# COMPSCI 614: Randomized Algorithms with Applications to Data Science

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2024.

Lecture 12

# Logistics

- The midterm is the Thursday after break in class.
- I will hold a review session Monday from 3-4:30pm and Tuesday in class.
- There is no real quiz this week, but see Weekly Quizzes section on Moodle for a single question quiz where you can mark if you attended Sally Dong's job talk for extra credit.

- Practice midterm in moodle.
  - solutions
  - study guide

# Summary

Last Time:

- Finish up fast low-rank approximation via approximate matrix multiplication.
- Start on stochastic trace estimation and motivation for matrix-vector query algorithms.

# Summary

Last Time:

- Finish up fast low-rank approximation via approximate matrix multiplication.
- Start on stochastic trace estimation and motivation for matrix-vector query algorithms.

Today:

- Finish stochastic trace estimation.
- Hutchinson's estimator and full analysis.
- k-~~means~~ ++

# Matrix Trace

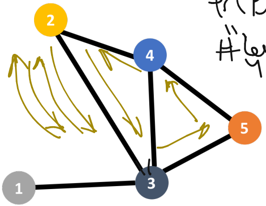
The trace of a matrix $A \in \mathbb{R}^{n \times n}$ is the sum of it diagonal entries.

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

When $A$ is diagonalizable (e.g., when it is symmetric) with eigenvalues $\lambda_1, \ldots, \lambda_n$, $\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i$.

**Main question:** How many matrix-vector multiplication "queries" $Ax_1, \ldots, Ax_m$ are required to approximate $\mathrm{tr}(A)$?

# Motivating Example




$\mathrm{tr}(B^4)$ = # length 4 paths

**$B^3$**

| | | | | |
|---|---|---|---|---|
| 0 | 1 | 4 | 2 | 1 |
| 1 | 2 | 6 | 5 | 2 |
| 4 | 6 | 4 | 6 | 6 |
| 2 | 5 | 6 | 4 | 5 |
| 1 | 2 | 6 | 5 | 2 |

sparse matrix approx

$$\frac{1}{6}\,\mathbf{tr}(B^3) = \#\text{ triangles.}$$

$B(B(Bx))$

$\mathrm{tr}(B^3)$ = # length 3 paths starting at a vertex ending at the same vertex

- Explicitly forming $B^3$ and computing $\mathbf{tr}(B^3)$ takes $O(n^3)$ time.
- Can multiply $B^3$ by a vector in $3 \cdot |E| = O(n^2)$ operations.
- So a trace estimation algorithm using $m$ queries, yields an $O(m \cdot |E|)$ time approximate triangle counting algorithm.

# Other Examples

**Example 2:** Hessian/Jacobian matrix-vector products.

$n \times n$ Jacobian  $n \times n$ Hessian

- For vector $x$, $\nabla \mathbf{f}(y)x$ and $\nabla^2 f(y)x$ can often be computed efficiently using finite difference methods or explicit differentiation (e.g., via backpropagation).
- Do not need to fully form $\nabla f(y)$ or $\nabla^2 f(y)$.
- Many applications of estimating the traces of these matrices, e.g., in analyzing neural network convergence, in optimization of score-based methods, etc.
- $\text{tr}(\nabla^2 f(y)x)$: Laplacian
- $\text{tr}(\nabla f(y)x)$: Divergence

# Other Examples

**Example 3:** $A$ is a function of another (explicit) matrix $B$, $A = f(B)$ that can be applied efficiently via an iterative method.

# Other Examples

$\Sigma^{1/2} \cdot \Sigma^{1/2} = \Sigma$

Example 3: $A$ is a function of another (explicit) matrix $B$, $A = f(B)$ that can be applied efficiently via an iterative method.

$n \times n$

$x \sim N(0, \Sigma)$

$\Sigma^{1/2} g$ – iid Gaussian

- Repeated multiplication to apply $A = B^3$.
- Conjugate gradient, MINRES, or any linear system solver:

$$A = B^{-1}.$$

$\Sigma^{1/2} \underbrace{E[gg^T]}_{I} \Sigma^{1/2}$

$\Sigma$

- Lanczos method, polynomial/rational approximation:

$$A = \exp(B),\ A = \sqrt{B},\ A = \log(B),\ \text{etc.}$$

- These methods run in $n^2 \cdot C$ time, where $C$ depends on properties of $B$. Typically $C \ll n$ so $n^2 \cdot C \ll n^3$.

# Matrix Function Examples

- Log-likelihood computation in Bayesian optimization, experimental design. $\mathrm{tr}(\log(B)) = \log\det(B)$. $= \sum \log(\lambda_i)$
- Estrada index, a measure of protein folding degree and more generally, network connectivity. $\mathrm{tr}(\exp(B))$.
- Trace inverse, which is important in uncertainty quantification and many other scientific computing applications. $\mathrm{tr}(B^{-1})$


- Information about the matrix eigenvalue spectrum, since $\mathrm{tr}(f(B)) = \sum_{i=1}^{n} f(\lambda_i)$, where $\lambda_i$ is $B$'s $i^{th}$ eigenvalue.
- E.g., counting the number of eigenvalues in an interval, spectral density estimation, matrix norms
- See e.g., [Ubaru, and Saad 2017].

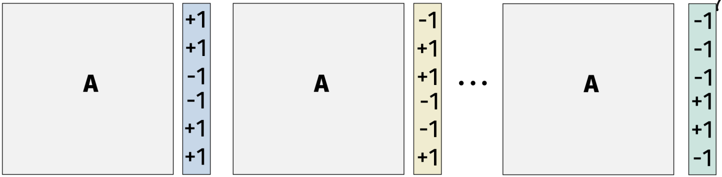
# Hutchinson's Method

Hutchinson 1991, Girard 1987:

$[1\ -1\ -1\ 1]\,[A]\begin{bmatrix}1\\-1\\1\\1\end{bmatrix}$

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \underline{\mathbf{x}_i^T A \mathbf{x}_i}$ as an approximation to $\mathrm{tr}(A)$.

gaussian




- One of the earliest examples I know of a randomized algorithm for linear algebraic computation.

# Hutchinson's Method Error Bound

**Theorem**

*Let $\overline{\mathsf{T}}$ be the trace estimate returned by Hutchinson's method. If $m = O\left(\frac{1}{\delta\epsilon^2}\right)$, then with probability $\geq 1 - \delta$,*

$$\left|\overline{\mathsf{T}} - \mathrm{tr}(A)\right| \leq \epsilon \|A\|_F$$

# Hutchinson's Method Error Bound

**Theorem**

*Let $\bar{\mathsf{T}}$ be the trace estimate returned by Hutchinson's method. If $m = O\left(\frac{1}{\delta\epsilon^2}\right)$, then with probability $\geq 1-\delta$,*

tr(A)

$$\left|\bar{\mathsf{T}} - \mathrm{tr}(A)\right| \leq \epsilon \|A\|_F$$

– non-negative eigenvalues

If $A$ is symmetric positive semidefinite (PSD) then

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \lambda_i^2} \leq \sum_{i=1}^{n} \lambda_i = \mathrm{tr}(A).$$

So for PSD $A$: $(1-\epsilon)\,\mathrm{tr}(A) \leq \bar{\mathsf{T}} \leq (1+\epsilon)\,\mathrm{tr}(A).$

# Proof Approach

**Theorem**

*Let $\overline{\mathsf{T}}$ be the trace estimate returned by Hutchinson's method. If $m = O\left(\frac{1}{\delta\epsilon^2}\right)$, then with probability $\geq 1 - \delta$,*

$$\left|\overline{\mathsf{T}} - \mathrm{tr}(A)\right| \leq \epsilon \|A\|_F$$

1. Show that $\mathbb{E}[\overline{\mathsf{T}}] = \mathrm{tr}(A)$.
2. Bound $\mathrm{Var}[\overline{\mathsf{T}}]$.
3. Apply Chebyshev's inequality.

# Proof Approach

**Theorem**

*Let $\overline{\mathsf{T}}$ be the trace estimate returned by Hutchinson's method. If $m = O\left(\frac{1}{\delta\epsilon^2}\right)$, then with probability $\geq 1 - \delta$,*

$$\left|\overline{\mathsf{T}} - \mathrm{tr}(A)\right| \leq \epsilon \|A\|_F$$

1. Show that $\mathbb{E}[\overline{\mathsf{T}}] = \mathrm{tr}(A)$.
2. Bound $\mathrm{Var}[\overline{\mathsf{T}}]$.
3. Apply Chebyshev's inequality.

$$\begin{bmatrix}1 & -1 & 1 & 1\end{bmatrix}\begin{bmatrix}A\end{bmatrix}\begin{bmatrix}1\\-1\\1\end{bmatrix}$$

A tighter proof that uses the Hanson-Wright inequality, an exponential concentration inequality for quadratic forms, can improve the $\delta$ dependence to $\log(1/\delta)$ – we'll cover this later in the class.

# Expectation Analysis

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \underline{\mathbf{x}_i^T A \mathbf{x}_i}$ as an approximation to $\text{tr}(A)$.

---

By linearity of expectation, $\mathbb{E}[\bar{\mathsf{T}}] = \mathbb{E}[\mathbf{x}^T A \mathbf{x}]$ $= \text{tr}(A)$ for a single random $\pm 1$ vector $\mathbf{x}$.

# Expectation Analysis

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\operatorname{tr}(A)$.

---

By linearity of expectation, $\mathbb{E}[\bar{\mathsf{T}}] = \mathbb{E}[\mathbf{x}^T A \mathbf{x}]$ for a single random $\pm 1$ vector $\mathbf{x}$.

$$\mathbb{E}[\mathbf{x}^T A \mathbf{x}] = \mathbb{E} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \cdot \mathbb{E}[\mathbf{x}_i \mathbf{x}_j]$$

0 if $i \neq j$

1 if $i = j$

# Expectation Analysis

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

By linearity of expectation, $\mathbb{E}[\bar{\mathsf{T}}] = \mathbb{E}[\mathbf{x}^T A \mathbf{x}]$ for a single random $\pm 1$ vector $\mathbf{x}$.

$$\mathbb{E}[\mathbf{x}^T A \mathbf{x}] = \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} \cdot \mathbb{E}[\mathbf{x}_i \mathbf{x}_j]$$

- When $i \neq j$, $\mathbf{x}_i \mathbf{x}_j = 1$ with probability 1/2 and $-1$ with probability 1/2, so $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = 0$. When $i = j$, $\mathbf{x}_i \mathbf{x}_j = 1$, so $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = 1$.

# Expectation Analysis

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\operatorname{tr}(A)$.

---

By linearity of expectation, $\mathbb{E}[\bar{\mathsf{T}}] = \mathbb{E}[\mathbf{x}^T A \mathbf{x}]$ for a single random $\pm 1$ vector $\mathbf{x}$.

$$\mathbb{E}[\mathbf{x}^T A \mathbf{x}] = \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} \cdot \mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = \sum_{i=1}^{n} A_{ii}.$$

- When $i \neq j$, $\mathbf{x}_i \mathbf{x}_j = 1$ with probability $1/2$ and $-1$ with probability $1/2$, so $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = 0$. When $i = j$, $\mathbf{x}_i \mathbf{x}_j = 1$, so $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = 1$.

# Expectation Analysis

**Hutchinson's Estimator::** $(xx^T)_{ij} = x_i x_j = 1$ if $i = j$ only.

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

$\mathbb{E}\,\mathrm{tr}(x^T A x) = \mathbb{E}\,\mathrm{tr}(x x^T A) = \mathrm{tr}(\mathbb{E}\,x x^T A) = \mathrm{tr}(A)$

---

By linearity of expectation, $\mathbb{E}[\bar{\mathrm{T}}] = \mathbb{E}[\mathbf{x}^T A \mathbf{x}]$ for a single random $\pm 1$ vector $\mathbf{x}$.

$$\mathbb{E}[\mathbf{x}^T A \mathbf{x}] = \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} \cdot \mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = \sum_{i=1}^{n} A_{ii}. = \mathrm{tr}(A)$$

- When $i \neq j$, $\mathbf{x}_i \mathbf{x}_j = 1$ with probability $1/2$ and $-1$ with probability $1/2$, so $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = 0$. When $i = j$, $\mathbf{x}_i \mathbf{x}_j = 1$, so $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = 1$.
- So the estimator is correct in expectation: $\mathbb{E}[\bar{\mathrm{T}}] = \mathrm{tr}(A)$.

# Variance Bound

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

problem on pset 7 looks a lot like this.

---

$$\mathrm{Var}[\bar{\mathrm{T}}]$$

# Variance Bound

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\overline{\mathsf{T}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\operatorname{tr}(A)$.

---

$$\operatorname{Var}[\,\overline{\mathsf{T}}\,] = \frac{1}{m} \operatorname{Var}[\mathbf{x}^T A \mathbf{x}]$$

# Variance Bound

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

$$\mathrm{Var}[\bar{\mathsf{T}}] = \frac{1}{m} \mathrm{Var}[\mathbf{x}^T A \mathbf{x}] = \frac{1}{m} \mathrm{Var}\left[\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij}\right]$$

# Variance Bound

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

$$\mathrm{Var}[\bar{\mathsf{T}}] = \frac{1}{m}\mathrm{Var}[\mathbf{x}^T A \mathbf{x}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij}\right]$$

Can we apply linearity of variance here?

$\boxed{x_1 x_1 A_{11}}$

$\boxed{x_1 x_2 A_{12}}$

$\boxed{x_2 x_1 A_{21}}$

# Variance Bound

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

$$\mathrm{Var}[\bar{\mathsf{T}}] = \frac{1}{m}\mathrm{Var}[\mathbf{x}^T A \mathbf{x}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij}\right]$$

Can we apply linearity of variance here? Almost – need to remove repeated terms, and then can use pairwise independence.

### Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

$n^2$ terms

$$\mathrm{Var}[\bar{\mathrm{T}}] = \frac{1}{m}\mathrm{Var}[\mathbf{x}^T A \mathbf{x}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij}\right]$$

"reorganized"

Can we apply linearity of variance here? Almost – need ~~to remove repeated terms~~, and then can use pairwise independence.

$n$ terms $\quad \binom{n}{2} = \frac{n\cdot(n-1)}{2}$

$$\mathrm{Var}[\bar{\mathrm{T}}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n} A_{ii} + \sum_{i=1}^{n}\sum_{j>i} \mathbf{x}_i \mathbf{x}_j (A_{ij} + A_{ji})\right]$$

$x_1 x_2 (A_{12} + A_{21})$

# Variance Bound

**Hutchinson's Estimator::**

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\overline{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

$$\mathrm{Var}[\,\overline{\mathrm{T}}\,] = \frac{1}{m}\mathrm{Var}[\mathbf{x}^T A \mathbf{x}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij}\right]$$

Can we apply linearity of variance here? Almost – need to remove repeated terms, and then can use pairwise independence.

$$\mathrm{Var}[\,\overline{\mathrm{T}}\,] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n} A_{ii} + \sum_{i=1}^{n}\sum_{j>i} \mathbf{x}_i \mathbf{x}_j (A_{ij} + A_{ji})\right]$$

$$= \frac{1}{m}\sum_{i=1}^{n}\sum_{j>i} \underbrace{\mathrm{Var}[\mathbf{x}_i \mathbf{x}_j]}_{\stackrel{?}{=} 1} \cdot \underbrace{(A_{ij} + A_{ji})^2}_{\leq}$$

# Variance Bound

**Hutchinson's Estimator::**

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

$(A-B)^2$
$= A^2 + B^2 - 2AB$
$2AB \le A^2 + B^2$

---

$$\mathrm{Var}[\bar{\mathrm{T}}] = \frac{1}{m}\mathrm{Var}[\mathbf{x}^T A \mathbf{x}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i \mathbf{x}_j A_{ij}\right]$$

Can we apply linearity of variance here? Almost – need to remove repeated terms, and then can use pairwise independence.

"arithmetic geometric mean inequality"

$\sqrt{AB} \le \frac{A+B}{2}$

$$\mathrm{Var}[\bar{\mathrm{T}}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^n A_{ii} + \sum_{i=1}^n \sum_{j>i} \mathbf{x}_i \mathbf{x}_j (A_{ij} + A_{ji})\right]$$

1

$$= \frac{1}{m}\sum_{i=1}^n \sum_{j>i} \mathrm{Var}[\mathbf{x}_i \mathbf{x}_j] \cdot (A_{ij} + A_{ji})^2 \le \frac{1}{m}\sum_{i=1}^n \sum_{j>i} 2A_{ij}^2 + 2A_{ji}^2$$

$A_{ij}^2 + A_{ji}^2 + 2A_{ij}A_{ji} \le 2 \cdot \frac{(A_{ij}^2 + A_{ji}^2)}{2}$

# Variance Bound

**Hutchinson's Estimator::**

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

$$\begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}$$

---

$$\mathrm{Var}[\bar{\mathrm{T}}] = \frac{1}{m}\mathrm{Var}[\mathbf{x}^T A \mathbf{x}] = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{x}_i \mathbf{x}_j A_{ij}\right]$$

Can we apply linearity of variance here? Almost – need to remove repeated terms, and then can use pairwise independence.

$$\boxed{\mathrm{Var}[\bar{\mathrm{T}}]} = \frac{1}{m}\mathrm{Var}\left[\sum_{i=1}^{n} A_{ii} + \sum_{i=1}^{n}\sum_{j>i} \mathbf{x}_i \mathbf{x}_j (A_{ij} + A_{ji})\right]$$

$$\frac{2\|A - \mathrm{diag}(A)\|_F^2}{m}$$

$$= \frac{1}{m}\sum_{i=1}^{n}\sum_{j>i} \mathrm{Var}[\mathbf{x}_i \mathbf{x}_j] \cdot (A_{ij} + A_{ji})^2 \leq \frac{1}{m}\sum_{i=1}^{n}\sum_{j>i} 2A_{ij}^2 + 2A_{ji}^2 \leq \boxed{\frac{2\|A\|_F^2}{m}}.$$

$$\leq \frac{1}{m} \cdot 2 \cdot \sum_{i=1}\sum_{j=1} A_{ij}^2$$

## Final Analysis

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\overline{\mathsf{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

Chebyshev's inequality implies that, for $m = \frac{2}{\delta\epsilon^2}$:

$$\Pr\left[\left|\overline{\mathsf{T}} - \mathrm{tr}(A)\right| \geq \epsilon\|A\|_F\right] \leq \frac{2\|A\|_F^2/m}{\epsilon^2\|A\|_F^2} = \delta.$$

$$\frac{2}{m\varepsilon^2}$$

# Final Analysis

## Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\bar{\mathsf{T}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathsf{tr}(A)$.

---

Chebyshev's inequality implies that, for $m = \frac{2}{\delta \epsilon^2}$:

$x^T A x$

$$\Pr\left[\left|\bar{\mathsf{T}} - \mathsf{tr}(A)\right| \geq \epsilon \|A\|_F\right] \leq \frac{2\|A\|_F^2/m}{\epsilon^2 \|A\|_F^2} = \delta.$$

$x_1 x_2 = 1 \quad x_2 x_3 = 1 \Rightarrow x_1 x_3 = 1$

Could we have gotten a better bound by applying Bernstein's inequality to $\sum_{i=1}^{n} \sum_{j>i} \mathbf{x}_i \mathbf{x}_j (A_{ij} + A_{ji})$?

$x^T A x$

- upper bounds on $A_{ij} + A_{ji}$? ✓

- pairwise independent ✗

# Final Analysis

Hutchinson's Estimator::

- Draw $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ i.i.d. with random $\{+1, -1\}$ entries.
- Return $\overline{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^T A \mathbf{x}_i$ as an approximation to $\mathrm{tr}(A)$.

---

Chebyshev's inequality implies that, for $m = \frac{2}{\delta\epsilon^2}$:

$$\Pr\left[\left|\overline{\mathrm{T}} - \mathrm{tr}(A)\right| \geq \epsilon\|A\|_F\right] \leq \frac{2\|A\|_F^2/m}{\epsilon^2\|A\|_F^2} = \delta.$$

Could we have gotten a better bound by applying Bernstein's inequality to $\sum_{i=1}^{n}\sum_{j>i} \mathbf{x}_i\mathbf{x}_j(A_{ij} + A_{ji})$?

Hanson-Wright is an exponential concentration bound that can be used in the specific case – improves bound to $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$.

# Optimality of Hutchinson's Method

The $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ bound given by the Hanson-Wright inequality is tight.

- Any algorithm that only uses queries of the form $\mathbf{x}_i^T A \mathbf{x}_i$ requires $\Omega\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ samples to estimate $\mathrm{tr}(A)$ to error $\pm\epsilon\,\mathrm{tr}(A)$ for PSD A [Wimmer, Wu, Zhang 2014].

$\epsilon\|A\|_F$

# Optimality of Hutchinson's Method

The $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ bound given by the Hanson-Wright inequality is tight.

- Any algorithm that only uses queries of the form $\underline{\mathbf{x}_i^T A \mathbf{x}_i}$ requires $\Omega\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ samples to estimate $\mathrm{tr}(A)$ to error $\pm\epsilon\,\mathrm{tr}(A)$ for PSD A [Wimmer, Wu, Zhang 2014]. $\underline{Ax_1} \ldots Ax_m$
- We recently showed that using the full power of matrix-vector queries, one can achieve $O\left(\frac{\log(1/\delta)}{\epsilon}\right)$ queries for PSD matrices.

(Hutch++ Meyer et al.