Stochastic Chebyshev Gradient Descent for Spectral Optimization

Insu Han\textsuperscript{1} \quad Haim Avron\textsuperscript{2} \quad Jinwoo Shin\textsuperscript{1}

\textsuperscript{1}Korea Advanced Institute of Science and Technology (KAIST)
\textsuperscript{2}Tel Aviv University

NeurIPS 2018 Motreal
Spectral Optimization

• For a scalar function $f : \mathbb{R} \rightarrow \mathbb{R}$ and matrix $A \in \mathbb{R}^{d \times d}$, spectral-sum is defined as:

\[
\Sigma_f(A) := \sum_{i=1}^{d} f(\lambda_i) = \text{tr}(f(A)),
\]

$\lambda_1, \lambda_2, \ldots, \lambda_d$ : eigenvalues of $A$
Spectral Optimization

- For a scalar function $f : \mathbb{R} \to \mathbb{R}$ and matrix $A \in \mathbb{R}^{d \times d}$, the spectral-sum is defined as:

\[
\Sigma_f(A) := \sum_{i=1}^{d} f(\lambda_i) = \text{tr}(f(A)),
\]

$\lambda_1, \lambda_2, \ldots, \lambda_d$ : eigenvalues of $A$

- If $f(x) = \log x$, it is the log-determinant
- If $f(x) = x^{-1}$, it is the trace of inverse
- If $f(x) = x^p$, it is the Schatten norm (the nuclear norm is the case $p = 1$)
- If $f(x) = x \log x$, it is the von-Neumann entropy
- If $f(x) = \exp(x)$, it is the Estrada index
- If $f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$, it is rank or testing positive definiteness
Spectral Optimization

• For a scalar function \( f : \mathbb{R} \to \mathbb{R} \) and matrix \( A \in \mathbb{R}^{d \times d} \), spectral-sum is defined as:

\[
\Sigma_f(A) := \sum_{i=1}^{d} f(\lambda_i) = \text{tr}(f(A)),
\]

\( \lambda_1, \lambda_2, \ldots, \lambda_d \) : eigenvalues of \( A \)

• **Goal**: solve the optimization

\[
\min_{\theta} \Sigma_f(A(\theta)) + g(\theta)
\]

\( A(\theta) \) is a parameterized symmetric matrix, \( g \) is a simple function.

\( \Box \text{easy to compute } g, \nabla g \)
Challenges

• Gradient-based methods:

\[
\theta \leftarrow \theta - \eta \nabla \theta \left( \sum_f (A(\theta)) + g(\theta) \right)
\]

easy to compute

• Computing exact \( \nabla \theta \sum_f (A(\theta)) \) requires \( O(d^3) \) operations, \( d \) : matrix dimension
Challenges

• Gradient-based methods:
  \[ \theta \leftarrow \theta - \eta \nabla_\theta \left( \Sigma_f(A(\theta)) + g(\theta) \right) \]
  easy to compute

• Computing exact \( \nabla_\theta \Sigma_f(A(\theta)) \) requires \( \mathcal{O}(d^3) \) operations, \( d \) : matrix dimension

• A recent work [1] can approximate \( \nabla_\theta \Sigma_f(A(\theta)) \) using \( \mathcal{O}(\|A\|_0) \) (# of non-zeros of \( A \))

• But, the gradient estimator is biased, which hurts stable/fast convergence of SGD

Challenges

• Gradient-based methods:

$$\theta \leftarrow \theta - \eta \nabla_\theta \left( \sum_f (A(\theta)) + g(\theta) \right)$$

easy to compute

• Computing exact $\nabla_\theta \sum_f (A(\theta))$ requires $O(d^3)$ operations, $d$ : matrix dimension

• A recent work [1] can approximate $\nabla_\theta \sum_f (A(\theta))$ using $O(\|A\|_0)$ (# of non-zeros of $A$)

• But, the gradient estimator is biased, which hurts stable/fast convergence of SGD

“We propose a fast unbiased gradient estimator with convergence guarantees of SGD/SVRG”

Randomized Chebyshev Expansion

• Why biased? Spectral sums approximation [1] itself is biased since it combines
  (1) randomized trace estimator (unbiased)
  (2) Chebyshev polynomial expansion of $f \approx p_n$ (biased)

\[
\Sigma_f(A(\theta)) = \text{tr}(f(A)) = \mathbb{E}_v [v^\top f(A)v] \approx \mathbb{E}_v [v^\top p_n(A)v]
\]

($v$: random vector)

Randomized Chebyshev Expansion

- Why biased? Spectral sums approximation [1] itself is biased since it combines
  (1) randomized trace estimator (unbiased)
  (2) Chebyshev polynomial expansion of \( f \approx p_n \) (biased)

\[
\sum f(A(\theta)) = \text{tr}(f(A)) = \mathbb{E}_v [v^\top f(A)v] \approx \mathbb{E}_v [v^\top p_n(A)v] \quad (v: \text{random vector})
\]

- To make it unbiased, we consider the following randomized Chebyshev expansions

\[
f(x) = \sum_{j=0}^{\infty} a_j T_j(x), \quad p_n(x) = \sum_{j=0}^{n} a_j T_j(x) \quad n \sim q_n \quad \text{random sampling}
\]

\[
\hat{p}_n(x) = \sum_{j=0}^{n} \frac{a_j}{1 - \sum_{i=0}^{j-1} q_i} T_j(x)
\]

- Then, \( \mathbb{E}_n [\hat{p}_n(x)] = f(x) \) and the gradient estimator with \( \hat{p}_n \) is unbiased 😊

Randomized Chebyshev Expansion

- Why biased? Spectral sums approximation [1] itself is biased since it combines
  (1) randomized trace estimator (unbiased)
  (2) Chebyshev polynomial expansion of $f \approx p_n$ (biased)

$$\Sigma_f(A(\theta)) = \text{tr}(f(A)) = \mathbb{E}_v[v^T f(A)v] \approx \mathbb{E}^\text{biased}_v[v^T p_n(A)v] \quad (v: \text{random vector})$$

- To make it unbiased, we consider the following randomized Chebyshev expansions

$$f(x) = \sum_{j=0}^{\infty} a_j T_j(x), \quad p_n(x) = \sum_{j=0}^{n} a_j T_j(x) \quad \text{random sampling}$$

$$\hat{p}_n(x) = \sum_{j=0}^{n} \frac{a_j}{1 - \sum_{i=0}^{j-1} q_i} T_j(x)$$

- Then, $\mathbb{E}_n[\hat{p}_n(x)] = f(x)$ and the gradient estimator with $\hat{p}_n$ is unbiased 😊

- Question: what is a good distribution $q_n$?

Optimal Degree Distribution

• An estimator with small variance leads to faster convergence.
• **Problem**: minimize the variance of estimator given the expected degree by $N$

$$\min_{q_n} \text{Var}_n[\hat{p}_n] \quad \text{s.t.} \quad E_n[n] = N$$
**Optimal Degree Distribution**

- An estimator with small variance leads to faster convergence.
- **Problem**: minimize the variance of estimator given the expected degree by $N$

$$
\min_{q_n} \text{Var}_n [\hat{p}_n] \quad \text{s.t.} \quad E_n[n] = N
$$

**Theorem [Han, Avron and Shin 2018]**. The optimal degree distribution is

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

- $\rho > 1$ : defined by $f$
- $k = \min\{N, \lfloor \frac{\rho}{\rho-1} \rfloor\}$

- Under the optimal distribution, we prove the convergence guarantees of SGD/SVRG

(see paper for Details)
Experimental Results for Two Applications

1. Matrix completion via **nuclear norm** regularization (left)
2. Gaussian process regression via **log-determinant** optimization (right)

![Graph](image1.png)

MovieLens 10M dataset, $f(x) = x^{1/2}$

![Graph](image2.png)

Szeged Humid dataset, $f(x) = \log x$
Thank you

Stochastic Chebyshev Gradient Descent for Spectral Optimization

Key words: Matrix optimization, Randomized Chebyshev truncation, Variance minimization

Poster # 6
Thursday Dec 6th 5:00 – 7:00 PM
@ Room 210 & 230 AB