Randomized Matrix Computations Lecture 3

Daniel Kressner
Chair for Numerical Algorithms and HPC
Institute of Mathematics, EPFL
daniel.kressner@epfl.ch

Trace Estimation

- Motivation
- Analysis

Literature:

Tropp'2023 Joel A. Tropp. *Probability Theory & Computational Mathematics*, Lecture notes, Caltech, 2023.

MT'2020 Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. Acta Numerica'2020.

pdf available on Moodle

Computing determinants

For $n \times n$ matrix A consider determinant

$$\det(A) := \sum_{\text{permutation } \sigma} \operatorname{sign}(\sigma) \cdot a_{1,\sigma(1)} \cdot a_{2,\sigma(2)} \cdots a_{n,\sigma(n)}.$$

How would you compute/approximate $\log\det(A)$ numerically?

Computing determinants

For $n \times n$ matrix A consider determinant

$$\det(A) := \sum_{\text{permutation } \sigma} \operatorname{sign}(\sigma) \cdot a_{1,\sigma(1)} \cdot a_{2,\sigma(2)} \cdots a_{n,\sigma(n)}.$$

How would you compute/approximate $\log\det(A)$ numerically?

- For $n = O(10^3)$: Via LU factorization $PA = LU$.
  $$\det(A) = \det(L)\det(U)/\det(P)$$
- For large n and sparse A: Via sparse LU factorization.
- For large n and A can only be accessed through matvec products: Assuming that A is symm pos def $\rightsquigarrow$
  Stochastic trace estimator applied to $\operatorname{trace}\log(A) = \log\det(A)$.

EFY: Verify $\operatorname{trace}\log(A) = \log\det(A)$. What if A is symmetric but not pos def?

Goal of trace estimation:

Compute (an approximation of) $\mathrm{trace}(A)$
for large-scale symmetric matrix $A \in \mathbb{R}^{n \times n}$.

Only matvec products with A are available.

Motivation

Example 1: Trace of matrix functions / Determinant

Matrix functions: For symmetric A, given a spectral decomposition $A = Q \cdot \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \cdot Q^\top$, the matrix function $f(A)$ is defined as

$$f(A) = Q \cdot \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n)) \cdot Q^\top.$$

Computing $f(A)v$ is faster than computing $f(A)$!

- Trace(A^{-1}) (Uncertainty quantification [Kalantzis/Bekas/Curioni/Gallopoulos'2013], Lattice quantum chromodynamics [Wu et al.'2016])
- Network analysis ($\exp(A)$, Estrada index)
- Determinant of sym. positive definite A via $\log\det(A) = \mathrm{trace}(\log A)$

Example 2: Frobenius norm estimation $\|B\|_F^2 = \mathrm{trace}(B^\top B)$ and other Schatten-p norms [Gratton/Titley-Peloquin'2018, Dudley/Saibaba/Alexanderian'2021]

Applications of log determinant

- 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

Motivation

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 kernel $k(\cdot,\cdot)$.

Goal: Train Gaussian process on *lots* 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

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

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

Cholesky decomposition of K too expensive $\rightsquigarrow$ stochastic 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)

Randomized trace estimation

We call a real random vector X *isotropic* if $\mathbb{E}[XX^\top] = I$. Most common choices:

- Rademacher vectors (± 1 entries);
- Gaussian vectors ($X \sim \mathcal{N}(0, I_n)$)

Theorem [Girard-Hutchinson trace estimation] For an isotropic random vector X it holds that

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

Proof. $\mathbb{E}[X^\top AX] = \sum_{i,j} \mathbb{E}[X_i X_j] a_{ij} = \sum_i a_{ii} = \mathrm{trace}(A)$. ◇

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

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

Randomized trace estimation: Example

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

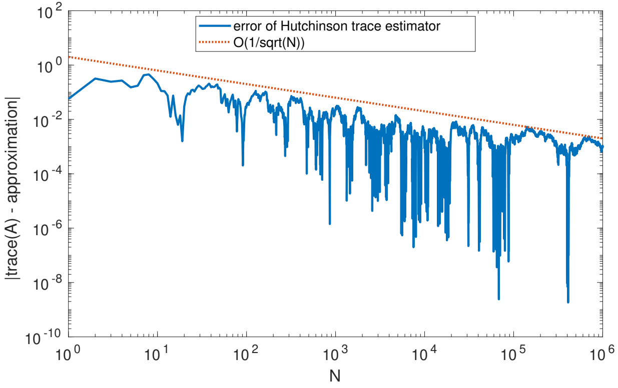


Figure: Behavior of $|\mathrm{trace}(A) - \mathrm{trace}_N(A)|$ when increasing N (# probe vectors).

Chebyshev

Aim to study convergence of $\mathrm{trace}_N(A)$ as N increases.

$$\mathrm{Var}\big[\mathrm{trace}_N(A)\big] = \frac{1}{N^2}\sum \mathrm{Var}\big[(X^{(i)})^\top A X^{(i)}\big] = \frac{1}{N}\mathrm{Var}\big[X^\top A X\big]$$

Variance depends on choice of X (EFY: Verify these identities):

Rademacher $\rightsquigarrow$ $\mathrm{Var}\big(X^\top A X\big) = 2\|A - \mathrm{diag}(A)\|_F^2$

Gaussian $\rightsquigarrow$ $\mathrm{Var}\big(X^\top A X\big) = 2\|A\|_F^2$.

Chebyshev's inequality gives

$$\mathbb{P}\big\{|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \epsilon\big\} \leq \mathrm{Var}\big(X^\top A X\big)\epsilon^{-2}N^{-1} \leq 2\|A\|_F^2\epsilon^{-2}N^{-1}.$$

If A is spd, we have $\|A\|_F^2 \leq \|A\|_2 \mathrm{trace}(A)$ and can get a relative error bound:

$$\mathbb{P}\big\{|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \epsilon \cdot \mathrm{trace}(A)\big\} \leq 2\rho(A)^{-1}\epsilon^{-2}N^{-1},$$

where $\rho(A) := \mathrm{trace}(A)/\|A\|_2$ is called intrinsic dimension of A.

Chebyshev: $N = O\big(\frac{1}{\delta\rho(A)\epsilon^2}\big)$ needed for rel acc ϵ^{-2} with prob $1-\delta$.

Improved bounds: Rademacher

Chebyshev only takes variance into account
$\rightsquigarrow$ suboptimal dependence on δ.

Now assume that X is Rademacher.

EFY: Show that Hoeffding's inequality leads to

$$\mathbb{P}\left\{|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \epsilon\right\} \leq \exp\left(-\frac{2\epsilon^2 N}{\|A\|_2^2 n^2}\right).$$

Conclude that for spd A, $N = O\left(\frac{\log \delta^{-1} n^2}{\rho(A)^2 \epsilon^2}\right)$ needed for rel acc ϵ^{-2} with prob $1 - \delta$. Improved depend. on δ but unpleasant depend. on n.

EFY: Show that Bernstein's inequality leads for traceless A to

$$\mathbb{P}\left\{|\mathrm{trace}_N(A) - \mathrm{trace}(A)| \geq \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2 N}{4\|A\|_F^2 + 2/3 n\epsilon \|A\|_2}\right).$$

How can one incorporate a nonzero trace?
Maintains improved depend. on δ and less crazy dependence on n.

Improved bounds: Rademacher

SOTA: Use Hanson-Wright inequality.

Theorem [Cortinovis/DK'2021] Let A be symmetric with $\mathrm{diag}(A) = 0$. Then

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

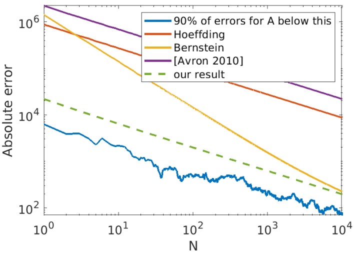
Embedding trick: write $\mathrm{trace}_N(A) = \frac{1}{N}\sum_{i=1}^{N}(X^{(i)})^\top A X^{(i)}$

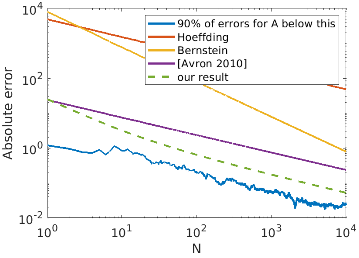
$$= \begin{bmatrix} X^{(1)\top} \cdots X^{(N)\top} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{N}A & & & & \\ & \frac{1}{N}A & & & \\ & & \ddots & & \\ & & & \frac{1}{N}A & \\ & & & & \frac{1}{N}A \end{bmatrix} \cdot \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(N)} \end{bmatrix} \longleftarrow \text{Rademacher}$$

Corollary. 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

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

Comparison of estimates





2000×2000 matrices:

- 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';`

Blue curve: for each value of *N*, compute 20 trace estimates, throw away the worst 10%, and plot the largest error of the remaining estimates.

Improved bounds: Gaussian

Now, suppose that X is Gaussian and that A is spd with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

By unitary invariance

$$X^\top AX - \mathrm{trace}(A) = \lambda_1(Z_1^2 - 1) + \cdots + \lambda_n(Z_n^2 - 1), \quad Z_k \sim N(0,1) \text{ i.i.d.}$$

We already know that

$$\log \mathbb{E}\big[\exp\big(\theta(Z^2-1)\big)\big] = \log \frac{e^{-\theta}}{\sqrt{1-2\theta}} = -\frac{1}{2}\log(1-2\theta) + 2\theta \leq \frac{\theta^2}{1-2\theta}$$

for $\theta < 1/2$. Could continue with sub-exponential properties, but would give slightly sub-optimal bounds. Instead, one concludes that

$$\log \mathbb{E}\big[X^\top AX - \mathrm{trace}(A)\big] \leq \frac{\theta^2\|A\|_F^2}{1-2\theta\|A\|_2}, \quad \theta < 1/(2\|A\|_2).$$

Chernoff gives Hanson-Wright inequality

$$\mathbb{P}\left(|X^\top BX| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{4\|A\|_F^2 + 4\varepsilon\|A\|_2}\right).$$

Conclude with embedding trick. (Vershynin: Technique works for all sub-Gaussian vectors!).

The curse of Monte Carlo

For Rademacher/Gaussian, we were able to improve dependence on δ and properties of A, but *not* on ϵ^{-2}!

The central limit theorem tells us that

$$\sqrt{N}\big(\mathrm{trace}_N(A) - \mathrm{trace}(A)\big) \to N(0, \mathrm{Var}\big[X^\top AX\big]) \quad \text{(in distribution)}$$

For sufficiently large N, we expect the error to behave like $\sqrt{N} \cdot Z \sim N(0, \mathrm{Var}\big[X^\top AX\big])$. We already know that

$$\mathbb{P}\{\sqrt{N}Z \geq \epsilon\} \leq \exp(-\epsilon^2/(4\|A\|_F^2))$$

and, hence,

$$\mathbb{P}\{Z \geq \epsilon\} \leq \exp(-\epsilon^2/(4N\|A\|_F^2)).$$

It follows that $N \sim \epsilon^{-2}$ is needed to attain constant failure probability.

In order to improve dependence on ϵ^{-2}, need to reduce variance!