. 2006). "Entanglement and the foundations of statistical mechanics". In: *Nature Physics* 2.11, pp. 754–758.

Reimann, Peter (Oct. 2007). "Typicality for Generalized Microcanonical Ensembles". In: *Physical Review Letters* 99.16.

Roosta-Khorasani, Farbod and Uri Ascher (Sept. 2014). "Improved Bounds on Sample Size for Implicit Matrix Trace Estimators". In: *Foundations of Computational Mathematics* 15.5, pp. 1187–1212.

Saibaba, Arvind K, Alen Alexanderian, and Ilse CF Ipsen (2017). "Randomized matrix-free trace and log-determinant estimators". In: *Numerische Mathematik* 137.2, pp. 353–395.

Schrödinger, E. (1927). "Energieaustausch nach der Wellenmechanik". In: *Annalen der Physik* 388.15, pp. 956–968.

Silver, R.N. and H. Röder (Aug. 1994). "Densities of states of mega-dimensional Hamiltonian matrices". In: *International Journal of Modern Physics C* 05.04, pp. 735–753.

Silver, R.N. et al. (Mar. 1996). "Kernel Polynomial Approximations for Densities of States and Spectral Functions". In: *Journal of Computational Physics* 124.1, pp. 115–130.

Skilling, John (1989). "The Eigenvalues of Mega-dimensional Matrices". In: *Maximum Entropy and Bayesian Methods*. Springer Netherlands, pp. 455–466.

Ubaru, Shashanka, Jie Chen, and Yousef Saad (2017). "Fast Estimation of $tr(f(A))$ via Stochastic Lanczos Quadrature". In: *SIAM Journal on Matrix Analysis and Applications* 38.4, pp. 1075–1099.

Weiße, Alexander et al. (Mar. 2006). "The kernel polynomial method". In: *Reviews of Modern Physics* 78.1, pp. 275–306.