Stochastic Chebyshev Gradient Descent for Spectral Optimization

Insu Han¹, Haim Avron², and Jinwoo Shin¹

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

Optimizing Spectral-sums

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

\[ \theta - \theta - n \nabla_x \text{tr} (f(A)) \]  

- Computing \( \nabla_x \text{tr} (f(A)) \) requires matrix decompositions with \( O(d^2) \) costs.
- One can use stochastically approximate the gradient (with random \( v \)) as

\[ \theta - \theta - n \nabla_x \text{tr} (\left< f(A), v \right>) \]  

- 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 \( \nabla_v \text{tr} \left( p_n(A) v \right) \) is fast and accurate, it is biased: 

\[ E \left[ \nabla_v \text{tr} \left( p_n(A) v \right) \right] = \nabla_v \text{tr} (f(A)) \neq 0 \]  

- The bias slows the convergence since errors accumulate over iterations.

Randomized Chebyshev Expansions

So far, we deterministically choose the truncation degree:

\[ f(x) = \sum_{j=0}^{d} b_j T_j(x), \quad p_n(x) = \sum_{j=0}^{d} b_j T_j(x). \]

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

\[ \tilde{p}_n(x) = \sum_{j=0}^{\min(n, d)} b_j T_j(x). \]

- We get an unbiased estimator: 

\[ E \left[ \tilde{p}_n(x) \right] = f(x). \]
- SGD for spectral-sums: \( \theta - \theta - n \nabla_v \text{tr} \left( \tilde{p}_n(A) v \right) \) (random \( v \) and \( n \)).
- Under mild assumptions on \( q_n \), an estimator with small variance leads to faster convergence.

Optimal Degree Minimization for Minimum Variance

We aim to minimize the Chebyshev weighted variance:

\[ \min_{n \in \mathbb{N}} \text{Var} \left[ \tilde{p}_n \right] = E \left[ \int_0^1 \left( \tilde{p}_n(x) - f(x) \right)^2 \, dx \right] \]  

with constraint the average degree \( E \left[ n \right] \) is given by \( N \).

Theorem (Han, Avron and Shin, 2018). Suppose analytic function \( f \) is \( |f(x)| \leq U \) and bounded by ellipse with foci \( +1, -1 \) and sum of major and minor semi-axes lengths equals to \( p > 1 \). Let \( k = \min \left[ N, \frac{1}{2} p^2 \right] \), then the distribution that minimizes the variance is given by:

\[ q_n = \begin{cases} 0 & \text{for } n < N - k \\ \frac{1 - k(p - 1)}{p} & \text{for } n = N - k \\ \frac{k(p - 1)^2}{p^2 + 1} & \text{for } n > N - k \end{cases} \]

In short: the optimal \( q_n \) minimizes the variance of unbiased estimator.

Algorithm and Analysis

We consider general spectral-sums optimization:

\[ \min_{\theta} \text{tr} \left( f(A) \theta + g(\theta) \right) \]

where \( C \) is a parameter space and \( g \) is some simple function. For analysis, we assume:

1. All eigenvalues of \( A \) for \( \theta \in C \) are bound in some interval.
2. The objective is \( 1 \)-strongly convex and continuous function of \( \theta \).
3. \( A(\theta) \) is \( L_1 \)-Lipschitz for \( \| \cdot \|_1 \), and \( g(\theta) \) is \( L_2 \)-Lipschitz and \( \beta \)-smooth.

Algorithm 1. Stochastic gradient descent (SGD) with \( \{ \theta_k \}_{k=1}^{\infty} \) and \( \eta_n \rightarrow q_n \).

Theorem (Han, Avron and Shin, 2018). Let \( \theta(0) \in C \) be the parameter after \( t \) updates. If one chooses the step-size \( \eta_n = \eta / n \), then it holds that

\[ \text{E} \left[ \theta(T) - \theta^{*} \right] \leq \frac{4}{\eta^2} \max \left( E \left[ \| \frac{1}{n} \sum_{k=0}^{n-1} f(\theta_k) \|_2^2 \right], \frac{d L_2^2}{\eta^2} \right) \]

where \( C_1, C_2 > 0 \) are constants independent of \( M, N \), \( \eta \) 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 \( \{ \theta_k \}_{k=1}^{\infty} \).

Theorem. Let \( \beta^2 = \frac{2}{\eta^2} \left( \frac{L_2^2}{\eta^2} + \frac{d L_2^2}{\eta^2} \right) \) for some constants \( D_1, D_2 \geq 0 \) independent of \( M, N \). Choose \( \eta = \frac{1}{2 \beta} \) and \( T \geq 2(\beta^2 / \eta)^2 \). Then, it holds

\[ \text{E} \left[ \theta(T) - \theta^{*} \right] \leq \frac{8}{T^2} \left( \frac{\text{E} \left[ \| \theta(0) - \theta^{*} \|_2^2 \right]}{T^2} \right), \]

where \( 0 < \eta / 2 < 1 \) 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.

- 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

- 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.

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