Stochastic Chebyshev Gradient Descent for Spectral Optimization

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

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

NeurIPS 2018 Motréal

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 :

$$\Sigma_f(A) := \sum_{i=1}^{d} f(\lambda_i) \;=\; \mathtt{tr}(f(A)),$$

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

Spectral Optimization

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

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

$\lambda_1, \lambda_2, \ldots, \lambda_d$: 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-p 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 :

$$\Sigma_f(A) := \sum_{i=1}^{d} f(\lambda_i) \;=\; \mathtt{tr}(f(A)),$$

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

- Goal : solve the optimization

$$\min_{\theta} \Sigma_f(A(\theta)) + g(\theta)$$

easy to compute g, ∇g

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

- E.g., collaborative filtering, hyperparameter learning and etc.

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

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

- [Han et al., 2017, Dong et al., 2017] can approximate $\nabla_\theta \Sigma_f(A(\theta))$ using $\mathcal{O}(\|A\|_0)$ 😊
- But, the gradient estimator is biased, which hurts stable/fast convergence of SGD 😭

SGD: stochastic gradient descent

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

- [Han et al., 2017, Dong et al., 2017] can approximate $\nabla_\theta \Sigma_f(A(\theta))$ using $\mathcal{O}(\|A\|_0)$ 😊
- 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

SVRG: stochastic variance reduced gradient, SGD: stochastic gradient descent

Randomized Chebyshev Expansion

- Why biased? The prior spectral-sum approximations are **biased** on combining
  (1) randomized trace estimator (unbiased)
  (2) Chebyshev polynomial expansion of $f \approx p_n$ (biased) 😭

$$\Sigma_f(A(\theta)) = \mathtt{tr}(f(A)) = \underbrace{\mathbf{E}_{\mathbf{v}}[\mathbf{v}^\top f(A)\mathbf{v}]}_{\text{unbiased}} \approx \underbrace{\mathbf{E}_{\mathbf{v}}[\mathbf{v}^\top p_n(A)\mathbf{v}]}_{\text{biased estimator}}$$

($\mathbf{v}$: random vector)

Randomized Chebyshev Expansion

- Why biased? The prior spectral-sum approximations are **biased** on combining
  (1) randomized trace estimator (unbiased)
  (2) Chebyshev polynomial expansion of $f \approx p_n$ (biased) 😭

$$\Sigma_f(A(\theta)) = \mathtt{tr}(f(A)) = \underbrace{\mathbf{E}_{\mathbf{v}}[\mathbf{v}^\top f(A)\mathbf{v}]}_{\text{unbiased}} \approx \underbrace{\mathbf{E}_{\mathbf{v}}[\mathbf{v}^\top p_n(A)\mathbf{v}]}_{\text{biased estimator}}$$

($\mathbf{v}$: random vector)

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

$$\boxed{f(x) = \sum_{j=0}^{\infty} a_j T_j(x), \quad p_n(x) = \sum_{j=0}^{n} a_j T_j(x) \xrightarrow[\text{random sampling}]{n \sim q_n} \widehat{p}_n(x) = \sum_{j=0}^{n} \frac{a_j}{1 - \sum_{i=0}^{j-1} q_i} T_j(x)}$$

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

Randomized Chebyshev Expansion

- Why biased? The prior spectral-sum approximations are **biased** on combining
  (1) randomized trace estimator (unbiased)
  (2) Chebyshev polynomial expansion of $f \approx p_n$ (biased) 😭

$$\Sigma_f(A(\theta)) = \mathbf{tr}(f(A)) = \underbrace{\mathbf{E}_{\mathbf{v}}[\mathbf{v}^\top f(A)\mathbf{v}]}_{\text{unbiased}} \approx \underbrace{\mathbf{E}_{\mathbf{v}}[\mathbf{v}^\top p_n(A)\mathbf{v}]}_{\text{biased estimator}}$$

($\mathbf{v}$: random vector)

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

$$\boxed{f(x) = \sum_{j=0}^{\infty} a_j T_j(x), \quad p_n(x) = \sum_{j=0}^{n} a_j T_j(x) \xrightarrow[\text{random sampling}]{n \sim q_n} \widehat{p}_n(x) = \sum_{j=0}^{n} \frac{a_j}{1 - \sum_{i=0}^{j-1} q_i} T_j(x)}$$

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

- <u>Question :</u> 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 N

$$\min_{q_n} \mathrm{Var}_n \left[\widehat{p}_n\right] \qquad \text{s.t.} \quad \mathbf{E}_n[n] = N$$

Optimal Degree Distribution

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

$$\min_{q_n} \mathrm{Var}_n\left[\widehat{p}_n\right] \qquad \text{s.t.} \quad \mathbf{E}_n[n] = N$$

Theorem 1 [Han, Avron and Shin 2018]. The optimal degree distribution is

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

$\rho > 1$: defined by f

$k = \min\{N, \left\lfloor \frac{\rho}{\rho-1} \right\rfloor\}$

Optimal Degree Distribution

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

$$\min_{q_n} \mathrm{Var}_n\left[\widehat{p}_n\right] \qquad \text{s.t.} \quad \mathbf{E}_n[n] = N$$

Theorem 1 [Han, Avron and Shin 2018]. The optimal degree distribution is

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

$\rho > 1$: defined by f

$k = \min\{N, \left\lfloor \frac{\rho}{\rho-1} \right\rfloor\}$

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

Theorem 2 [Han, Avron and Shin 2018].

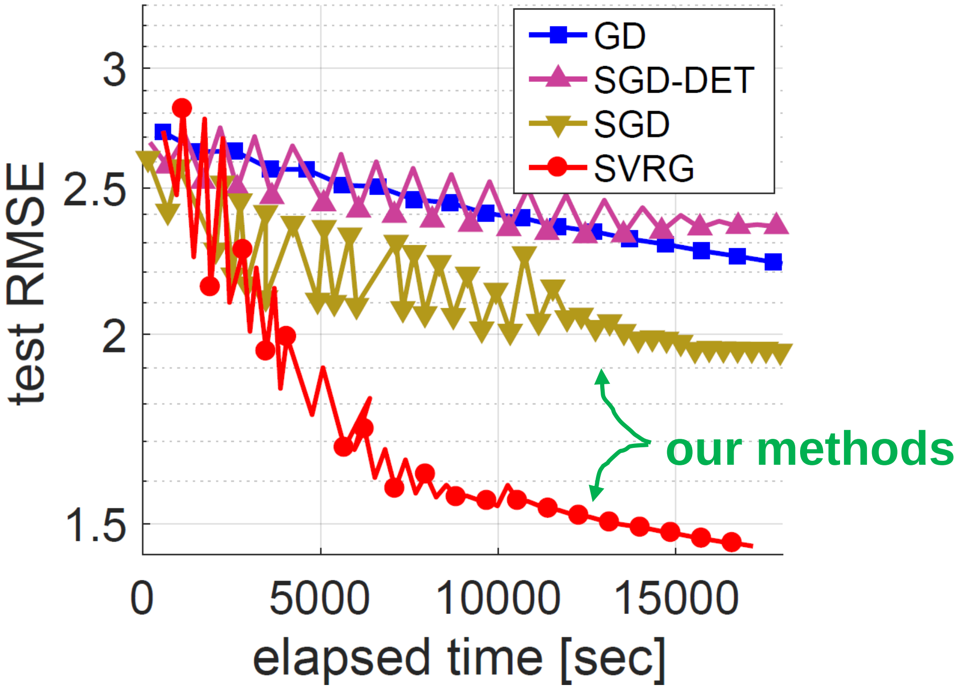
$$\mathbf{E}[\|\theta^* - \theta^{(T)}\|_2^2] \leq \frac{\mathcal{O}(1)}{T}\|\theta^* - \theta^{(0)}\|_2^2$$

θ^* : optimal

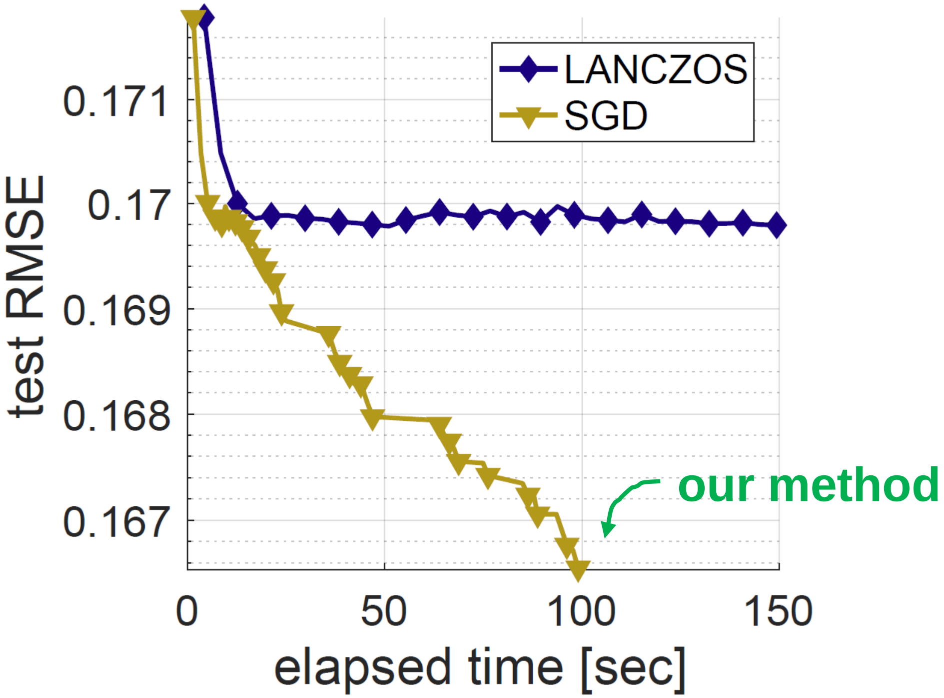
$\theta^{(T)}$: θ in T^{th} iteration of SGD

Experimental Results for Two Applications

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



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



Szeged Humid datset, $f(x) = \log x$

Our algorithms run at least **6 times** faster than other gradient descent methods

Thank you

Stochastic Chebyshev Gradient Descent for Spectral Optimization

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

Poster # 6

Thurday Dec 6^{th} 5:00 – 7:00 PM

@ Room 210 & 230 AB