Randomized trace estimation and determinants

Alice Cortinovis

Joint work with Daniel Kressner

4 May 2022 – University of Arkansas Spring Lecture Series

Motivation

When do we need to compute the determinant?
What does it have to do with trace estimation?

Example: Gaussian process regression

J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. GPy-Torch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. (NeurIPS 2018)

Gaussian process: Distribution with mean $\mu(\cdot)$ and covariance $k(\cdot,\cdot)$.

Goal: Train a Gaussian process on *a lot* of data (up to 500k points), that is, given a class of possible k depending on some *hyperparameters* θ, find θ that best fit the data by minimizing

$$L(\theta \mid \text{training data } X, y) := \log\det(K) - y^T K^{-1} y,$$

where K is the discretization of $k(\cdot,\cdot)$ on $X \times X$.

They use randomized trace estimation!

S. Ubaru, J. Chen, and Y. Saad. Fast estimation of tr(f(A)) via stochastic Lanczos quadrature. (SIAM J. Matrix Anal. Appl., 2017)

Applications for computing determinants

- Statistical learning
  - Y. Zhang & W. E. Leithead Approximate implementation of the logarithm of the matrix determinant in Gaussian process regression (Journal of Statistical Computation and Simulation, 2007)
  - R. H. Affandi, E. Fox, R. Adams, and B. Taskar. Learning the parameters of determinantal point process kernels. (International Conference on Machine Learning 2014)
  - I. Han, D. Malioutov, and J. Shin Large-scale log-determinant computation through stochastic Chebyshev expansions. (International Conference on Machine Learning 2015)
  - K. Dong, D. Eriksson, H. Nickisch, D. Bindel, and A. Wilson. Scalable Log Determinants for Gaussian Process Kernel Learning. (NeurIPS 2017)
  - J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. GPy-Torch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. (NeurIPS 2018)
- Lattice quantum chromodynamics
  - C. Thron, S. J. Dong, K. F. Liu, and H. P. Ying. Padé-Z_2 estimator of determinants. Physical Review D - Particles, Fields, Gravitation and Cosmology, 1998)
- Markov random fields models
- Graph theory: $\det =$ number of spanning trees

Which algorithm would you use to compute the determinant of a symmetric positive definite (SPD) matrix B?

Which algorithm would you use to compute the determinant of a symmetric positive definite (SPD) matrix B?

- Definition by Laplace expansion (exponential time!)
- Compute the eigenvalues and take their product (cubic time)
- Take Cholesky factorization $B = LL^T$ for lower triangular matrix L and take $\det(B) = L_{11}^2 \cdots L_{nn}^2$ (this is what Matlab does – still cubic time!)

The determinant as the trace of the matrix logarithm

Definition (Matrix function [Higham'2008 book])

For a symmetric matrix $B = \begin{bmatrix} Q \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} Q^T \end{bmatrix}$

and a real-valued function f defined on the eigenvalues of B we define

$$f(B) := \begin{bmatrix} Q \end{bmatrix} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_n) \end{bmatrix} \begin{bmatrix} Q^T \end{bmatrix}$$

Theorem

For an SPD matrix B, we have $\log\det(B) = \mathrm{trace}(\log(B))$.

Note: We *do not* want to compute the whole $\log B$.

Other applications of trace estimation

- Frobenius norm estimation: $\|B\|_F^2 = \mathrm{trace}(B^T B)$.
  [Gratton/Titley-Peloquin'2018], [Roosta-Khorasani/Székely/Ascher'2015].
- Other Schatten-p norms: $\|B\|_p^p = \mathrm{trace}((B^T B)^{p/2})$
  [Dudley/Saibaba/Alexanderian'2020].
- Trace of other matrix functions:
  - Trace of the inverse (Uncertainty quantification [Kalantzis/Bekas/Curioni/Gallopoulos'2013], Lattice quantum chromodynamics [Wu et al.'2016])
  - Counting #triangles in a graph ($\frac{1}{6}\mathrm{trace}(B^3)$) [Avron'2010]
  - Trace of $\exp(B)$ is the Estrada index of a graph (see Prof. Saad's lecture tomorrow).

The Hutchinson trace estimator

1. Want to estimate $\mathrm{trace}(A)$ for symmetric (indefinite) $A \in \mathbb{R}^{n \times n}$
2. Can only do matrix-vector products with A

$\rightsquigarrow$ Why does this setting make sense?
When $A = f(B)$, computing $f(B)v$ is faster than computing $f(B)$!

Outline of the (rest of the) talk

1. The Hutchinson trace estimator
   - The estimator
   - Convergence analysis
   - Numerical examples
2. Approximation of the quadratic forms
   - Relation to quadrature
   - Convergence results (briefly)
3. Combined results for log-determinant approximation and numerical experiments
4. Conclusion

Hutchinson trace estimator

Theorem

For a random vector $X = (X_1, \dots, X_n)^T$ *such that* $\mathbb{E}[XX^T] = I$ *it holds*

$$\mathbb{E}\big[X^T A X\big] = \mathrm{trace}(A).$$

Most common choices for X:

- Gaussian vectors ($X \sim \mathcal{N}(0, I_n)$) $\rightsquigarrow$ variance $= 2\|A\|_F^2$.
- Rademacher vectors (± 1 entries) $\rightsquigarrow$ variance $= 2\|A - \mathrm{diag}(A)\|_F^2$;

Idea: Take N independent copies $X^{(1)}, \dots, X^{(N)}$ of X and approximate

$$\mathrm{trace}(A) \approx \mathrm{trace}_N(A) := \frac{1}{N}\sum_{i=1}^{N} (X^{(i)})^T A X^{(i)}.$$

Hutchinson's trace estimator: Example

$$\text{trace}(A) \approx \text{trace}_N(A) := \frac{1}{N}\sum_{i=1}^{N}(X^{(i)})^T A X^{(i)}$$

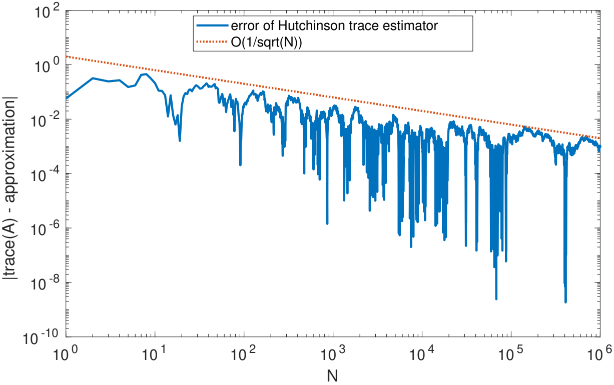


Figure: The behavior of $|\text{trace}(A) - \text{trace}_N(A)|$ when increasing N (# probe vectors).

Existing bounds for indefinite matrices

For SPD matrices, bounds on relative error are well studied: [Avron/Toledo'2011], [Roosta-Khorasani/Ascher'2015], [Gratton/Titley-Peloquin'2018].

For indefinite case (e.g. $A = \log B$), we cannot expect bound on relative error!

Theorem ([Avron'2010])

We have $\mathbb{P}(|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \varepsilon) \leq \delta$ *for*

$$N \geq \frac{6}{\varepsilon^2}\|A\|_*^2 \log\frac{2n}{\delta} \text{ (Rademacher)}, \quad N \geq \frac{20}{\varepsilon^2}\|A\|_*^2 \log\frac{4}{\delta} \text{ (Gaussian)}$$

where $\|A\|_* =$ *nuclear norm (sum of singular values).*

In the context of log-determinant:

Theorem ([Ubaru/Chen/Saad'2017])

For SPD matrix B *with condition number* $\kappa(B)$ *and* $N \geq \frac{24n^2}{\varepsilon^2}\log\frac{2}{\delta}\log(1+\kappa(B)^2)$ *we have (Rademacher case)*

$$\mathbb{P}\left(|\mathrm{trace}_N(\log B) - \log\det B| \geq \varepsilon\right) \leq \delta.$$

One quadratic form instead of N

Embedding trick: write $\operatorname{trace}_N(A) = \frac{1}{N}\sum_{i=1}^{N}(X^{(i)})^T A X^{(i)}$

$$= \begin{bmatrix} {X^{(1)}}^T & \dots & {X^{(N)}}^T \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{N}A & & & & \\ & \frac{1}{N}A & & & \\ & & \ddots & & \\ & & & \frac{1}{N}A & \\ & & & & \frac{1}{N}A \end{bmatrix} \cdot \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(N)} \end{bmatrix}$$

⟵Rad./Gauss. v.

⇝ we only need tail bounds for $X^T A X$.

Analysis for Gaussian vectors

Theorem

For Gaussian vectors, $\mathbb{P}(|X^T AX - \mathrm{trace}(A)| \geq \varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2}{4\|A\|_F^2 + 4\varepsilon\|A\|_2}\right)$.

Proof ingredients:

1. Moment generating function of random variable $Y := X^T AX - \mathrm{trace}(A)$ can be written explicitly using rotational invariance of Gaussian distribution + properties of χ^2 distributions
2. Use Chernoff bounds to conclude [Book by Boucheron/Lugosi/Massart'2013]

Theorem ([C./Kressner'2021])

For Gaussian vectors and $N \geq \frac{4}{\varepsilon^2}\left(\|A\|_F^2 + \varepsilon\|A\|_2\right)\log\frac{2}{\delta}$ *we have*

$$\mathbb{P}\Big(|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \varepsilon\Big) \leq \delta.$$

Analysis for Rademacher vectors (1)

Main problem wrt Gaussian case: No rotational invariance!

First try: Exploit boundedness. We want

$$\mathbb{P}(|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \varepsilon) \leq \delta.$$

- Hoeffding's inequality: given $Y_1, \dots, Y_N$ zero-mean independent r.v. with $|Y_i| \leq C$ we have $\mathbb{P}(|\sum_{i=1}^N Y_i| \geq \varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2}{2nC^2}\right)$. This gives

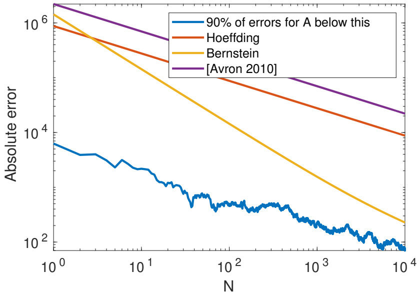
$$N \geq \frac{2n^2}{\varepsilon^2}\|A\|_2^2 \log\frac{2}{\delta};$$

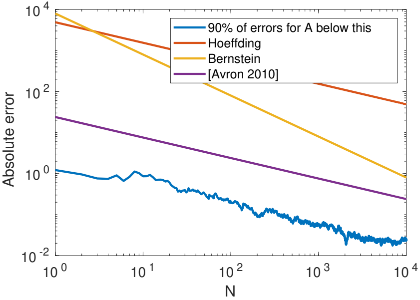
- Bernstein's inequality: if, in addition, $\mathbb{E}[|Y_i|^2] \leq V$ then $\mathbb{P}(|\sum_{i=1}^N Y_i| \geq \varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2}{2NV + \frac{2}{3}C\varepsilon}\right)$. This gives

$$N \geq \frac{4}{\varepsilon^2}\left(\|A\|_F^2 + \frac{\varepsilon}{3}n\|A\|_2\right)\log\frac{2}{\delta}.$$

Existing estimates for indefinite matrices

With Rademacher random vectors





y-axis: Value of ε such that $\mathbb{P}(|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \varepsilon) \leq 0.1$.

These are 2000×2000 matrices obtained by

- Left: `A = randn(n); A = A + A';`
- Right: `d=[(1:n/2).^(-2), -((n/2+1):n).^(-2)];`
  `[Q, ]=qr(randn(n)); A=Q*diag(d)*Q';`

Analysis for Rademacher vectors (2)

Assume A symmetric with zero diagonal.
Hanson-Wright inequality (in [Rudelson/Vershynin'2013]): for $X = (X_1, \dots, X_n)$ with X_i independent and sub-Gaussian[1] random variables with sub-Gaussian norm K we have

$$\mathbb{P}(|X^T AX| \geq \varepsilon) \leq 2\exp\left(-c\min\left(\frac{\varepsilon^2}{K^4\|A\|_F^2}, \frac{\varepsilon}{K^2\|A\|_2}\right)\right).$$

We want explicit constants for Rademacher random vectors!

- [Book by Boucheron/Lugosi/Massart'2013]: $2\exp\left(-\frac{\varepsilon^2}{32\|A\|_F^2+128\varepsilon\|A\|_2}\right)$
- [Book by Foucart/Rauhut'2013]: $2\exp\left(-\min\left\{\frac{3\varepsilon^2}{128\|A\|_F^2}, \frac{\varepsilon}{32\|A\|_2}\right\}\right)$
- [Adamczak'2003]: $2\exp\left(-\frac{\varepsilon^2}{16\|A\|_F^2+16\varepsilon\|A\|_2}\right)$

[1] A random variable Y is sub-Gaussian if there exists $s > 0$ such that $\mathbb{E}[\exp((Y/s)^2)] < +\infty$. The sub-Gaussian norm of Y is defined as $\inf\{s > 0 \mid \mathbb{E}[\exp((Y/s)^2)] \leq 2\}$

Analysis for Rademacher vectors (3)

Theorem ([C./Kressner'2021])

Let A be symmetric with $\mathrm{diag}(A) = 0$. Then

$$\mathbb{P}\left(|X^T AX| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{8\|A\|_F^2 + 8\varepsilon\|A\|_2}\right).$$

Proof ingredients:

1. Bound entropy of random variable $X^T AX$ using logarithmic Sobolev inequalities from [Adamczak'2003]
2. Get a bound on the moment generating function via entropy method / Herbst argument [Boucheron/Lugosi/Massart'2013]
3. Use Chernoff bounds to conclude

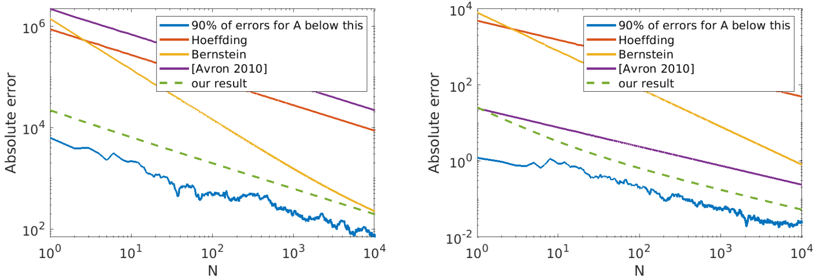
Corollary ([C./Kressner'2021])

Let A be symmetric (indefinite), use Rademacher vectors.
For $N \geq \frac{8}{\varepsilon^2}\left(\|A - \mathrm{diag}(A)\|_F^2 + \varepsilon\|A - \mathrm{diag}(A)\|_2\right)\log\frac{2}{\delta}$ we have

$$\mathbb{P}\Big(|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \varepsilon\Big) \leq \delta.$$

New estimates for indefinite matrices

With Rademacher random vectors



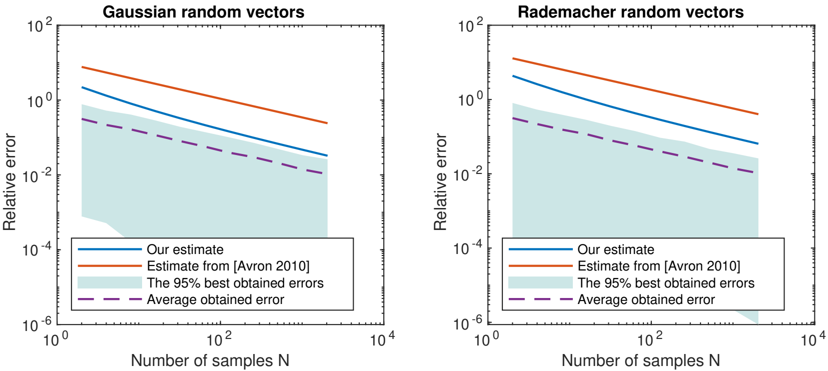
y-axis: Value of ε such that $\mathbb{P}(|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \varepsilon) \leq 0.1$.

These are 2000×2000 matrices obtained by

- Left: `A = randn(n); A = A + A';`
- Right: `d=[(1:n/2).^(-2), -((n/2+1):n).^(-2)];`
  `[Q, ]=qr(randn(n)); A=Q*diag(d)*Q';`

Numerical example: triangle counting

$B =$ adjacency matrix of undirected graph $\rightsquigarrow$ #triangles = $\frac{1}{6}\mathrm{trace}(B^3)$.



Upper border of shaded area = value of ε for which N guarantees accuracy ε with failure probability $\delta = 0.05$.

Matrix from `https://snap.stanford.edu/data/ca-GrQc.html`, $n = 5242$.

Summary

We now have (tight) tail bounds for Hutchinson's trace estimator with indefinite matrices.

$$\mathrm{trace}(A) \approx \mathrm{trace}_N(A) := \frac{1}{N}\sum_{i=1}^{N}(X^{(i)})^T A X^{(i)}.$$

To get $\mathbb{P}(|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \varepsilon) \leq \delta$ we can choose

- $N \geq \frac{8}{\varepsilon^2}\left(\|A - \mathrm{diag}(A)\|_F^2 + \varepsilon\|A - \mathrm{diag}(A)\|_2\right)\log\frac{2}{\delta}$ for Rademacher;
- $N \geq \frac{4}{\varepsilon^2}\left(\|A\|_F^2 + \varepsilon\|A\|_2\right)\log\frac{2}{\delta}$ for Gaussian.

Recall that we are interested in $\log\det(B) = \mathrm{trace}(\log B)$ therefore we need a way to compute $(X^{(i)})^T \log(B) X^{(i)}$.

Approximating the quadratic forms

Quadrature and Lanczos method

G. H. Golub and G. Meurant. Matrices, moments and quadrature with applications. (2010)

Let $B = Q \cdot D \cdot Q^T$ spectral decomposition and let $x \in \mathbb{R}^n$:

$$x^T \log(B) x = (Q^T x)^T \cdot \log(D) \cdot (Q^T x) = \int_{\lambda_{\min}}^{\lambda_{\max}} \log(t) \mathrm{d}\mu(t),$$

where

$$\mathrm{d}\mu(t) = \sum_{i=1}^{n} z_i^2 \delta_{\lambda_i}(t), \qquad z = Q^T x.$$

Gauss quadrature:

$$\text{integral} \approx \sum_{i=1}^{m} w_i \log(\theta_i) =: I_m.$$

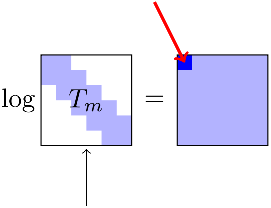
Approximation of quadratic forms via Lanczos method

Theorem ([Golub/Meurant'2010])

Let T_m be matrix obtained after m steps of Lanczos method applied to B with starting vector x. Then $I_m := \sum_{i=1}^{m} w_i \log(\theta_i) = e_1^T \log(T_m) e_1$ where

- *nodes θ_i = eigenvalues of T_m*
- *weights w_i = squares of first elements of normalized eigenvectors of T_m*

For unit vector x,
approximate $x^T \log(B) x$ as



Obtained after m steps of Lanczos

Lanczos algorithm:

Initialize $u_1 \leftarrow x/\|x\|_2$ and $\beta_0 \leftarrow 0$
for $i = 1, \ldots, m$ **do**
  $\alpha_i \leftarrow u_i^T B u_i$
  $r_i \leftarrow B u_i - \alpha_i u_i - \beta_{i-1} u_{i-1}$
  $\beta_i \leftarrow \|r_i\|_2$
  $u_{i+1} \leftarrow r_i / \beta_i$
end for

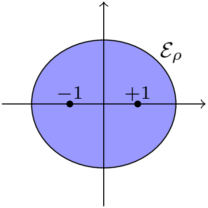
$$T_m \leftarrow \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ & & \beta_{m-1} & \alpha_m \end{bmatrix}$$

Quadrature and polynomial approximation

Theorem

Assume $[\lambda_{\min}, \lambda_{\max}] \subseteq [-1, 1]$. *Let* f *be analytic on Bernstein ellipse* $\mathcal{E}_\rho$, *let* $M_\rho = \max_{\mathcal{E}_\rho} |f(z)|$, *then*

$$|x^T f(A) x - I_m| \leq \frac{4M_\rho}{1 - 1/\rho} \rho^{-2m}.$$



Gauss quadrature is exact for polynomials up to degree $2m - 1$; if we take Chebyshev polynomial P_{2m-1}, thanks to analyticity jth coefficient is bounded by $2M_\rho/\rho^j$.

It also holds for general spectral intervals $[\lambda_{\min}, \lambda_{\max}]$, with a shifted and scaled ellipse (see, e.g., [Ubaru/Chen/Saad'2017, C./Kressner'2021]).

Polynomial approximation of logarithm

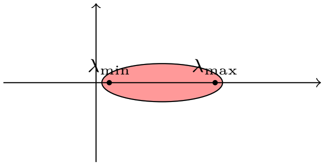
For B SPD with condition number $\kappa(B)$ and $x \neq 0$ we have

$$|x^T \log(B)x - \|x\|_2^2 \cdot e_1^T \log(T_m)e_1| \leq 2\|x\|_2^2(\sqrt{\kappa(B)+1}+1)\log(2\kappa(B))\left(\frac{\sqrt{\kappa(B)+1}-1}{\sqrt{\kappa(B)+1}+1}\right)^{2m}.$$

For $f = \log$:

- Reduce to case $\lambda_{\max} = 1/\lambda_{\min} = \sqrt{\kappa(B)}$ via $\log(\lambda B) = \log(B) + \log\lambda \cdot I$.
- Take Bernstein ellipse with radius

$$\rho = \frac{\sqrt{\kappa(B)+1}-1}{\sqrt{\kappa(B)+1}+1};$$



- Note that maximum of $\log$ is on the real axis.

Combined results and numerical experiments

Putting everything together

$$\text{approximation} := \frac{1}{N}\sum_{i=1}^{N} \|X^{(i)}\|_2^2 \,(\log(T_m))_{1,1} \approx \frac{1}{N}\sum_{i=1}^{N}(X^{(i)})^T \log(B) X^{(i)},$$

where T_m is obtained from Lanczos with starting vector $X^{(i)}$.

Theorem ([C./Kressner'2021])

Suppose that we use Rademacher random vectors and:

(i) *Number of samples*

$$N \geq 32\varepsilon^{-2}\left(\|\log B - \operatorname{diag}(\log B)\|_F^2 + \frac{\varepsilon}{2}\|\log B - \operatorname{diag}(\log B)\|_2\right)\log\frac{2}{\delta};$$

(ii) *Number of Lanczos iterations* $m \geq \frac{\sqrt{\kappa(B)+1}}{4}\log\left(8\varepsilon^{-1} n\sqrt{\kappa(B)}\right)$.

Then

$$\mathbb{P}(|\text{approximation} - \log\det(B)| \geq \varepsilon) \leq \delta.$$

Gaussian random vectors

Theorem ([C./Kressner'2021])

Suppose that we use Gaussian random vectors and :

(i) *Number of samples*

$$N \geq 16\varepsilon^{-2}(\|\log(B)\|_F^2 + \varepsilon\|\log(B)\|_2)\log\frac{4}{\delta};$$

(ii) *Number of Lanczos iterations*

$$m \geq \frac{\sqrt{\kappa(B)+1}}{4}\log\left(4\varepsilon^{-1}n^2(\sqrt{\kappa(B)+1}+1)\log(2\kappa(B))\right).$$

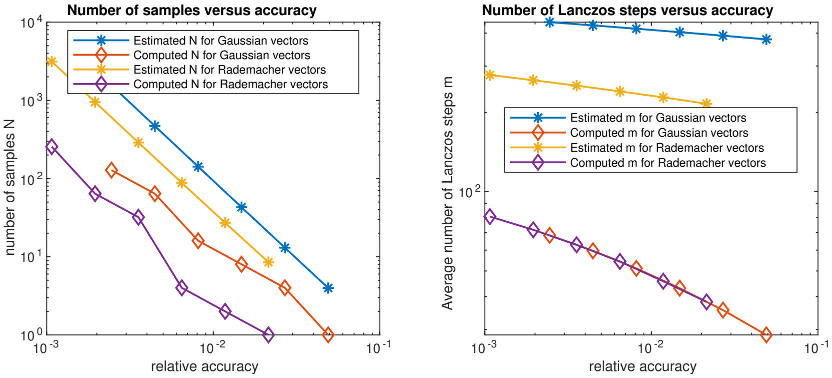
If, additionally, $n \geq 2$ *and* $N \leq \frac{\delta}{2}\exp\left(\frac{n^2}{16}\right)$*, then*

$$\mathbb{P}(|\text{approximation} - \log\det(B)| \geq \varepsilon) \leq \delta.$$

Need to have $\|X\|_2^2$ under control.

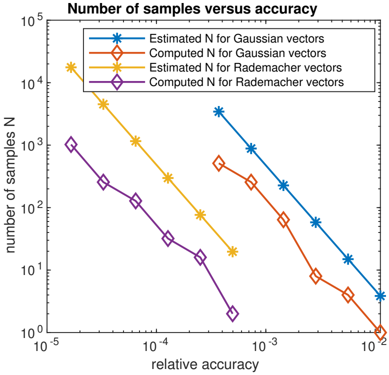
Some numerical experiments

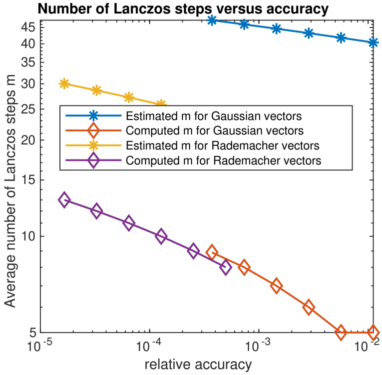
$B \in \mathbb{R}^{6400 \times 6400}$ is a 2D Gaussian kernel from [Meyer/Musco/Musco/Woodruff'2020].



Some numerical experiments

$B \in \mathbb{R}^{102158 \times 102158}$ from SuiteSparse matrix collection (`thermomec_TC`).





Conclusion

Conclusions

Summary:

a) In several applications one wants the trace of a matrix function, e.g. $\log\det(B) = \mathrm{trace}(\log B)$.

b) Cheap approximations of $v^T f(B)v$ can be obtained via Lanczos/quadrature.

c) Hutchinson's trace estimator $\mathrm{trace}_N(A) = \frac{1}{N}\sum_{i=1}^{N} X_i^T A X_i$ can be used for estimating the trace of a matrix which is available through matrix-vector multiplications
$\rightsquigarrow$ Presented improved convergence analysis of Hutchinson estimator.

Alice Cortinovis and Daniel Kressner, On randomized trace estimates for indefinite matrices with an application to determinants. Foundations of Computational Mathematics, 2021.