CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin

Spring 2024

Lecture 14: Stochastic Trace Estimation

Outline

Matrix Trace

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

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

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

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

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

Spectral Sums

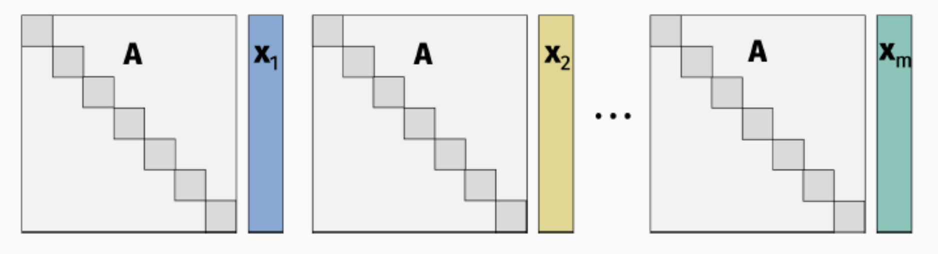
Given a symmetric positive semidefinite (PSD) matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ with eigen-decompostion $\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^T$ and eigenvalues $\{\lambda_i\}_{i=1}^d$, and desired function $f(\cdot)$, compute the *trace of the matrix function* $f(\boldsymbol{A})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

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

- *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

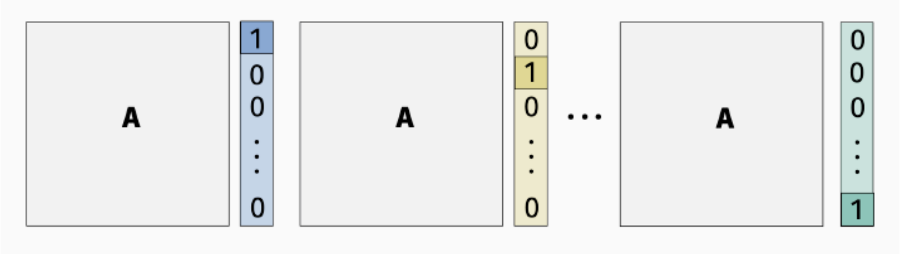
- Access to $\boldsymbol{A}$ implicitly through a *matrix-vector multiplication oracle*.
- Typically useful when $\boldsymbol{A}$ is not stored explicitly, but we have an efficient algorithm for multiplying $\boldsymbol{A}$ by a vector.
- Matrix-vector products (Matvecs) cost $O(\text{nnz}(\boldsymbol{A}))$.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



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

A naive approach

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



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

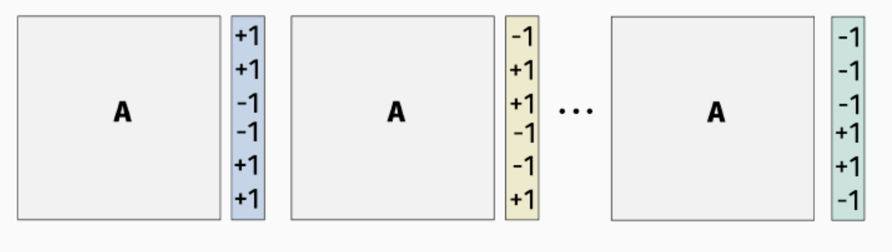
Stochastic Trace Estimation

Hutchinson's stochastic trace estimator

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

$$\mathrm{Tr}(A) \approx \frac{1}{m} \sum_{l=1}^{m} \boldsymbol{x}_l^\top A \boldsymbol{x}_l, \tag{1}$$

  where $\boldsymbol{x}_l, l = 1, \ldots, m$, 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].



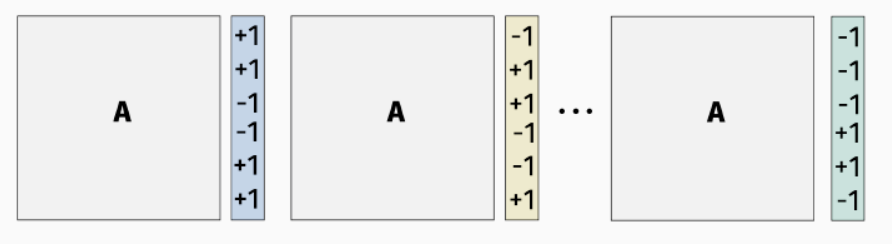
Stochastic trace estimator

Theorem

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

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

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



Expected Value Analysis

Hutchinson's Estimator:

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

Expected value analysis:
For a single random ± 1 vector $\boldsymbol{x}$, we have

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

So the estimator is correct in expectation:

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

Variance Analysis

Hutchinson's Estimator:

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

Variance analysis:

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

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

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

Analysis

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

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

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

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 $\boldsymbol{A}$ and random vector $\boldsymbol{x}$ with i.i.d sub-Gaussian entries, with constant sub-Gaussian parameter C, we have for $t \geq 0$:

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

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

Improved Analysis

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

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

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 $\tau = (2\epsilon - \epsilon^2) \cdot \mathrm{Tr}(\boldsymbol{A})$ 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 $\boldsymbol{x}_l$'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 $\boldsymbol{y}_l = \boldsymbol{U}\boldsymbol{x}_l$ are also Gaussian for unitary $\boldsymbol{U}$. χ^2-distribution.)

Hutch++

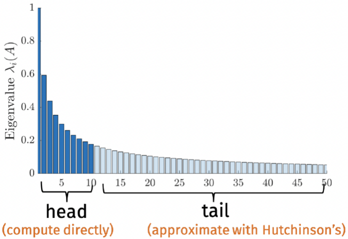
Hutch++ : Improved trace estimator

- Hutchinson's estimator is powerful, and gives a nice 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.

$$\mathrm{Tr}(\boldsymbol{A}) = \mathrm{Tr}(\boldsymbol{A}_k) + \mathrm{Tr}(\boldsymbol{A} - \boldsymbol{A}_k).$$

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

Hutch++



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

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

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

Good low rank approximation

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

Lemma (Woo14)

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

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

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

$$\mathrm{Tr}(\tilde{\boldsymbol{A}}_k) = \mathrm{Tr}(\boldsymbol{QQ}^T\boldsymbol{A}) = \mathrm{Tr}(\boldsymbol{Q}^T(\boldsymbol{AQ}))$$

Hutch++ Algorithm

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

We have the following result:

Lemma

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

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

Hutch++ mean and variance

Theorem

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

$$\mathbb{E}[Hutch++(\boldsymbol{A})] = \mathrm{Tr}(\boldsymbol{A})$$

$$\mathrm{Var}[Hutch++(\boldsymbol{A})] \leq \frac{1}{km}\mathrm{Tr}^2(\boldsymbol{A})$$

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

For variance, we use the Conditional Variance Formula,

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

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

Exercise

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.hosting.nyu.edu/hutchplusplus/`

Hints for Problem 4 in HW2: Write $\|\boldsymbol{A} - \boldsymbol{A}_k\|_F$ and $\mathrm{Tr}(\boldsymbol{A})$ in terms of eigenvalues. Next, use the Holder's inequality $\|v\|_2^2 \leq \|v\|_1 \|v\|_\infty$. Note the function $\gamma \to \frac{\sqrt{a\gamma}}{b+\gamma}$ 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 $\frac{\|\boldsymbol{A}-\boldsymbol{A}_k\|_F}{\mathrm{Tr}(\boldsymbol{A})}$.

Matlab Demo