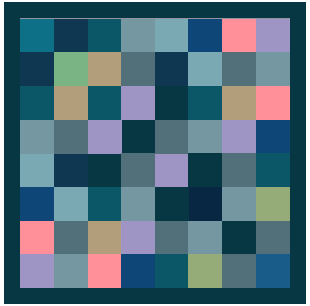
Analysis of stochastic Lanczos quadrature for spectrum approximation

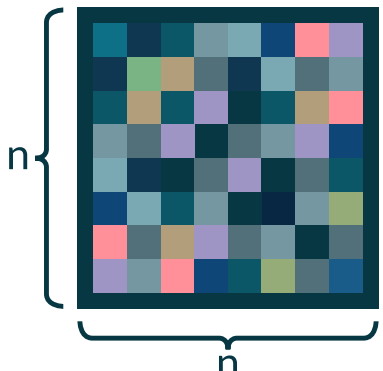
Tyler Chen, Thomas Trogdon, and Shashanka Ubaru

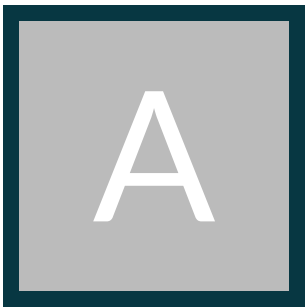
Tyler Chen

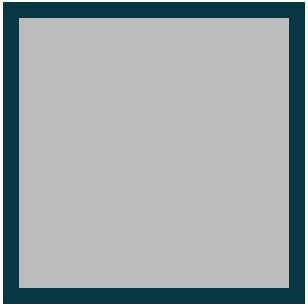
Thomas Trogdon

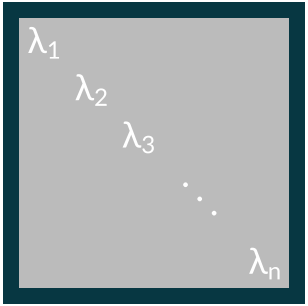
Shashanka Ubaru




n
n


A




λ_1
λ_2
λ_3
$\ddots$
λ_n

Goal: approximate spectrum

Traditionally, approximate a few eigenvalues to high accuracy.

Local spectrum approximation via power method, vanilla Lanczos, etc.

CESM contains **all** information about the spectrum

CESM contains **all** information about the spectrum
and gives a natural target for a **global** approximation

Goal: approximate spectrum globally

The **Wasserstein distance** between distribution functions μ and ν is

$$d_{\mathrm{W}}(\mu, \nu) = \int |\mu(x) - \nu(x)| \mathrm{d}x.$$

Goal: Find approximation to Φ with small Wasserstein distance, accessing $\mathbf{A}$ only through matrix-vector products.

The **cumulative emperical spectral measure** (CESM) of $\mathbf{A}$ is the distribution function

$$\Phi(x) = \sum_{j=1}^{n} \frac{1}{n} \mathbb{1}[\lambda_j \leq x].$$

The **weighted CESM** of $\mathbf{A}$ and $\mathbf{v}_i$ is the distribution function

$$\Psi_i(x) = \sum_{j=1}^{n} [w_i]_j \, \mathbb{1}[\lambda_j \leq x]$$

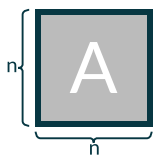
where $[w_i]_i$ is the square of the projection of $\mathbf{v}_i$ onto the i-th eigenvector of $\mathbf{A}$.

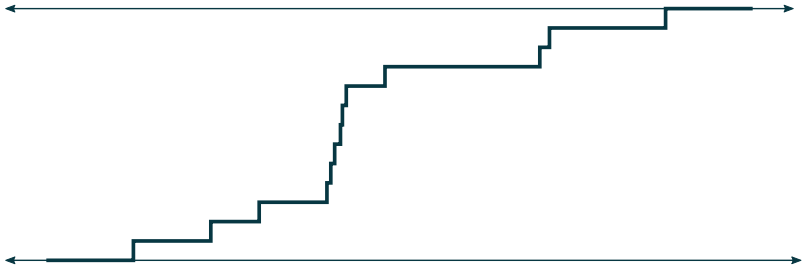
The **SLQ output** is the average of the Gaussian quadrature approximations to the weighted CESMs,

$$\left\langle [\Psi_i]_k^{\mathrm{gq}} \right\rangle := \sum_{i=1}^{n_{\mathrm{v}}} \frac{1}{n_{\mathrm{v}}} [\Psi_i]_k^{\mathrm{gq}}.$$

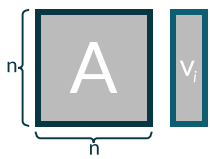
Questions: For good Wasserstein approximation

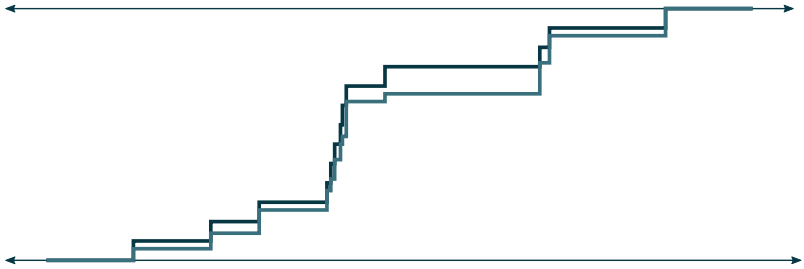
- how many samples do we need?
- how many Lanczos iterations for each sample?



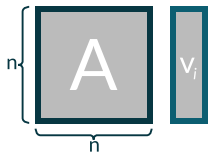


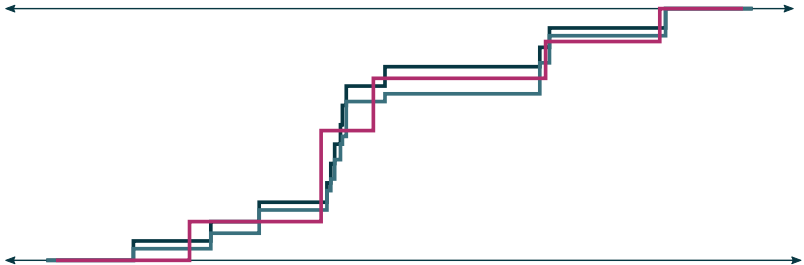
$$\Phi(x) = \sum_{j=1}^{n} \frac{1}{n} \mathbb{1}[\lambda_j \leq x]$$





$$\Phi(x) = \sum_{j=1}^{n} \frac{1}{n} \mathbb{1}[\lambda_j \leq x] \qquad \Psi_i(x) = \sum_{j=1}^{n} [w_i]_j \mathbb{1}[\lambda_j \leq x]$$





$$\Phi(x) = \sum_{j=1}^{n} \frac{1}{n} \mathbb{1}[\lambda_j \leq x] \qquad \Psi_i(x) = \sum_{j=1}^{n} [w_i]_j \mathbb{1}[\lambda_j \leq x] \qquad [\Psi_i]_k^{\mathrm{gq}}(x) = \sum_{j=1}^{k} [d_i]_j \mathbb{1}[[\theta_i]_j \leq x]$$

The main cost of SLQ is running **Lanczos** on $(\mathbf{A}, \mathbf{v}_i)$

If we run k iterations of Lanczos on n_{V} vectors, total number of matrix vector products is $k\, n_{\mathrm{V}}$.

Lanczos requires $O(n)$ storage without reorthogonalization and $O(kn)$ with.

Our analysis assumes an implementation of Lanczos which is close to exact arithmetic; for instance using reorthogonalization.

There is evidence SLQ still seems to work without reorthogonalization, but our analysis isn't necessarily directly applicable in these cases.

Lanczos (and therefore SLQ) only requires access to $\mathbf{A}$ through matrix-vector products!

$$\mathbb{1}[\texttt{true}] = 1 \qquad\qquad \mathbb{1}[\texttt{false}] = 0$$

$$\Phi(x) = \sum_{j=1}^{n} \frac{1}{n} \mathbb{1}[\lambda_j \leq x] = n^{-1} \operatorname{tr}(\mathbb{1}[\mathbf{A} \leq x])$$

$$\Psi_i(x) = \sum_{j=1}^{n} [w_i]_j \mathbb{1}[\lambda_j \leq x] = \mathbf{v}_i^\top \mathbb{1}[\mathbf{A} \leq x] \mathbf{v}_i$$

SLQ overview:

- sample $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n_\mathrm{v}}$ iid from unit hypersphere
- for each $\mathbf{v}_i$, approximate weighted CESM $\boldsymbol{\Psi}_i$ by $[\boldsymbol{\Psi}_i]_k^{\mathrm{gq}}$, obtained via k iterations of Lanczos
- Output average $\langle [\boldsymbol{\Psi}_i]_k^{\mathrm{gq}} \rangle = \sum_{i=1}^{n_\mathrm{v}} \frac{1}{n_\mathrm{v}} [\boldsymbol{\Psi}_i]_k^{\mathrm{gq}}$

Remark. The moments of the k-point Gaussian quadrature rule $[\Psi_i]_k^{\mathrm{gq}}$ are equal to those of Ψ_i through degree $2k-1$.

Theorem. The k-point Gaussian quadrature rule $[\Psi_i]_k^{\mathrm{gq}}$ for Ψ_i can be computed using k iterations of the Lanczos algorithm on $\mathbf{A}$ and $\mathbf{v}_i$.

Remark. The runtime for k iterations of the Lanczos algorithm on $\mathbf{A}$ and $\mathbf{v}_i$ is $O(k\,(T_{\text{mv}} + n))$

Remark. The runtime for k iterations of the Lanczos algorithm on $\mathbf{A}$ and $\mathbf{v}_i$ is $O(k\,(T_{\text{mv}} + n))$, and the required storage is $O(n)$ without reorthogonalization and $O(nk)$ with.

$$
\begin{aligned}
d_{\mathrm{W}}(\Phi, \langle [\Psi_i]_k^{\mathrm{gq}} \rangle) &\leq d_{\mathrm{W}}(\Phi, \langle \Psi_i \rangle) + d_{\mathrm{W}}(\langle \Psi_i \rangle, \langle [\Psi_i]_k^{\mathrm{gq}} \rangle) \\
&\leq d_{\mathrm{W}}(\Phi, \langle \Psi_i \rangle) + \langle d_{\mathrm{W}}(\Psi_i, [\Psi_i]_k^{\mathrm{gq}}) \rangle.
\end{aligned}
$$

Lemma. Let $m = n\,\Phi(x)$. Then,

$$\Psi_i(x) \sim \text{Beta}\left(\frac{m}{2}, \frac{n-m}{2}\right).$$

The uniform distribution on the unit hypersphere is unitarily invariant. Let $\mathbf{U}$ be the eigenvectors of $\mathbf{A}$. We may therefore assume

$$\mathbf{U}^\top \mathbf{v}_i \overset{\mathrm{d}}{=} \frac{\mathbf{x}}{\|\mathbf{x}\|},$$

where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Thus, the $[w_i]_j$ have joint distribution given by,

$$[w_i]_j \overset{\text{d}}{=} \left(\frac{[\mathbf{x}]_j}{\|\mathbf{x}\|} \right)^2 = \frac{([\mathbf{x}]_j)^2}{([\mathbf{x}]_1)^2 + \cdots + ([\mathbf{x}]_n)^2},$$

for $j = 1, \ldots, n$, where $[\mathbf{x}_i]$ are iid standard normals.

For independent chi-square random variables $Y \sim \chi^2_\alpha$ and $Z \sim \chi^2_\beta$,

$$\frac{Y}{Y+Z} \sim \text{Beta}\left(\frac{\alpha}{2}, \frac{\beta}{2}\right).$$

Let $Y = \sum_{j=1}^{m}([\mathbf{x}]_j)^2$ and $Z = \sum_{j=m+1}^{n}([\mathbf{x}]_j)^2$. Then

$$\Phi(x) = \sum_{j=1}^{m}[w_i]_j = \frac{([\mathbf{x}]_1)^2 + \cdots + ([\mathbf{x}]_m)^2}{([\mathbf{x}]_1)^2 + \cdots + ([\mathbf{x}]_n)^2} = \frac{Y}{Y+Z}.$$

Theorem. Let $X \sim \mathrm{Beta}(\alpha, \beta)$. Then $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ and X is $4(\alpha + \beta + 1)$-sub-Gaussian.

Lemma. Suppose X is σ^2-sub-Gaussian. Let $X_1, \dots, X_{n_\vee}$ be iid samples of X. Then for all $t > 0$,

$$\mathbb{P}[|\langle X_i \rangle - \mathbb{E}[X]| > t] \leq 2 \exp\left(-\frac{n_\vee}{2\sigma^2} t^2\right).$$

Theorem. For all $t > 0$,

$$\max_x \mathbb{P}\Big[|\Phi(x) - \langle \Psi_i(x)\rangle| > t\Big] \le 2\exp(-n_{\mathrm{v}}(n+2)t^2)$$

$$\mathbb{P}\Big[\max_x |\Phi(x) - \langle \Psi_i(x)\rangle| > t\Big] \le 2n\exp(-n_{\mathrm{v}}(n+2)t^2).$$

Let $I[\mathbf{A}] = \lambda_{\max} - \lambda_{\min}$. Then,

$$\begin{aligned} d_{\mathrm{W}}(\Phi, \langle \Psi_i \rangle) &= \int |\Phi(x) - \langle \Psi_i(x) \rangle| \mathrm{d}x \\ &\leq I(\mathbf{A}) \max_x |\Phi(x) - \langle \Psi_i(x) \rangle|. \end{aligned}$$

Given $0 < \eta < 1$, if $n_{\mathrm{v}} \geq 4(n+2)t^{-2}\ln(2n\eta^{-1})$,

$$\mathbb{P}\Big[d_{\mathrm{W}}(\Phi, \langle \Psi_i \rangle) > tI[\mathbf{A}]/2\Big] \leq \eta.$$

Proposition. If μ and ν are supported on (a, b) and have equal moments through degree s, then

$$d_{\mathrm{W}}(\mu, \nu) \leq 12(b-a)s^{-1}.$$

Therefore, if $k > 12t^{-1} + \frac{1}{2}$,

$$\langle d_{\mathrm{W}}(\Psi_i, [\Psi_i]_k^{\mathrm{gq}}) \rangle \leq d_{\mathrm{W}}(\Psi_i, [\Psi_i]_k^{\mathrm{gq}}) \leq tI[\mathbf{A}]/2.$$

Theorem. Given $0 < \eta < 1$ and $t > 0$, set $n_{\mathrm{v}} \geq 4(n+2)t^{-2}\ln(2n\eta^{-1})$ and $k > 12t^{-1} + \frac{1}{2}$. Then, using $k\, n_{\mathrm{v}}$ matrix-vector products, SLQ outputs an approximation $\langle [\Psi_i]_k^{\mathrm{gq}} \rangle$ to Φ satisfying,

$$\mathbb{P}\left[d_{\mathrm{W}}(\Phi, \langle [\Psi_i]_k^{\mathrm{gq}} \rangle) > tI[\mathbf{A}] \right] \leq \eta.$$

Theorem. If μ and ν are supported on (a, b) and have equal moments through degree s, then the function $\mu - \nu$ changes sign at least s times on (a, b).

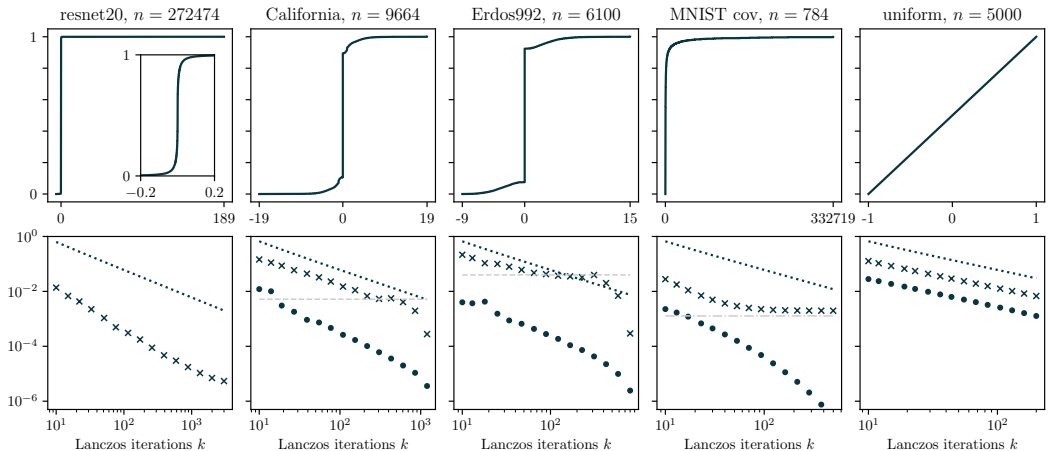
Since $[\Psi_i]_k^{\mathrm{gq}}$ and Ψ_i share $2k-1$ moments, then $[\Psi_i]_k^{\mathrm{gq}} - \Psi_i$ changes sign at least $2k-1$ times.

The Gaussian quadrature $[\Psi_i]_k^{\mathrm{gq}}$ is piecewise constant except at k points of increase

The Gaussian quadrature $[\Psi_i]_k^{\mathrm{gq}}$ is piecewise constant except at k points of increase, and both Ψ_i and $[\Psi_i]_k^{\mathrm{gq}}$ are weakly increasing.

Thus, the only possible points of sign change in $[\mathbf{\Psi}_i]_k^{\mathrm{gq}} - \mathbf{\Psi}_i$ are at the points of increase and constant regions of $[\mathbf{\Psi}_i]_k^{\mathrm{gq}}$

Therefore, there is a sign change in $[\Psi_i]_k^{\mathrm{gq}} - \Psi_i$ at every point of increase and every constant region of $[\Psi]_k^{\mathrm{gq}}$.


resnet20, $n = 272474$
California, $n = 9664$
Erdos992, $n = 6100$
MNIST cov, $n = 784$
uniform, $n = 5000$
1
0
−0.2
0.2
189
-19
19
-9
15
332719
-1
10^0
10^{-2}
10^{-4}
10^{-6}
10^1
10^2
10^3
Lanczos iterations k