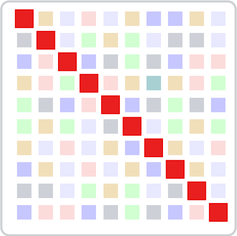
Stochastic Trace Estimation

Monte Carlo estimators, structure-aware probing, variance reduction, and Lanczos quadrature

Alice Cortinovis
MoRAT Summer School



Road map

1. Basic Girard-Hutchinson estimator

Based on quadratic forms, unbiased estimator

2. Concentration inequalities

How accurate is our trace estimate? What tools can we use to show these results?

3. Variance reduction

Structured probing and low-rank control variates

4. Trace of matrix functions

Stochastic Lanczos quadrature

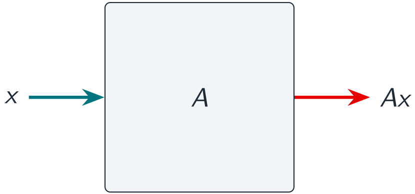
Trace estimation from matrix–vector products

The matrix-free model

Let $A \in \mathbb{R}^{n \times n}$ be too large or too expensive to form.
We can only query

$$x \longmapsto Ax.$$

How can we recover the trace? $A \cdot e_1, Ae_2, \ldots, Ae_n$



One application of A is the cost unit.

Goal

Get APPROXIMATION of $\mathrm{tr}(A)$
using $<n$ matrix-vector multiplications

Classical stochastic trace estimation: [Gir87; Hut89; AT11].

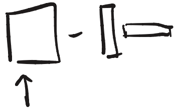
Why estimate a trace?

$(A^k)_{ij}$ = # paths of length k that start from i and finish at vertex j

Norms

$$\|B\|_F^2 = \mathrm{tr}(B^T B).$$

A trace estimator therefore estimates Frobenius norms of implicit operators.



sparse

Spectral sums

For symmetric A,

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

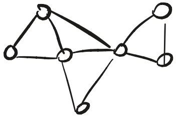
Examples include log determinants and eigenvalue counts.

$\log(\det(A)) = \mathrm{tr}(\log(A))$

Graphs

For the adjacency matrix of a simple undirected graph,

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



The key identity: isotropic probing

Theorem (Unbiased quadratic form)

Let $\omega \in \mathbb{R}^n$ satisfy

$$\mathbb{E}[\omega\omega^T] = I. \longrightarrow \omega = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_n \end{bmatrix} \qquad \mathbb{E}[\omega_i \omega_j] = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Then, for every $A \in \mathbb{R}^{n\times n}$,

$$\mathbb{E}[\omega^T A\omega] = \mathrm{tr}(A).$$

Proof.

$$\mathbb{E}[\omega^T A \omega] = \mathbb{E}[\mathrm{tr}(\omega^T A \omega)] = \mathbb{E}[\mathrm{tr}(A\omega\omega^T)] = \mathrm{tr}(\mathbb{E}[A\omega\omega^T])$$

$$= \mathrm{tr}\big(A \cdot \underbrace{\mathbb{E}[\omega\omega^T]}_{I}\big) = \mathrm{tr}(A)$$

□

The Girard–Hutchinson estimator

Draw independent isotropic vectors $\omega_1, \ldots, \omega_m$ and set

$$\boxed{\widehat{t}_{\mathrm{GH}}(A) = \frac{1}{m} \sum_{j=1}^{m} \omega_j^T A \omega_j.}$$

Cost

m quadratic forms with A

Variance

$$\mathrm{Var}(\hat{t}_{GH}(A)) = \frac{1}{m} \mathrm{Var}(\omega^T A \omega)$$

Girard [Gir87]; Hutchinson [Hut89].

Which probes can we use?

In theory: any isotropic random vector!

Probe ω	Isotropic?	Main advantage	Main drawback
Gaussian $N(0, I)$	yes	rotation invariance and clean analysis	diagonal entries contribute to variance
Rademacher $\mathrm{Unif}\{\pm 1\}^n$	yes	diagonal contributes no variance	depends on the coordinate system
Uniform on the sphere	yes	removes radial fluctuation	slightly less convenient to generate
Coordinate $\sqrt{n}\, e_J$	yes	useful for special sparsity patterns	can have enormous variance

Exact variance formulas

Theorem

Let A be a symmetric matrix.

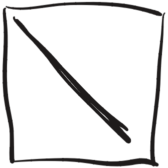
$$\begin{aligned} \mathrm{Var}(\omega^T A\omega) &= 2\,\|A\|_F^2\,, && \omega \sim N(0, I), \\ \mathrm{Var}(\omega^T A\omega) &= 2\,\|\mathrm{offdiag}(A)\|_F^2\,, && \omega \sim \textit{Rademacher}. \end{aligned}$$

For m independent probes, divide each variance by m.

A diagonal

$\omega \sim$ Rademacher vector

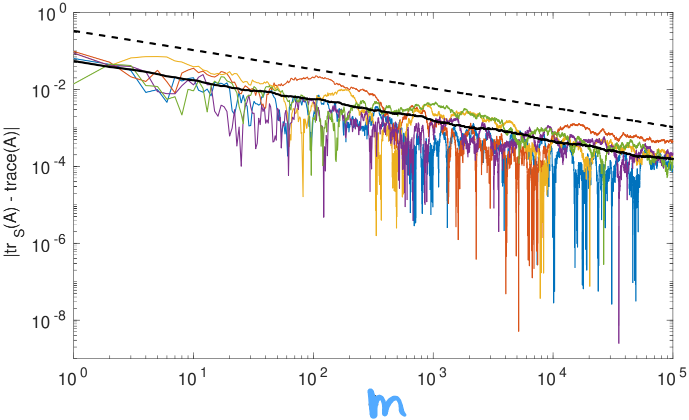
$\omega^T A \omega = \sum (\omega_i)^2 a_{ii} = \mathrm{tr}(A)$



Exact formulas go back to [Gir87; Hut89].

An example (with Gaussian random vectors)

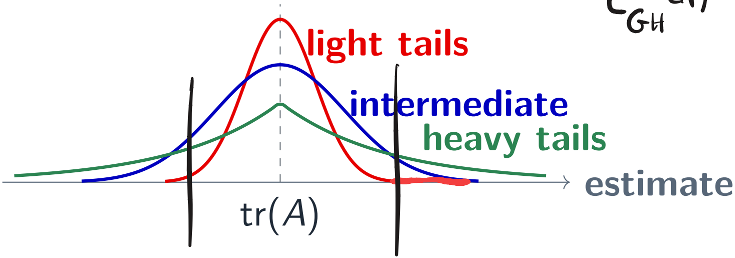
- $A \in \mathbb{R}^{1000\times 1000}$, random SPD.
- Gaussian random vectors
- Black line: average over 100 runs
- Black dotted line: $O(1/\sqrt{m})$



Concentration of quadratic forms

How many probes guarantee ε accuracy with failure probability at most δ?

$$\hat{t}_{GH}(A) = \frac{1}{m} \sum_{i=1}^{m} w_i^T A w_i$$



First attempt: Markov inequality

Theorem (Markov inequality)

If $X \geq 0$, then, for every $t > 0$,

$$\boxed{\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t}.}$$

Consequence

$$\mathbb{P}\left((X-\mathbb{E}[X])^2 \geq t^2\right) \leq \frac{\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

Second attempt: Chebyshev inequality

Theorem (Chebyshev inequality)

If X has finite variance, then, for every $t > 0$,

$$\boxed{\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathrm{Var}(X)}{t^2}.}$$

Consequence for Girard–Hutchinson

$$X = \hat{t}_{GH}(A) = \frac{1}{m}\sum_{i=1}^{m} w_i^T A w_i \qquad w_i \sim \text{Rademacher}$$

$$\mathbb{P}\left(|\hat{t}_{GH}(A) - \mathrm{tr}(A)| \geq \varepsilon\right) \leq \underbrace{\frac{2}{m}\|\mathrm{offdiag}(A)\|_F^2 \cdot \frac{1}{\varepsilon^2}}_{} = \delta$$

$m \geq \frac{2}{\delta\varepsilon^2}\|\mathrm{offdiag}(A)\|_F^2$ are sufficient to guarantee accuracy ε up to failure prob. δ

$m \geq \frac{2n^2}{\varepsilon^2} \|A\|_2^2 \log 2/\delta$

Theorem (Hoeffding inequality)

If $X_1, \ldots, X_m$ are independent and $a_j \le X_j \le b_j$, then

$$\mathbb{P}\left\{ \left| \sum_j (X_j - \mathbb{E}X_j) \right| \ge t \right\} \le 2\exp\left(-\frac{2t^2}{\sum_j (b_j - a_j)^2} \right).$$

Consequence for Girard–Hutchinson

$X_i = \frac{1}{m} w_i^T A w_i$

$|X_i| = \frac{1}{m} | w_i^T A w_i | \le \frac{1}{m} \|A\|_2 \|w_i\|^2$

$|X_i| \le \frac{n}{m} \|A\|_2$

$\mathbb{P}\left\{ \left| \hat{t}_{GH}(A) - \operatorname{tr}(A) \right| \ge \varepsilon \right\} \le 2 \exp\left(- \frac{\varepsilon^2 m}{2 n^2 \|A\|_2^2} \right)$

$m \geq \frac{2n^2}{\varepsilon^2} \|A\|_2^2 \log 2/\delta$

Bernstein inequality and Rademacher probes

Theorem (Bernstein inequality)

If $X_1, \ldots, X_m$ are independent, centred, $|X_j| \le b$, and $V = \sum_j \mathbb{E}[X_j^2]$, then

$$\mathbb{P}\left\{\left|\sum_j X_j\right| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2(V + bt/3)}\right).$$

Consequence for Girard–Hutchinson

$$m \geq \left(\frac{4}{\varepsilon^2}\|\operatorname{offdiag}(A)\|_F^2 + \frac{3n}{\varepsilon}\|A\|_2\right)\log 2/\delta$$

Chernoff: moment generating functions control tails

For a centred random variable Y and $\theta > 0$, Markov applied to $e^{\theta Y}$ gives

$$\mathbb{P}\{Y \geq t\} \leq e^{-\theta t}\mathbb{E}e^{\theta Y} = \exp\bigl(-\theta t + \log \mathbb{E}e^{\theta Y}\bigr) .$$

Optimizing over θ gives the *Chernoff bound*:

$$\mathbb{P}\{Y \geq t\} \leq \inf_{\theta > 0} \exp\bigl(-\theta t + \log \mathbb{E}e^{\theta Y}\bigr) .$$

The message

we want to contrвe the (log of the) MGF of $w^T A w$

Sub-Gaussian, sub-exponential, and sub-gamma tails

Let Y be centred.

Sub-Gaussian (σ^2)

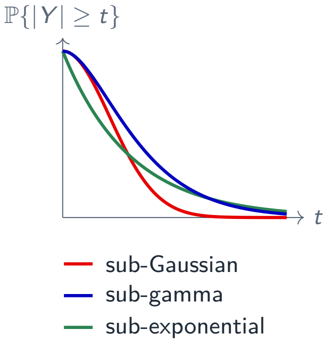
$$\log \mathbb{E} e^{\theta Y} \le \frac{\sigma^2\theta^2}{2}, \qquad \theta \in \mathbb{R}.$$

Sub-exponential (ν, b)

$$\log \mathbb{E} e^{\theta Y} \le \frac{\nu^2\theta^2}{2}, \qquad |\theta| < 1/b.$$

Sub-gamma (v, c)

$$\log \mathbb{E} e^{\theta Y} \le \frac{v\theta^2}{2(1 - c\,|\theta|)}, \qquad |\theta| < 1/c.$$



Terminology varies slightly: here sub-exponential means local quadratic MGF control.

sub-Gaussian behavior and Hoeffding inequality

Hoeffding: boundedness $\Longrightarrow$ sub-Gaussian $\Longrightarrow$ Gaussian tail

If X_j are independent and $a_j \le X_j \le b_j$, set $Z = \sum_j (X_j - \mathbb{E}X_j)$ and $S = \sum_j (b_j - a_j)^2$. Then

$$\log \mathbb{E} e^{\theta Z} \le \frac{\theta^2 S}{8}.$$

Thus Z is sub-Gaussian with parameter $S/4$; Chernoff gives

$$\boxed{\mathbb{P}\{|Z| \ge t\} \le 2e^{-2t^2/S}.}$$

$$\mathbb{P}(Z \ge t) \le \inf_{\theta > 0} \exp\left(-\theta t + \log \mathbb{E} e^{\theta Z}\right)$$

$$\le \inf_{\theta > 0} \exp\left(-\theta t + \frac{\theta^2 S}{8}\right)$$

$\hookrightarrow$ choose $\theta = \frac{4t}{S}$

sub-Gamma behavior and Bernstein inequality

Bernstein: variance and boundedness $\implies$ sub-gamma $\implies$ two-regime tail

If X_j are independent, centred, $|X_j| \le b$, and $V = \sum_j \mathbb{E}[X_j^2]$, then

$$\log \mathbb{E} e^{\theta \sum_j X_j} \le \frac{V\theta^2}{2(1 - b\,|\theta|\,/3)}.$$

The sum is sub-gamma with $(v, c) = (V, b/3)$; Chernoff gives

$$\boxed{\mathbb{P}\left\{\left|\sum_j X_j\right| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2(V + bt/3)}\right).}$$

The MGF of a Gaussian quadratic form

$g^T A g \qquad g \sim N(0, I)$

Let $g \sim N(0, I)$ and write $A = U \Lambda U^T$.

- Rotation invariance: $g^T A g = g^T U \Lambda \underbrace{U^T g}_{\sim g} \sim g^T \Lambda g = \sum_{i=1}^{n} \lambda_i g_i^2$

  $g_i \sim N(0,1)$ iid

- MGF of Chi-square random variables:

$$\mathbb{E}\left[\exp(\theta g_i^2)\right] = \frac{1}{\sqrt{1-2\theta}} \quad \text{for } \theta < \frac{1}{2}$$

- Showing sub-Gamma behavior:

$$\mathbb{E}\left[\exp\left(\theta\left(\sum(\lambda_i g_i^2 - \lambda_i)\right)\right)\right] \leq \exp\left(\frac{\theta^2 \|A\|_F^2}{1-2|\theta|\,\|A\|_2}\right) \quad \text{for } |\theta| < \frac{1}{2\|A\|_2}$$

Concentration of a Gaussian quadratic form

The sub-gamma tail bound gives, for every $t > 0$,

$$\boxed{\mathbb{P}\left\{\left|g^T A g - \mathrm{tr}(A)\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{4\left\|A\right\|_F^2 + 4t\left\|A\right\|_2}\right).}$$

More generally, if $x_1, \ldots, x_n$ are independent, centered, and $\left\|x_i\right\|_{\psi_2} \leq K$, then the Hanson–Wright inequality states

$$\mathbb{P}\left\{\left|x^T A x - \mathbb{E}(x^T A x)\right| \geq t\right\} \leq 2\exp\left[-c\min\left\{\frac{t^2}{K^4\left\|A\right\|_F^2}, \frac{t}{K^2\left\|A\right\|_2}\right\}\right],$$

where $c > 0$ is an absolute constant.

General form: [RV13]; explicit Gaussian constants: [CK22].

Averaging m Gaussian probes

For an average of m independent copies, the sub-gamma parameters become

$$v_m = \frac{2\|A\|_F^2}{m}, \qquad c_m = \frac{2\|A\|_2}{m}.$$

Consequently,

$$\boxed{\mathbb{P}\left\{\left|\widehat{t}_{\mathrm{GH}}(A) - \mathrm{tr}(A)\right| \geq t\right\} \leq 2\exp\left(-\frac{mt^2}{4\|A\|_F^2 + 4t\|A\|_2}\right).}$$

Cortinovis–Kressner [CK22].

$$m \geq \left(\frac{4}{\varepsilon^2}\|A\|_F^2 + \frac{4}{\varepsilon}\|A\|_2\right)\log 2/\delta$$

The symmetric positive semidefinite case

Let $A \succeq 0$, $A \neq 0$, and

$$\mu = \frac{\|A\|_2}{\mathrm{tr}(A)}.$$

Since $\|A\|_F^2 \leq \|A\|_2 \, \mathrm{tr}(A)$, it is sufficient that

$$m \geq \frac{4(1+\varepsilon)\mu}{\varepsilon^2} \log \frac{2}{\delta}.$$

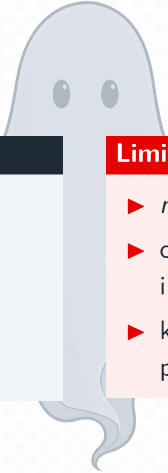
Since $\mu \leq 1$, a universal rule for $0 < \varepsilon \leq 1$ is

$$\boxed{m \geq \frac{8}{\varepsilon^2} \log \frac{2}{\delta}.}$$

independent from A!

[CK22]; classical PSD analysis: [AT11].

Takeaway for the basic Girard-Hutchinson trace estimator



Strengths

- one line of code;
- embarrassingly parallel;
- unbiased;
- dimension-free relative guarantees for $A \succeq 0$.

Limitations

- $m = O(\varepsilon^{-2})$ at fixed confidence;
- cancellation can make relative error ill-conditioned;
- known structure in A is unused unless probes are adapted.

Small-ball estimates: a different question

Concentration controls moderate deviations around the mean. A *small-ball estimate* instead asks how likely a random quantity is to be extremely close to zero:

$$\mathbb{P}\{\|X\|_2 \leq \theta\} \leq \phi(\theta), \qquad \phi(\theta) \longrightarrow 0 \quad (\theta \downarrow 0).$$

For $g \sim \mathcal{N}(0, I_n)$ and fixed $x \neq 0$,

$$\mathbb{P}\{\left|x^T g\right| \leq \theta \, \|x\|_2\} \leq \sqrt{\frac{2}{\pi}}\,\theta.$$

With m independent Gaussian probes, $\Omega^T x / \|x\|_2 \sim \mathcal{N}(0, I_m)$ and

$$\boxed{\mathbb{P}\left\{ \frac{\left\|\Omega^T x\right\|_2}{\sqrt{m}} \leq \theta \, \|x\|_2 \right\} = F_m(m\theta^2) \leq \frac{(m/2)^{m/2}}{\Gamma(m/2+1)}\,\theta^m.}$$

Small samples can rule out severe underestimation

Let $B \in \mathbb{R}^{p \times n}$ and define

$$\psi_m(B) = \frac{1}{\sqrt{m}} \|B\Omega\|_F\,, \qquad \Omega \in \mathbb{R}^{n \times m} \text{ Gaussian.}$$

Since $A = B^T B \succeq 0$,

$$\psi_m(B)^2 = \frac{1}{m}\operatorname{tr}(\Omega^T A \Omega) = \widehat{t}_{\mathrm{GH}}(A), \qquad \|B\|_F^2 = \operatorname{tr}(A).$$

The rank-one case is worst for lower tails, so uniformly for every B,

$$\boxed{\mathbb{P}\{\psi_m(B) \le \theta\,\|B\|_F\} \le F_m(m\theta^2), \qquad 0 < \theta < 1.}$$

Equivalently, $\mathbb{P}\{\widehat{t}_{\mathrm{GH}}(A) \le \alpha \operatorname{tr}(A)\} \le F_m(m\alpha)$.

A rough two-sided certificate from very few probes

Overestimation is even easier: with stable rank $\rho(B) = \|B\|_F^2 \,/\, \|B\|_2^2$, Chernoff bound gives

$$\mathbb{P}\{\psi_m(B) \geq \tau \|B\|_F\} \leq \exp\left[-\frac{m\rho(B)}{2}(\tau - 1)^2\right], \qquad \tau > 1.$$

Using only $\rho(B) \geq 1$ and a union bound yields universal factor-ten guarantees:

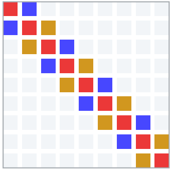
number of probes m	1	2	3
lower bound for $\mathbb{P}\{\|B\|_F/10 \leq \psi_m(B) \leq 10\|B\|_F\}$	0.920	0.990	0.999

Use case

Cheap certificates and stopping tests may need only to exclude catastrophic error, not deliver a high-accuracy estimate.

Variance reduction

Structured probing: can we quickly recover the trace exactly?



a tridiagonal matrix

the corresponding path graph

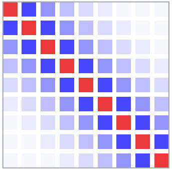
Question for you

How should we choose deterministic probe vectors so that the trace is recovered exactly?

$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \vdots \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \vdots \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \vdots \qquad \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$

Randomized probing when entries decay away from the band

exp(banded)



strong off-diagonal decay

Use random signs!

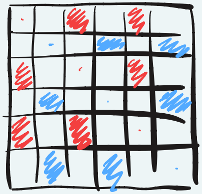
$$\omega_1 = \begin{bmatrix} \xi_1 & 0 & \xi_3 & 0 & \cdots \end{bmatrix}^T,$$

$$\omega_2 = \begin{bmatrix} 0 & \xi_2 & 0 & \xi_4 & \cdots \end{bmatrix}^T,$$

$\xi_i \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$.

$$\widehat{t}_{\text{SP}} = \omega_1^T A \omega_1 + \omega_2^T A \omega_2.$$

Variance



Stochastic probing analysis: [FRS25].

The control-variate principle

Suppose $\widetilde{A}$ is a symmetric approximation whose trace is cheap to compute. Then

$$\mathrm{tr}(A) = \mathrm{tr}(\widetilde{A}) + \mathrm{tr}(A - \widetilde{A}).$$

Estimate only the residual:

$$\widehat{t} = \mathrm{tr}(\widetilde{A}) + \frac{1}{m}\sum_{j=1}^{m} \gamma_j^T (A - \widetilde{A})\gamma_j.$$

Unbiasedness and variance

With independent Gaussian probes,

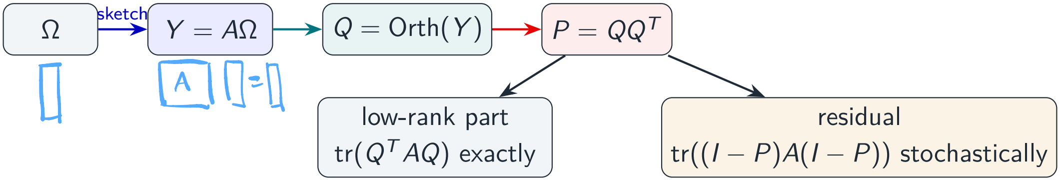
$$\mathbb{E}[\widehat{t} \mid \widetilde{A}] = \mathrm{tr}(A), \qquad \mathrm{Var}(\widehat{t} \mid \widetilde{A}) = \frac{2}{m}\left\|A - \widetilde{A}\right\|_F^2.$$

The algorithmic question

How do we obtain a useful $\widetilde{A}$ using only matrix–vector products with A?

Hutch++: sketch, capture, estimate

$A = QQ^TA + (A - QQ^TA)$



With $R = (I-P)A(I-P)$ and independent probes γ_j,

$$\boxed{\widehat{t}_{\mathrm{H}++}(A) = \underbrace{\mathrm{tr}(Q^TAQ)}_{=\,\mathrm{tr}(QQ^TA)} + \frac{1}{m}\sum_{j=1}^{m} \gamma_j^T R \gamma_j.}$$

Hutch++: [Mey+21].

Why Hutch++ improves the rate

For $A \succeq 0$, let A_r be a best rank-r approximation. Then

$$\|A - A_r\|_F \le \frac{\operatorname{tr}(A)}{\sqrt{r+1}}.$$

A Gaussian rangefinder with $k = 2r + 1$ satisfies

$$\mathbb{E}_\Omega \left\|(I - QQ^T)A\right\|_F^2 \le 2\,\|A - A_r\|_F^2 .$$

Combining this with the conditional variance of the residual estimator gives

$$\boxed{\frac{\mathbb{E}\left|\widehat{t}_{\mathrm{H}++}(A) - \operatorname{tr}(A)\right|^2}{\operatorname{tr}(A)^2} \le \frac{4}{m(r+1)}.}$$

Balanced allocation

Choose $m \asymp r \asymp \varepsilon^{-1}$: relative error $O(\varepsilon)$ with $O(\varepsilon^{-1})$ matvecs, versus $O(\varepsilon^{-2})$ for Girard–Hutchinson.

Rangefinder bound: [HMT11]; Hutch++ analysis: [Mey+21].

Randomized Nyström versus randomized SVD

Let $A \succeq 0$, draw $\Omega \in \mathbb{R}^{n\times k}$, and form $Y = A\Omega$.

Randomized SVD approximation

$$Q = \mathrm{Orth}(Y), \qquad \widetilde{A}_{\mathrm{SVD}} = QQ^TAQQ^T.$$

Forming the small core Q^TAQ requires another application of A to Q: a *second pass*.

Randomized Nyström approximation

$$\widetilde{A}_{\mathrm{Nys}} = Y(\Omega^TY)^\dagger Y^T.$$

It uses only $Y = A\Omega$: one pass!

Nyström++: one-pass variance reduction

Draw independent sketches $\Omega \in \mathbb{R}^{n\times k}$ and $\Psi \in \mathbb{R}^{n\times m}$, and make one block call

$$[Y\ Z] = A[\Omega\ \Psi], \qquad \widetilde{A}_{\mathrm{Nys}} = Y(\Omega^T Y)^\dagger Y^T.$$

Then correct the Nyström trace with a stochastic residual estimate:

$$\boxed{\widehat{t}_{\mathrm{Nys++}} = \operatorname{tr}(\widetilde{A}_{\mathrm{Nys}}) + \frac{1}{m}\operatorname{tr}\Big(\Psi^T (A - \widetilde{A}_{\mathrm{Nys}})\Psi\Big)\,.}$$

Nyström++: [PCK22].

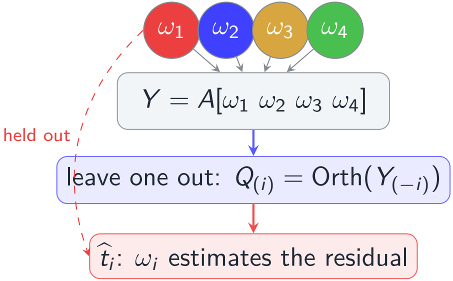
XTrace: every random vector does both jobs

Draw exchangeable probes $\omega_1, \ldots, \omega_m$ and compute $Y = A\Omega$ once. For each i,

$$Q_{(i)} = \mathrm{Orth}(Y_{(-i)}), \qquad Y_{(-i)} = A\Omega_{(-i)}.$$

The held-out probe estimates the residual; all other probes build its control variate:

$$\widehat{t}_i = \mathrm{tr}(Q_{(i)}^T A Q_{(i)}) + \omega_i^T (I - Q_{(i)} Q_{(i)}^T) A (I - Q_{(i)} Q_{(i)}^T) \omega_i.$$



Rotate the holdout: every vector plays both roles.

$$\boxed{\widehat{t}_{\mathrm{XTrace}} = \frac{1}{m} \sum_{i=1}^{m} \widehat{t}_i.}$$

XTrace: [ETW24].

XNysTrace: exchangeable Nyström control variates

For $A \succeq 0$, draw isotropic Gaussian probes $\Omega = [\omega_1 \ \cdots \ \omega_m]$ and compute $Y = A\Omega$ once. For every holdout i, define

$$Y_{(-i)} = A\Omega_{(-i)}, \qquad \widetilde{A}_{(i)} = Y_{(-i)}\big(\Omega_{(-i)}^T Y_{(-i)}\big)^\dagger Y_{(-i)}^T.$$

The held-out vector is independent of its Nyström control variate, so

$$\widehat{t}_i = \operatorname{tr}(\widetilde{A}_{(i)}) + \omega_i^T(A - \widetilde{A}_{(i)})\omega_i, \qquad \boxed{\widehat{t}_{\text{XNysTrace}} = \frac{1}{m}\sum_{i=1}^m \widehat{t}_i}$$

is unbiased.

Advantages

XNysTrace: [ETW24].

Which trace estimator should one use?

Method	Best suited for	Main caveat
Girard–Hutchinson	crude estimates; slowly decaying PSD spectra; minimal implementation effort	Monte Carlo rate $m^{-1/2}$
Structure-aware probing	known sparsity or graph-distance decay	quality and size of the colouring
Hutch++ / Nyström++	PSD matrices; moderate or fast spectral decay; tighter accuracy budgets	orthogonalization and low-rank post-processing

Trace of matrix functions

Why traces of matrix functions?

For $B = B^T = U\Lambda U^T$,

$$\mathrm{tr}(f(B)) = \sum_{i=1}^{n} f(\lambda_i),$$

so one trace encodes a global spectral statistic without computing the spectrum.

Log determinant

For $B \succ 0$,

$$\log\det B = \mathrm{tr}(\log B).$$

Central in Gaussian-process likelihoods and Bayesian inverse problems.

Eigenvalue counting

$$N(\tau) = \mathrm{tr}\left(\mathbf{1}_{(-\infty,\tau]}(B)\right).$$

Gives spectral distribution and density-of-states information.

Network communicability

$$\mathrm{EE}(B) = \mathrm{tr}(e^B) = \sum_i e^{\lambda_i}.$$

The Estrada index measures the abundance of closed walks in a network.

The obstacle

Need quadratic forms $x^T f(A)\, x$

How do we compute the quadratic forms?

1. “Explicit” polynomial approximation of f, e.g. Chebyshev polynomials or Taylor expansion [Boutsidis et al.'2017, Han et al.'2015, Pace/LeSage'2004, Zhang/Leithead'2007].
2. Via quadrature / Lanczos method [Bai/Fahey/Golub'1996, Golub/Meurant'2010, Ubaru/Chen/Saad'2017].

Let $B = Q \cdot D \cdot Q^T$ spectral decomposition and let $x \in \mathbb{R}^n$. Then

$$x^T f(B) x = (Q^T x)^T \cdot f(D) \cdot (Q^T x) = \int_{\lambda_{\min}}^{\lambda_{\max}} f(t) \mathrm{d}\mu(t), \text{ where } \mathrm{d}\mu(t) = \sum_{i=1}^{n} z_i^2 \delta_{\lambda_i}(t),\ z = Q^T x.$$

Gauss quadrature: $\text{integral} \approx \sum_{i=1}^{t} w_i f(\theta_i) =: \mathcal{I}_t.$

Approximation of quadratic forms via Lanczos method

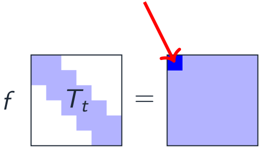
Golob / Meurant book

Theorem

Let T_t be matrix obtained after t steps of Lanczos method applied to B with starting vector x. Then $\mathcal{I}_t := \sum_{i=1}^{t} w_i f(\theta_i) = e_1^T f(T_t) e_1$ where

- *nodes θ_i = eigenvalues of T_t*
- *weights w_i = squares of first elements of normalized eigenvectors of T_t*

For unit vector x,
approximate $x^T f(B) x$ as



Obtained after t steps of Lanczos

Lanczos algorithm:
Set $q_1 = \omega / \|\omega\|$, $q_0 = 0$, and $\beta_0 = 0$.
For $j = 1, \ldots, t$,

$$z = Bq_j - \beta_{j-1} q_{j-1},$$
$$\alpha_j = q_j^T z,$$
$$z \leftarrow z - \alpha_j q_j,$$
$$\beta_j = \|z\|, \qquad q_{j+1} = z/\beta_j.$$

The coefficients α_j and β_j form tridiagonal matrix T_t.

Quadrature error is polynomial-approximation error

Assume $\sigma(B) \subset [a, b]$. Then

$$\left|\omega^T f(A)\omega - \|\omega\|^2 e_1^T f(T_t)e_1\right| \le 2\|\omega\|^2 \min_{p\in\mathcal{P}_{2t-1}} \|f - p\|_{\infty,[a,b]}.$$

Analytic f

Nonsmooth f

See [GM10; UCS17]; corrected analytic bounds in [CK22].

Stochastic Lanczos quadrature (SLQ)

For independent isotropic probes $\omega_1, \ldots, \omega_m$:

1. run t Lanczos steps for each ω_j to obtain $T_t^{(j)}$;
2. return

$$\widehat{s}_{m,t}(f) = \frac{1}{m} \sum_{j=1}^{m} \|\omega_j\|^2 \, e_1^T f(T_t^{(j)}) e_1.$$

Trace estimation error

GH analysis

Quadrature error

linked to polynomial approx.

Reuse

Pointers to recent literature

Open the Krylov black box

Krylov-aware trace estimation builds a block-Krylov space for A and reuses it both to approximate $f(A)$ and to construct a low-rank control variate.

- a larger projection space is obtained from the same matvecs;
- one Krylov run can serve several functions f;
- adaptive variants balance deflation and residual sampling.

Chen–Hallman: [CH23].

Other active directions

- **Multilevel estimators:** telescope across successively cheaper approximations and allocate probes by level [HT22].
- **Exchangeable estimators:** reuse samples across sketching and estimation, extending the XTrace principle to matrix functions [MAS26].
- **Certified/adaptive stopping:** estimate stochastic and Krylov errors separately rather than fixing m and t in advance.

A lesson we should learn

Try to look at the pieces of the estimator together

Summary

[AT11] H. Avron and S. Toledo. **"Randomized Algorithms for Estimating the Trace of an Implicit Symmetric Positive Semi-Definite Matrix".** In: *Journal of the ACM* 58.2 (2011), 8:1–8:34. DOI: 10.1145/1944345.1944349.

[CH23] T. Chen and E. Hallman. **"Krylov-Aware Stochastic Trace Estimation".** In: *SIAM Journal on Matrix Analysis and Applications* 44.3 (2023), pp. 1218–1244. DOI: 10.1137/22M1494257.

[CK22] A. Cortinovis and D. Kressner. **"On Randomized Trace Estimates for Indefinite Matrices with an Application to Determinants".** In: *Foundations of Computational Mathematics* 22 (2022), pp. 875–903. DOI: 10.1007/s10208-021-09525-9.

[EKN26] E. N. Epperly, D. Kressner, and Y. Nakatsukasa. **"Randomized Algorithms in Numerical Linear Algebra for Scientific Computing".** In: *Acta Numerica* (2026). To appear.

[ETW24] E. N. Epperly, J. A. Tropp, and R. J. Webber. **“XTrace: Making the Most of Every Sample in Stochastic Trace Estimation”.** In: *SIAM Journal on Matrix Analysis and Applications* 45.1 (2024), pp. 1–23. DOI: 10.1137/23M1548323.

[FRS25] A. Frommer, M. Rinelli, and M. Schweitzer. **“Analysis of Stochastic Probing Methods for Estimating the Trace of Functions of Sparse Symmetric Matrices”.** In: *Mathematics of Computation* 94.352 (2025), pp. 801–823. DOI: 10.1090/mcom/3984.

[Gir87] D. A. Girard. ***Un algorithme simple et rapide pour la validation croisée généralisée sur des problèmes de grande taille.*** Tech. rep. 665-M. IMAG, Grenoble, 1987.

[GM10] G. H. Golub and G. Meurant. ***Matrices, Moments and Quadrature with Applications.*** Princeton University Press, 2010.

[HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp. **“Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions”.** In: *SIAM Review* 53.2 (2011), pp. 217–288. DOI: 10.1137/090771806.

[HT22] E. Hallman and D. Troester. **“A Multilevel Approach to Stochastic Trace Estimation”.** In: *Linear Algebra and its Applications* 638 (2022). DOI: 10.1016/j.laa.2021.12.010.

[Hut89] M. F. Hutchinson. **“A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines”.** In: *Communications in Statistics—Simulation and Computation* 18.3 (1989), pp. 1059–1076. DOI: 10.1080/03610918908812806.

[MAS26] M. Madhavan, A. Alexanderian, and A. K. Saibaba. **“FlexTrace: Exchangeable Randomized Trace Estimation for Matrix Functions”.** In: (2026). arXiv: 2603.05721 [math.NA].

[Mey+21] R. A. Meyer et al. **“Hutch++: Optimal Stochastic Trace Estimation”.** In: *Symposium on Simplicity in Algorithms (SOSA)*. SIAM, 2021, pp. 142–155. DOI: 10.1137/1.9781611976496.16.

[PCK22] D. Persson, A. Cortinovis, and D. Kressner. **“Improved Variants of the Hutch++ Algorithm for Trace Estimation”.** In: *SIAM Journal on Matrix Analysis and Applications* 43.3 (2022), pp. 1162–1185. DOI: 10.1137/21M1447623.

[RV13] M. Rudelson and R. Vershynin. **“Hanson–Wright Inequality and Sub-Gaussian Concentration”.** In: *Electronic Communications in Probability* 18.82 (2013), pp. 1–9. DOI: 10.1214/ECP.v18-2865.

[UCS17] S. Ubaru, J. Chen, and Y. Saad. **“Fast Estimation of $\mathrm{tr}(f(A))$ via Stochastic Lanczos Quadrature”.** In: *SIAM Journal on Matrix Analysis and Applications* 38.4 (2017), pp. 1075–1099. DOI: 10.1137/16M1104974.