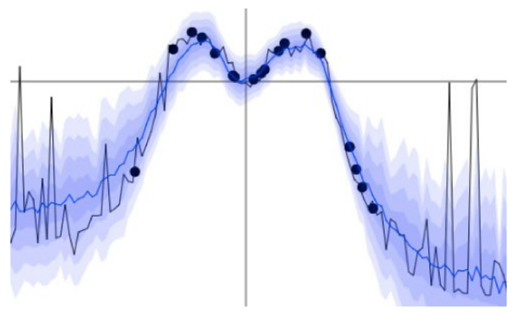
Stochastic Chebyshev Gradient Descent for Spectral Optimization

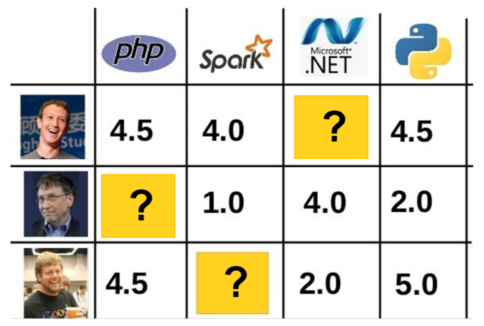
Insu Han[1], Haim Avron[2], and Jinwoo Shin[1]

[1]Korea Advanced Institute of Science and Technology (KAIST) and [2]Tel Aviv University

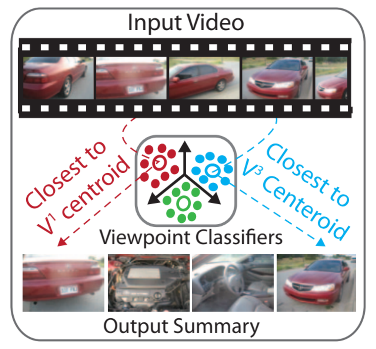
Problems of Interest



(a) regression



(b) recommender system



(c) video summarization

A variety of machine learning applications are involved in matrix spectral functions.

Definition of Spectral-sums

For a scalar function $f : \mathbb{R} \to \mathbb{R}$, **spectral-sums** is defined as

$$\sum_{i=1}^{d} f(\lambda_i) = \mathtt{tr}\left(f(A)\right),$$

where $\lambda_1, \lambda_2, \dots, \lambda_d$ are eigen (or singular) values of a symmetric $A \in \mathbb{R}^{d \times d}$.
Some examples :

- If $f(x) = \log x$, it is the log-determinant ($\to$ Gaussian process regression)
- If $f(x) = x^{-1}$, it is the trace of inverse ($\to$ the second-order optimization)
- If $f(x) = x^p$, it is the Schatten norm ($\to$ matrix completion for recommendation)
- if $f(x) = x \log x$, it is the Von-Neumann entropy ($\to$ quantum state tomography)
- If $f(x) = \exp(x)$, it is the Estrada index ($\to$ social network centrality)

Problems and Contributions

Challenges in spectral optimization:

approximating spectral sums	**optimizing** spectral sums
$\mathtt{tr}\left(f(A(\theta))\right) = \sum_i f(\lambda_i) \approx ?$	$\min_\theta \mathtt{tr}\left(f(A(\theta))\right)$

Both problems have at least $O(d^3)$ computational complexity for a $d \times d$ matrix.

Our contributions are following:

- **[Past works]** We developed a fast algorithm for **approximating** spectral-sums of large-scale matrices with rigorous provable guarantee.
- **[In this work]** We propose a fast unbiased gradient estimator for **optimizing** spectral-sums that guarantees to converge to the optimal.

Approximating Spectral-sums

- For appropriate random $\mathbf{v} \in \mathbb{R}^d$, we have for every B: $\mathtt{tr}(B) = \mathrm{E}\left[\mathbf{v}^\top B \mathbf{v}\right]$
  Generate M Rademacher random vectors $\mathbf{v}_1, \dots, \mathbf{v}_M \in \{-1, 1\}^d$ and estimate
  $$\mathtt{tr}\left(f(A)\right) \approx \frac{1}{M}\sum_{i=1}^{M} \mathbf{v}_i^\top f(A) \mathbf{v}_i.$$
- Truncated Chebyshev expansion of f: $f(x) = \sum_{j=0}^{\infty} b_j T_j(x) \approx \sum_{j=0}^{n} b_j T_j(x)$
  $$\mathtt{tr}\left(f(A)\right) \approx \frac{1}{M}\sum_{i=1}^{M}\sum_{j=0}^{n} b_j \mathbf{v}_i^\top T_j(A) \mathbf{v}_i$$
  where $T_j(A)\mathbf{v}_i$ can be computed efficiently using recursion.
- The overall running time is $O\left(M \times n \times \text{cost for multiplications } A\right)$.

Optimizing Spectral-sums

For optimization, the gradient-based methods are commonly used:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathtt{tr}\left(f(A(\theta))\right) \qquad (\eta : \text{step-size})$$

- Computing $\nabla_\theta \mathtt{tr}\left(f(A(\theta))\right)$ requires matrix decompositions with $O(d^3)$ costs.
- One can use stochastically approximate the gradient (with random $\mathbf{v}$) as
  $$\theta \leftarrow \theta - \eta \nabla_\theta \mathbf{v}^\top p_n(A(\theta))\mathbf{v}.$$
- It can be computed efficiently using matrix-vector multiplications with A and $\partial A / \partial \theta$ (details omitted), thus the complexity reduces to $O(d^2)$.
- With stochastic gradients, we can use SGD, SVRG, etc.

Critical issue: biased gradient estimator

- Even if the gradient estimate $\nabla_\theta \mathbf{v}^\top p_n(A)\mathbf{v}$ is fast and accurate, it is biased:
  $$\mathrm{E}\left[\nabla_\theta \mathbf{v}^\top p_n(A)\mathbf{v}\right] = \nabla_\theta \mathtt{tr}\left(p_n(A)\right) \neq \nabla_\theta \mathtt{tr}\left(f(A)\right)$$
  $$f(x) - p_n(x) \neq 0.$$
- The bias slows the convergence since errors accumulate over iterations.

Randomized Chebyshev Expansions

So far, we deterministically choose the truncation degree:

$$f(x) = \sum_{j=0}^{\infty} b_j T_j(x), \qquad p_n(x) := \sum_{j=0}^{n} b_j T_j(x).$$

Our proposal: randomly sample degree n with probability q_n and define

$$\widehat{p}_n(x) := \sum_{j=0}^{n} \frac{b_j}{1 - \sum_{i=0}^{j-1} q_i} T_j(x).$$

- We get an unbiased estimator: $\mathrm{E}\left[\widehat{p}_n(x)\right] = f(x)$.
- SGD for spectral-sums: $\theta \leftarrow \theta - \eta \nabla_\theta \mathbf{v}^\top \widehat{p}_n(A(\theta))\mathbf{v}$ (random $\mathbf{v}$ and n).
- Under mild assumptions on q_n, an estimator with small variance leads to faster convergence.

Optimal Degree Distribution for Minimum Variance

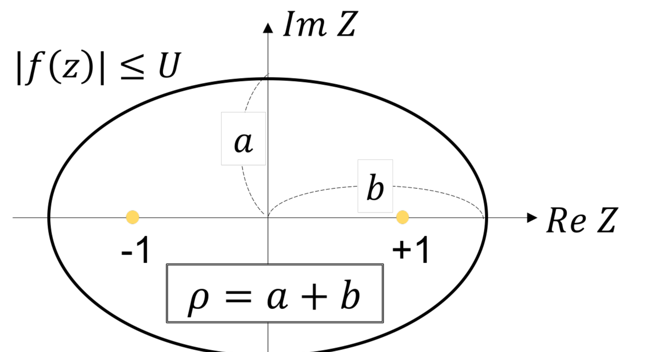
We aim to minimize the Chebyshev weighted variance:

$$\min_{\{q_n : n \geq 0\}} \mathrm{Var}\left[\widehat{p}_n\right] := \mathrm{E}\left[\int_{-1}^{1} \frac{\left(\widehat{p}_n(x) - f(x)\right)^2}{\sqrt{1 - x^2}} dx\right]. \qquad (1)$$

with constraint the average degree $\mathrm{E}[n]$ is given by N.

Theorem (Han, Avron and Shin, 2018). *Suppose analytic function f is $|f(z)| \leq U$ and bounded by ellipse with foci $+1, -1$ and sum of major and minor semi-axis lengths equals to $\rho > 1$. Let $k = \min\{N, \left\lfloor \frac{\rho}{\rho - 1} \right\rfloor\}$, then the distribution that minimizes the variance (1) is:*

$$q_n^* = \begin{cases} 0 & \text{for } n < N - k \\ 1 - \dfrac{k(\rho - 1)}{\rho} & \text{for } n = N - k \\ \dfrac{k(\rho - 1)^2}{\rho^{n+1}} & \text{for } n > N - k \end{cases}$$



In short: the optimal distribution q_n^* minimizes the variance of unbiased estimator.

Algorithm and Analysis

We consider general spectral-sums optimization:

$$\min_{\theta \in \mathcal{C}} \mathtt{tr}\left(f(A(\theta))\right) + g(\theta)$$

where $\mathcal{C}$ is a parameter space and g is some simple function. For analysis, we assume

1. All eigenvalues of $A(\theta)$ for $\theta \in \mathcal{C}$ are bound in some interval,
2. The objective is α-strongly convex and continuous function of θ,
3. $A(\theta)$ is L_A-Lipschitz for $\|\cdot\|_F$, $L_{\mathtt{nuc}}$-Lipschitz for $\|\cdot\|_{\mathtt{nuc}}$, and $g(\theta)$ is L_g-Lipschitz and β_g-smooth.

Algorithm 1. Stochastic gradient descent (SGD) with random $\{\mathbf{v}_i\}_{i=1}^M$ and $n \sim q_n$.

Theorem (Han, Avron and Shin, 2018). *Let $\theta^{(t)} \in \mathbb{R}^{d'}$ be the parameter after t updates. If one chooses the step-size $\eta_t = 1/\alpha t$, then it holds that*

$$\mathrm{E}\left[\left\|\theta^{(T)} - \theta^*\right\|_2^2\right] \leq \frac{4}{\alpha^2 T} \max\left(L_g^2, \left(\frac{2L_A^2}{M} + d' L_{\mathtt{nuc}}^2\right)\left(C_1 + \frac{C_2 N^4}{\rho^{2N}}\right)\right)$$

where $C_1, C_2 > 0$ are constants independent of M, N, and θ^* is the global optimum.

In short: the optimal q_n^* makes small variance and we can bound the error.

Algorithm 2. Stochastic variance reduction gradient (SVRG) with $\{\mathbf{v}_i\}_{i=1}^M$ and n.

Theorem. *Let $\beta^2 = 2\beta_g^2 + \left(\frac{L_A^4 + \beta_A^2}{M} + L_A^4\right)\left(D_1 + \frac{D_2 N^8}{\rho^{2N}}\right)$ for some constants $D_1, D_2 > 0$ independent of M, N. Choose $\eta = \frac{\alpha}{7\beta^2}$ and $T \geq 25\beta^2/\alpha^2$. Then, it holds*

$$\mathrm{E}\left[\left\|\widetilde{\theta}^{(S)} - \theta^*\right\|_2^2\right] \leq r^S \mathrm{E}\left[\left\|\theta^{(0)} - \theta^*\right\|_2^2\right],$$

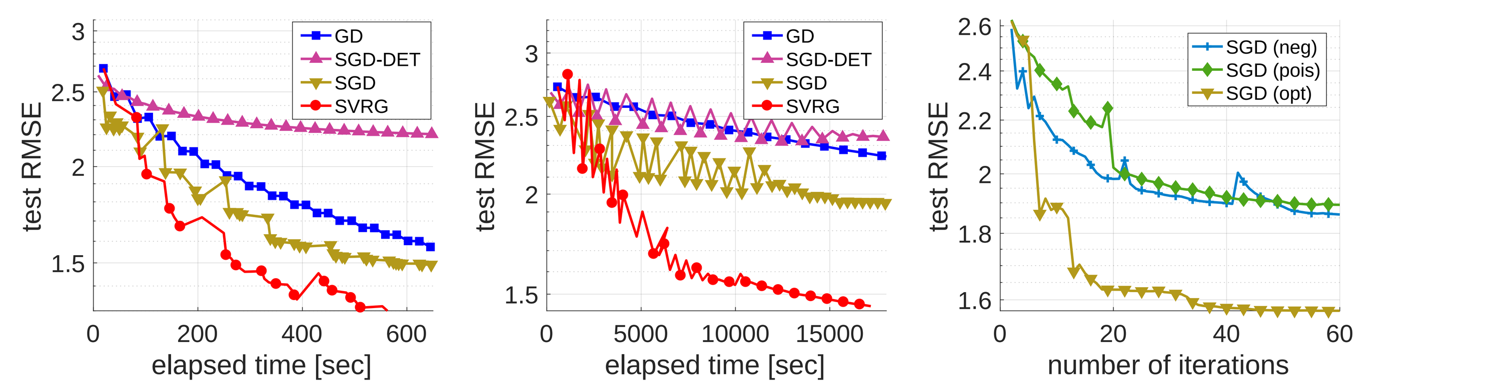
where $0 < r < 1$ is some constant.

In short: the optimal q_n^* with variance reduction yields better convergence rate.

Experiments

1. Schatten norm minimization for matrix completion under MovieLens 1M/10M.

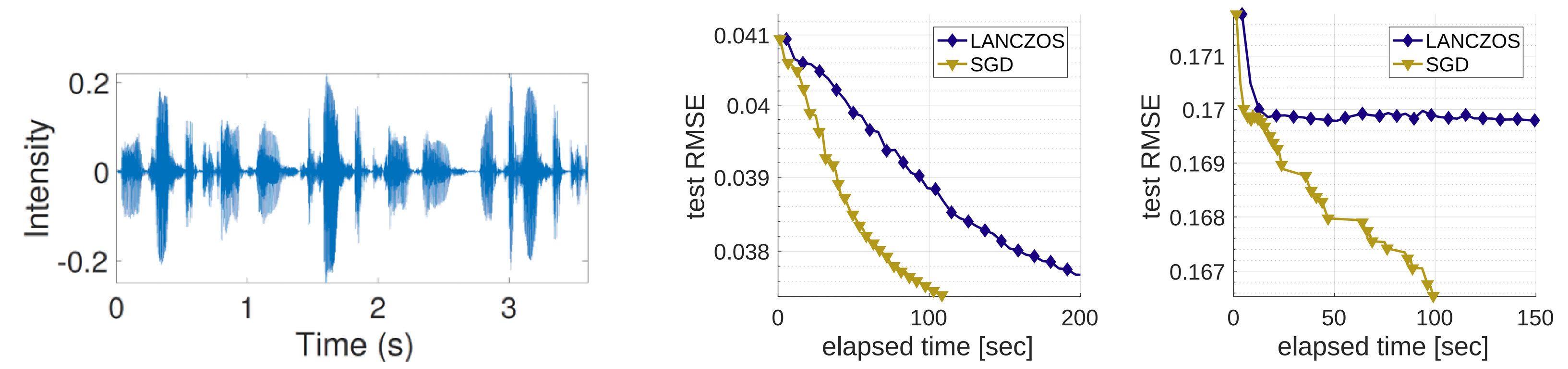
$$\min_{\theta \in \mathbb{R}^{d_1 \times d_2}} \mathtt{tr}(\sqrt{\theta^\top \theta}) + \lambda \sum_{(i,j) \in \mathtt{ENTRIES}} (\theta_{i,j} - B_{i,j})^2$$



- **Methods**: exact gradient descent (GD), deterministic Chebyshev expansion (SGD-DET), randomized approximation (SGD) and SVRG (best and up to 6 times faster than others).
- Optimal q_n^* shows much faster convergence than other distributions.

2. Log-determinant maximization for Gaussian process under sound/humid data

$$\min_\theta - \log\det \mathrm{Kernel}(X, \theta) + \mathbf{y}^\top \mathrm{Kernel}(X, \theta)\mathbf{y}$$



- LANCZOS (state-of-the-art) can be often stuck at alocal optimum, while SGD is more favorable to avoid it.