| | |
|---|---|
| **CSE 392: Matrix and Tensor Algorithms for Data** | Spring 2024 |
| Lecture 14 — Mar 4 | |
| *Instructor: Shashanka Ubaru* | *Scribe: Jeil Oh* |

# Stochastic Trace Estimation

## Matrix Trace

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

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

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

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

- In many situations, access to $A$ available only implicitly through a *matrix-vector multiplication oracle.*

## Spectral Sums

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

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

- **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 $\left(\frac{1}{x}\right)$.
- **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 $A$ implicitly through a *matrix-vector multiplication oracle.*
- Typically useful when $A$ is not stored explicitly, but we have an efficient algorithm for multiplying $A$ by a vector.
- Matrix-vector products (*Matvecs*) cost $O(\text{nnz}(A))$.
- *Examples*: Hessians in optimization, matrix functions as polynomials, structured matrices, etc.

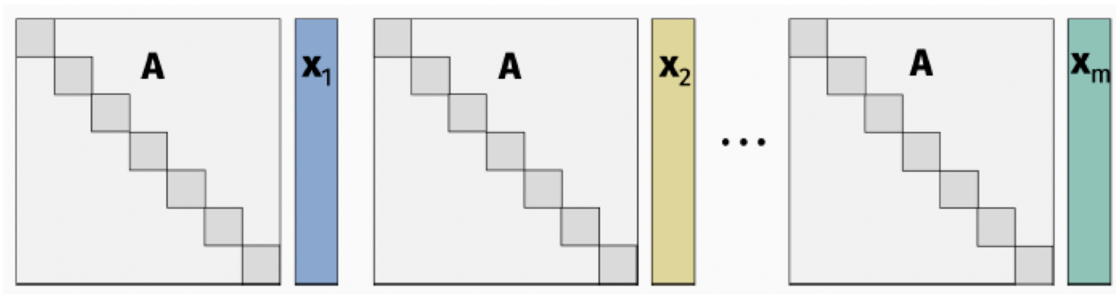


Figure 1: How many matvecs $A\mathbf{x}_1, \ldots, A\mathbf{x}_m$ are needed to estimate the trace?

## A naive approach

- Set $x_l = e_l$ for $l = 1, \ldots, d$.
- Return $\text{Tr}(A) = \sum_{l=1}^{d} x_l^T A x_l$.
- Total computational cost $O(\text{nnz}(A)d)$.

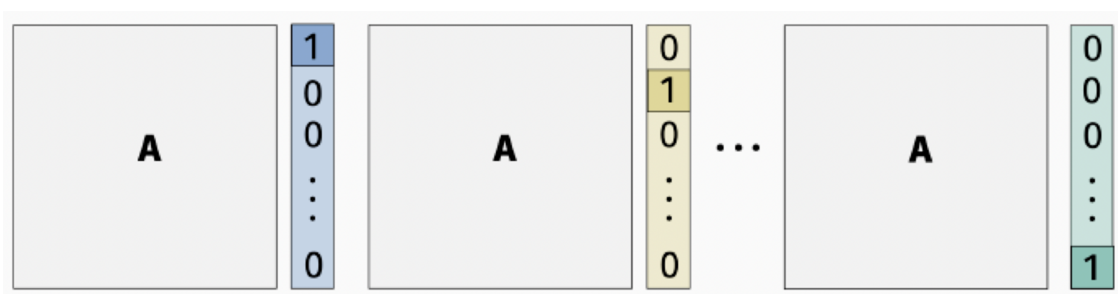


Figure 2: Exact solution, but required $d$ matvecs. Can we approximately estimate the trace with $\ll d$ matvecs?

## Hutchinson's Stochastic Trace Estimator

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

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

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

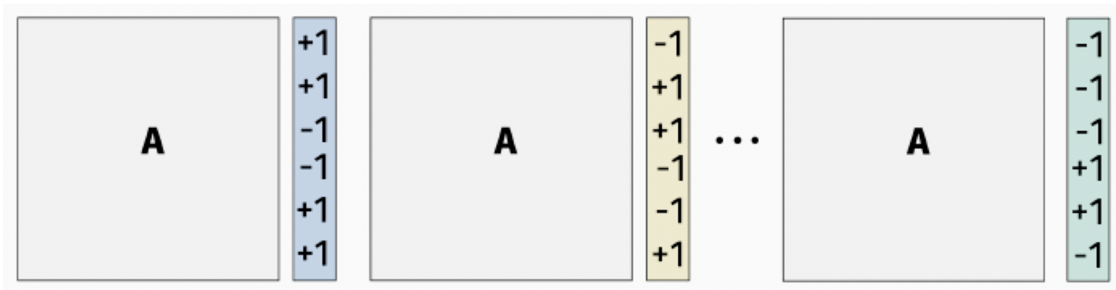


Figure 3: Radamacher distribution: vectors with $\pm 1$ entries with equal probabilities.

**Theorem 1.** *Let $A$ be an $d \times d$ symmetric positive semidefinite (PSD) matrix and $x_l, l = 1, \ldots, m$ be random starting vectors with Rademacher distribution. Then, for $\tilde{Tr}_m(A) = \frac{1}{m}\sum_{l=1}^m x_l^T A x_l$, with $m = O\left(\frac{\log(1/\eta)}{\varepsilon^2}\right)$, we have*

$$\mathbb{P}\left[\left|\tilde{Tr}_m(A) - Tr(A)\right| \leq \varepsilon \left|Tr(A)\right|\right] \geq 1 - \eta.$$

**Theorem 2** (Hutchinson's Estimator)**.** *Draw $x_l, l = 1, \ldots, m$, vectors with i.i.d. random $\{+1, -1\}$ entries. Return $\tilde{Tr}_m(A) = \frac{1}{m}\sum_{l=1}^m x_l^T A x_l$ as an approximation to $Tr(A)$.*

**Expected value analysis:**
For a single random $\pm 1$ vector $x$, we have

$$E[\tilde{Tr}_m(A)] = E[x^T A x] = E\left[\sum_{i=1}^d \sum_{j=1}^d x_i x_j A_{ij}\right] = \sum_{i=1}^d \sum_{j=1}^d E[x_i x_j A_{ij}] = \sum_{i=1}^d A_{ii}$$

So the estimator is correct in expectation:

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

It is unbiased estimator.

**Variance analysis:**

$$\mathrm{Var}[\tilde{\mathrm{Tr}}_m(A)] = \frac{1}{m}\mathrm{Var}[\mathbf{x}_l^T A \mathbf{x}_l] = \frac{1}{m}\left(\mathbb{E}[(\mathbf{x}_l^T A \mathbf{x}_l)^2] - \mathrm{Tr}(A)^2\right)$$

$$\mathbb{E}[(\mathbf{x}_l^T A \mathbf{x}_l)^2] = \mathbb{E}\left[\left(\sum_{i,j} x_i x_j A_{ij}\right)\left(\sum_{i',j'} x_{i'} x_{j'} A_{i'j'}\right)\right]$$

$$= \sum_{i \neq j} 2A_{ij}^2 + \sum_{i \neq j} A_{ij} A_{ji} + \sum_i A_{ii}^2$$

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

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

## Analysis

**Chebyshev's inequality:**

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

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

$$\Pr\left(\left|\tilde{\mathrm{Tr}}_m(A) - \mathrm{Tr}(A)\right| \geq \epsilon \cdot \mathrm{Tr}(A)\right) \leq \frac{\mathrm{Var}[\tilde{\mathrm{Tr}}_m(A)]}{(\epsilon \cdot \mathrm{Tr}(A))^2} \leq \frac{2}{m\epsilon^2}.$$

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

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

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 $\tau = (2\epsilon - \epsilon^2) \cdot \mathrm{Tr}(A)$ 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].

## Exercise:

- Would the proof using the Chebyshev inequality work if $x_l$'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 $y_l = Ux_l$ are also Gaussian for unitary $U$. $\chi^2$-distribution.)

## Exercise Solution:

For vectors $x_l$ drawn from an i.i.d Gaussian distribution $\mathcal{N}(0,1)$, the proof using the Chebyshev inequality would still be valid because Gaussian random variables have finite variance. The expectation and variance of the estimate can be computed as follows:

The expectation of the estimator $\tilde{\text{Tr}}_m(A)$ is:

$$\mathbb{E}[\tilde{\text{Tr}}_m(A)] = \mathbb{E}\left[\frac{1}{m}\sum_{l=1}^{m} x_l^T A x_l\right] = \mathbb{E}[x_l^T A x_l] = \text{Tr}(A),$$

since the expectation of $x_i^2$ is 1 and the expectation of $x_i x_j$ for $i \neq j$ is 0.

The variance of $x_l^T A x_l$ when $x_l$ has a Gaussian distribution is:

$$\text{Var}(x_l^T A x_l) = 2\sum_{i\neq j} A_{ij}^2,$$

which is 2 times the sum of the squares of the off-diagonal elements of $A$, due to the property that $y_l = U x_l$ are also Gaussian for unitary $U$, which maintains the distribution of $x_l$ due to the rotational invariance of the Gaussian distribution. The $\chi^2$-distribution of $y_l^T y_l$ has a variance of $2d$ where $d$ is the number of degrees of freedom.

Therefore, the variance of our estimator $\tilde{\text{Tr}}_m(A)$ is:

$$\text{Var}[\tilde{\text{Tr}}_m(A)] = \frac{1}{m}\text{Var}(x_l^T A x_l) = \frac{2}{m}\sum_{i\neq j} A_{ij}^2.$$

# 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.

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

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

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

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

## Good low rank approximation

Let $A_k$ be the best rank-$k$ approximation of $A$.

***Lemma (Woo14)***

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

$$\|A - \hat{A}_k\|_F \leq 2\|A - A_k\|_F.$$

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

$$\text{Tr}(\hat{A}_k) = \text{Tr}(QQ^T A) = \text{Tr}(Q^T(AQ))$$

## Hutch++ Algorithm

**Input**: Number of matvecs $m$ and input matrix $A$.

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

We have the following result:

**Lemma**

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

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

## Hutch++ mean and variance

**Theorem 3.** *Let $A \in \mathbb{R}^{d\times d}$ be a PSD matrix, for fixed $k$ and $m$, construct $Q \in \mathbb{R}^{d\times m}$ as before. Let $Hutch$++$(A) = Tr(Q^T(AQ)) + \tilde{Tr}_m((I - QQ^T)A)$. Then,*

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

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

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

For variance, we use the Conditional Variance Formula,

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

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

Given $Q$ fixed, $\mathrm{Tr}(Q^T(AQ))$ is a constant as it is the exact trace of the $k$-rank approximation of $A$. Therefore, the conditional expectation $\mathbb{E}[\text{Hutch++}(A)|Q]$ given $Q$ is also a constant. Hence, the conditional variance $\mathrm{Var}[\mathbb{E}[\text{Hutch++}(A)|Q]]$ is zero, because the variance of a constant is zero.

$$\mathrm{Var}[\mathbb{E}[\text{Hutch++}(A)|Q]] = 0$$

since the variance of a constant, which is the value of Hutch++$(A)$ given $Q$, is always zero regardless of the distribution of $Q$.

Now,

$$\begin{aligned}
\mathbb{E}[\mathrm{Var}[\text{Hutch++}(A)|Q]] &= \mathbb{E}[\mathrm{Var}[\mathrm{Tr}(Q^T(AQ))] + \mathbb{E}[\mathrm{Var}[\tilde{\mathrm{Tr}}_m((I - QQ^T)A)]] \\
&= 0 + \frac{2}{m}\mathbb{E}[\|\tilde{\mathrm{Tr}}_m((I - QQ^T)A)\|_F^2] \\
&\leq \frac{4}{m}\|A - A_k\|_F^2 \\
&\leq \frac{1}{km}\mathrm{Tr}^2(A)
\end{aligned}$$