

CSE 392/CS 395T/M 397C: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2025

Lecture 14: Stochastic Trace Estimation

- 1 Implicit trace estimation
- 2 Stochastic trace estimation
- 3 Hutch++

Matrix Trace

- Given a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ our goal is to compute the trace:

$$\text{Tr}(\mathbf{A}) = \sum_{i=1}^d \mathbf{A}_{ii}.$$

- In terms of the eigenvalues, if $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{\Lambda} = \text{diag}[\lambda_1, \dots, \lambda_d]$, we know:

$$\text{Tr}(\mathbf{A}) = \sum_{i=1}^d \lambda_i.$$

- In many situations, access to \mathbf{A} available only implicitly through a *matrix-vector multiplication oracle*. Estimate the trace implicitly (also called matrix-free)?

Spectral Sums

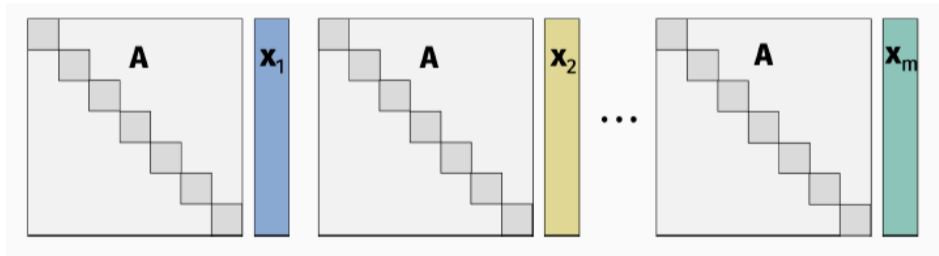
Given a symmetric positive semidefinite (PSD) matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ with eigen-decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and eigenvalues $\{\lambda_i\}_{i=1}^d$, and desired function $f(\cdot)$, compute the *trace of the matrix function* $f(\mathbf{A}) = \mathbf{U}f(\mathbf{\Lambda})\mathbf{U}^T$, i.e.,

$$\text{Tr}(f(\mathbf{A})) = \sum_{i=1}^d f(\lambda_i).$$

- *Popular examples:* log-determinant ($\log(x)$), numerical rank (step function), spectral density $\delta(x - \lambda_i)$, Schatten p -norms ($x^{p/2}$), von Neumann Entropy ($x \log(x)$), Estrada index ($\exp(x)$), trace of matrix inverse ($\frac{1}{x}$).
- *Applications:* machine learning, graph signal processing, quantum algorithms, scientific computing, statistics, computational biology and physics.
- *Naive approaches :* Eigenvalue decomposition, Cholesky Decomposition, singular value decomposition (SVD).
Cost: $O(d^3)$ or [Theory: $O(d^\omega)$ and $\omega = 2.373$].

Implicit trace estimation

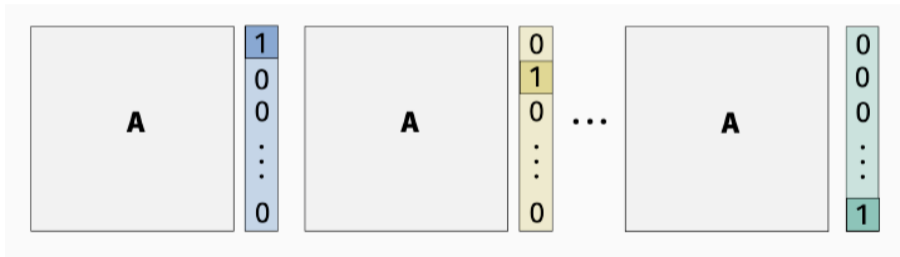
- Access to \mathbf{A} implicitly through a *matrix-vector multiplication oracle*.
- Typically useful when \mathbf{A} is not stored explicitly, but we have an efficient algorithm for multiplying \mathbf{A} by a vector.
- Matrix-vector products (Matvecs) cost $O(\text{nnz}(\mathbf{A}))$.
- *Examples:* Hessians in optimization, matrix functions as polynomials, structured matrices, etc.



How many matvecs $\mathbf{A}\mathbf{x}_1, \dots, \mathbf{A}\mathbf{x}_m$ are needed to estimate the trace?

A naive approach

- Set $\mathbf{x}_l = \mathbf{e}_l$ for $l = 1, \dots, d$.
- Return $\text{Tr}(\mathbf{A}) = \sum_{l=1}^d \mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l$.
- Total computational cost $O(\text{nnz}(\mathbf{A})d)$.



Exact solution, but required d matvecs. Can we approximately estimate the trace with $\ll d$ matvecs?

Stochastic Trace Estimation

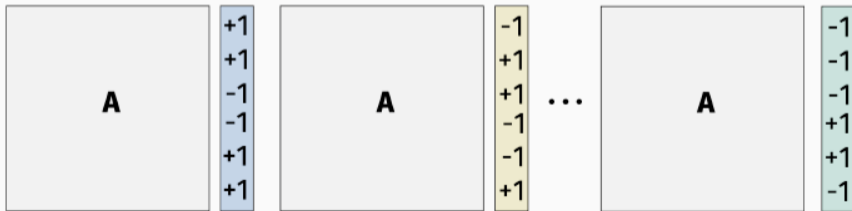
Hutchinson's stochastic trace estimator

- Hutchinson [Hutchinson, 1990] proposed a method for implicit matrix trace estimation:

$$\text{Tr}(A) \approx \frac{1}{m} \sum_{l=1}^m \mathbf{x}_l^\top A \mathbf{x}_l, \quad (1)$$

where $\mathbf{x}_l, l = 1, \dots, m$, are random vectors with i.i.d. random $\{+1, -1\}$ entries.

- *Randomized method*: Simple, powerful, and widely used method for trace estimation.
- Theoretical analyses were presented in [Avron, Toledo 2011], [Roosta, Ascher 2015].



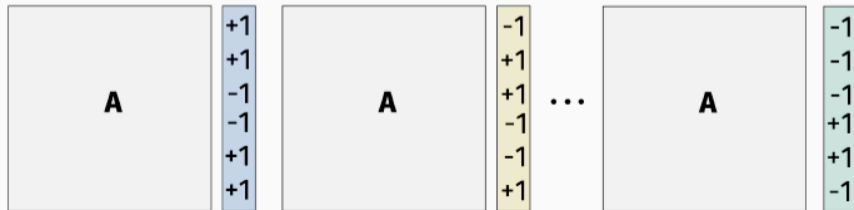
Stochastic trace estimator

Theorem

Let \mathbf{A} be an $d \times d$ symmetric positive semidefinite (PSD) matrix and $\mathbf{x}_l, l = 1, \dots, m$ be random starting vectors with Radamacher distribution. Then, for $\tilde{\text{Tr}}_m(\mathbf{A}) = \frac{1}{m} \sum_{l=1}^m \mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l$, with $m = O\left(\frac{\log(1/\eta)}{\epsilon^2}\right)$, we have

$$\Pr \left[\left| \tilde{\text{Tr}}_m(\mathbf{A}) - \text{Tr}(\mathbf{A}) \right| \leq \epsilon |\text{Tr}(\mathbf{A})| \right] \geq 1 - \eta.$$

Radamacher distribution: vectors with $\{\pm 1\}$ entries with equal probabilities.



Expected Value Analysis

Hutchinson's Estimator:

- Draw $\mathbf{x}_l, l = 1, \dots, m$, vectors with i.i.d. random $\{+1, -1\}$ entries.
- Return $\tilde{\text{Tr}}_m(\mathbf{A}) = \frac{1}{m} \sum_{l=1}^m \mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l$ as approximation to $\text{Tr}(\mathbf{A})$.

Expected value analysis:

For a single random ± 1 vector \mathbf{x} , we have

$$\mathbb{E}[\tilde{\text{Tr}}_m(\mathbf{A})] = \mathbb{E}[\mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l] = \mathbb{E} \sum_{i=1}^d \sum_{j=1}^d x_i x_j \mathbf{A}_{ij} = \sum_{i=1}^d \sum_{j=1}^d \mathbb{E}[x_i x_j \mathbf{A}_{ij}] = \sum_{i=1}^d \mathbf{A}_{ii}$$

So the estimator is correct in expectation:

$$\mathbb{E}[\tilde{\text{Tr}}_m(\mathbf{A})] = \text{Tr}(\mathbf{A}).$$

Variance Analysis

Hutchinson's Estimator:

- Draw $\mathbf{x}_l, l = 1, \dots, m$, vectors with i.i.d. random $\{+1, -1\}$ entries.
- Return $\tilde{\text{Tr}}_m(\mathbf{A}) = \frac{1}{m} \sum_{l=1}^m \mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l$ as approximation to $\text{Tr}(\mathbf{A})$.

Variance analysis:

$$\begin{aligned}\text{Var}[\tilde{\text{Tr}}_m(\mathbf{A})] &= \frac{1}{m} \text{Var}[\mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l] = \frac{1}{m} \left[\mathbb{E}[(\mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l)^2] - \text{Tr}(\mathbf{A})^2 \right] \\ \mathbb{E}[(\mathbf{x}_l^\top \mathbf{A} \mathbf{x}_l)^2] &= \mathbb{E} \left[\left(\sum_{i,j} x_i x_j \mathbf{A}_{ij} \right) \left(\sum_{i',j'} x_{i'} x_{j'} \mathbf{A}_{i'j'} \right) \right] \\ &= 2 \sum_{i \neq j} \mathbf{A}_{ij}^2 + \sum_{i \neq j} \mathbf{A}_{ii} \mathbf{A}_{jj} + \sum_i \mathbf{A}_{ii}^2\end{aligned}$$

We used that $x_i x_j$ and $x_{i'} x_{j'}$ are pairwise independent. Therefore,

$$\text{Var}[\tilde{\text{Tr}}_m(\mathbf{A})] = \frac{2}{m} \sum_{i \neq j} \mathbf{A}_{ij}^2 \leq \frac{2}{m} \|\mathbf{A}\|_F^2.$$

Chebyshev's inequality : $\Pr(|X - \mathbb{E}[X]| \geq \tau) \leq \frac{\text{Var}(X)}{\tau^2}$.

We have $\mathbb{E}[\tilde{\text{Tr}}_m(\mathbf{A})] = \text{Tr}(\mathbf{A})$ and $\text{Var}[\tilde{\text{Tr}}_m(\mathbf{A})] \leq \frac{2}{m} \|\mathbf{A}\|_F^2$. Choosing $\tau = \epsilon \cdot \text{Tr}(\mathbf{A})$:

$$\begin{aligned} \Pr\left(\left|\tilde{\text{Tr}}_m(\mathbf{A}) - \text{Tr}(\mathbf{A})\right| \geq \epsilon \cdot \text{Tr}(\mathbf{A})\right) &\leq \frac{\text{Var}(\tilde{\text{Tr}}_m(\mathbf{A}))}{(\epsilon \cdot \text{Tr}(\mathbf{A}))^2} \\ &\leq \frac{2}{m} \frac{\|\mathbf{A}\|_F^2}{(\epsilon \cdot \text{Tr}(\mathbf{A}))^2} = \frac{2}{m\epsilon^2}. \end{aligned}$$

For probability η , we can select $m \geq \frac{2}{\eta\epsilon^2}$.

Can improve this to $m = O\left(\frac{\log(1/\eta)}{\epsilon^2}\right)$, using *Hanson-Wright inequality*.

Improved Analysis

Hanson-Wright inequality [Hanson & Wright, 1971] : Given a symmetric matrix \mathbf{A} and random vector \mathbf{x} with i.i.d sub-Gaussian entries, with constant sub-Gaussian parameter C , we have for $t \geq 0$:

$$\Pr (|\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x}]| \geq t) \leq 2 \exp \left(-c \cdot \min \left(\frac{t^2}{\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|_2} \right) \right),$$

for some universal constant $c > 0$ that only depending on C .

Hanson-Wright inequality [Hanson & Wright, 1971] : Given a symmetric matrix \mathbf{A} and random vector \mathbf{x} with i.i.d sub-Gaussian entries, with constant sub-Gaussian parameter C , we have for $t \geq 0$:

$$\Pr(|\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x}]| \geq t) \leq 2 \exp\left(-c \cdot \min\left(\frac{t^2}{\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|_2}\right)\right),$$

for some universal constant $c > 0$ that only depending on C .

Markov's inequality :

$$\Pr(|X - \mathbb{E}[X]| \geq \tau) \leq \frac{\mathbb{E}[X^q]}{\tau^q}.$$

Choose $\tau = (2\epsilon - \epsilon^2) \cdot \text{Tr}(\mathbf{A})$ and $q = \log(1/\eta)$, then with some work we get the theorem with $m = O\left(\frac{\log(1/\eta)}{\epsilon^2}\right)$.

Alternatively, can also use the Markov's inequality (the exponential version) and some recent results, see [Roosta, Ascher 2015].

Further Reading:

- *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix.* by H. Avron and S. Toledo.
- *Improved bounds on sample size for implicit matrix trace estimators* by Roosta-Khorasani and Uri Ascher.

Exercise:

- Would the proof using the Chebyshev inequality work if \mathbf{x}_l 's are drawn from i.i.d Gaussian distribution $\mathcal{N}(0, 1)$? What are the expectation and the variance of the estimate? (Hint: Note that $\mathbf{y}_l = \mathbf{U}\mathbf{x}_l$ are also Gaussian for unitary \mathbf{U} . χ^2 -distribution.)

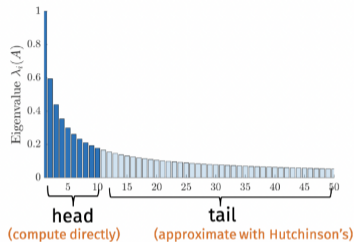
Hutch++

Hutch++ : Improved trace estimator

- Hutchinson's estimator is powerful, and gives a good rate of convergence. But requires $m = O(1/\epsilon^2)$ random vectors and matvecs.
- Recent results by Meyer et al., 2021, showed we can improve this to $m = O(1/\epsilon)$ matvecs.
- *Idea of Hutch++* - Matrices might have decaying eigenvalues. Trace of a low rank approximation of the matrix is a good approximation to the matrix trace.
- Split the trace (spectrum) as sum of trace of top k eigenvalues and bottom $n - k$ eigenvalues.

$$\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{A}_k) + \text{Tr}(\mathbf{A} - \mathbf{A}_k).$$

Meyer, Raphael A., et al. "Hutch++: Optimal stochastic trace estimation." Symposium on Simplicity in Algorithms (SOSA). Society for Industrial and Applied Mathematics, 2021.



Explicitly estimate the top few eigenvalues of \mathbf{A} . Use Hutchinson's for the rest.

- Find a good rank- k approximation $\tilde{\mathbf{A}}_k$.
- Observe $\text{Tr}(\mathbf{A}) = \text{Tr}(\tilde{\mathbf{A}}_k) + \text{Tr}(\mathbf{A} - \tilde{\mathbf{A}}_k)$.
- Compute $\text{Tr}(\tilde{\mathbf{A}}_k)$ exactly.
- Return $\text{Hutch}++(\mathbf{A}) = \text{Tr}(\tilde{\mathbf{A}}_k) + \tilde{\text{Tr}}_m(\mathbf{A} - \tilde{\mathbf{A}}_k)$.

If $k = m = O(1/\epsilon)$, then $|\text{Hutch}++(\mathbf{A}) - \text{Tr}(\mathbf{A})| \leq \epsilon \text{Tr}(\mathbf{A})$.

Good low rank approximation

Let \mathbf{A}_k be the best rank- k approximation of \mathbf{A} .

Lemma (Woo14)

Let $\mathbf{S} \in \mathbb{R}^{d \times m}$ have i.i.d. random entries from $\mathcal{N}(0, 1)$, $\mathbf{Q} = \text{orth}(\mathbf{A}\mathbf{S})$ and $\tilde{\mathbf{A}}_k = \mathbf{Q}\mathbf{Q}^T \mathbf{A}$. Then if $m = O(k + \log(1/\delta))$, with probability $1 - \delta$,

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_F \leq 2\|\mathbf{A} - \mathbf{A}_k\|_F.$$

We can compute $\text{Tr}(\tilde{\mathbf{A}}_k)$ with $2m$ matvecs with \mathbf{A} and $O(mn)$ space:

$$\text{Tr}(\tilde{\mathbf{A}}_k) = \text{Tr}(\mathbf{Q}\mathbf{Q}^T \mathbf{A}) = \text{Tr}(\mathbf{Q}^T (\mathbf{A}\mathbf{Q}))$$

Hutch++ Algorithm

- **Input:** Number of matvecs m and input matrix \mathbf{A} .
- Sample $\mathbf{S} \in \mathbb{R}^{d \times m/3}$ and $\mathbf{G} \in \mathbb{R}^{d \times m/3}$ with i.i.d. entries from $\mathcal{N}(0, 1)$.
- Compute $\mathbf{Q} = \text{orth}(\mathbf{AS})$.
- Return $\text{Hutch++}(\mathbf{A}) = \text{Tr}(\mathbf{Q}^T(\mathbf{AQ})) + \frac{3}{m} \text{Tr}(\mathbf{G}^T(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{G})$.

We have the following result:

Lemma

Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a PSD matrix and \mathbf{A}_k be its best rank- k approximation. Then,

$$\|\mathbf{A} - \mathbf{A}_k\|_F \leq \frac{1}{2\sqrt{k}} \text{Tr}(\mathbf{A})$$

Hutch++ mean and variance

Theorem

Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a PSD matrix, for fixed k and m , construct $\mathbf{Q} \in \mathbb{R}^{d \times m}$ as before. Let $Hutch++(\mathbf{A}) = \text{Tr}(\mathbf{Q}^T(\mathbf{A}\mathbf{Q})) + \tilde{\text{Tr}}_m((\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A})$. Then,

$$\begin{aligned}\mathbb{E}[Hutch++(\mathbf{A})] &= \text{Tr}(\mathbf{A}) \\ \text{Var}[Hutch++(\mathbf{A})] &\leq \frac{1}{km} \text{Tr}^2(\mathbf{A})\end{aligned}$$

For the mean, we have $\mathbb{E}[Hutch++(\mathbf{A})] = \mathbb{E}[\text{Tr}(\mathbf{Q}^T(\mathbf{A}\mathbf{Q}))] + \mathbb{E}[\mathbb{E}[\tilde{\text{Tr}}_m((\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A})|\mathbf{Q}]]$.

For variance, we use the Conditional Variance Formula,

$$\text{Var}[Hutch++(\mathbf{A})] = \mathbb{E}[\text{Var}[Hutch++(\mathbf{A})|\mathbf{Q}]] + \text{Var}[\mathbb{E}[Hutch++(\mathbf{A})|\mathbf{Q}]].$$

Can show $\text{Var}[\mathbb{E}[Hutch++(\mathbf{A})|\mathbf{Q}]] = 0$.

Further Reading:

- Meyer, Raphael A., et al. “Hutch++: Optimal stochastic trace estimation.” Symposium on Simplicity in Algorithms (SOSA). Society for Industrial and Applied Mathematics, 2021.
- <https://ram900.com/hutchplusplus/>

Hints for Problem 4 in HW2: Write $\|\mathbf{A} - \mathbf{A}_k\|_F$ and $\text{Tr}(\mathbf{A})$ in terms of eigenvalues. Next, use the Holder’s inequality $\|v\|_2^2 \leq \|v\|_1 \|v\|_\infty$. Note the function $\gamma \rightarrow \frac{\sqrt{a\gamma}}{b+\gamma}$ is maximized at $\gamma = b$, so $\frac{\sqrt{a\gamma}}{b+\gamma} \leq \frac{\sqrt{ab}}{2b}$. Choose appropriate a and b to bound the ratio $\frac{\|\mathbf{A} - \mathbf{A}_k\|_F}{\text{Tr}(\mathbf{A})}$.

Matlab Demo