Stochastic Spectral and Conjugate Descent Methods

Dmitry Kovalev

Joint work with P. Richtárik, E. Gorbunov and E. Gasanov

KAUST

March 26, 2019

Paper

This talk is based on paper:

Dmitry Kovalev, Peter Richtarik, Eduard Gorbunov, and Elnur Gasanov.
Stochastic spectral and conjugate descent methods.
In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3358–3367. Curran Associates, Inc., 2018.

Introduction

We consider the following problem:

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2} x^\top \mathbf{A} x - b^\top x \tag{1}$$

- $\mathbf{A}$ – $n \times n$ symmetric positive definite matrix
- Unique solution $x_* = \mathbf{A}^{-1} b$
- n can be huge

Randomized Coordinate Descent (RCD)

Algorithm 1

Parameters: probabilities $p_1, \ldots, p_n > 0$

Initialize: Choose $x_0 \in \mathbb{R}^n$

for $t = 0, 1, 2, \ldots$ **do**

Sample random $i \in [n]$ with probability $p_i > 0$

Set $x_{t+1} = x_t - \frac{\mathbf{A}_{i:}x_t - b_i}{\mathbf{A}_{ii}} e_i$, where e_i – i-th basis vector

end for

Randomized Coordinate Descent (RCD)

Algorithm 1

Parameters: probabilities $p_1, \ldots, p_n > 0$
Initialize: Choose $x_0 \in \mathbb{R}^n$
for $t = 0, 1, 2, \ldots$ **do**
    Sample random $i \in [n]$ with probability $p_i > 0$
    Set $x_{t+1} = x_t - \frac{\mathbf{A}_{i:}x_t - b_i}{\mathbf{A}_{ii}} e_i$, where e_i – i-th basis vector
end for

Theorem (Leventhal & Lewis (2010))

Let probabilities p_i be proportional to diagonal elements $\mathbf{A}_{ii}$. Then the random iterates of Algorithm 1 satisfy $\mathbb{E}\left[\|x_t - x_*\|_{\mathbf{A}}^2\right] \leq \epsilon$ as long as the number of iterations t is at least

$$\mathcal{O}\left(\frac{\mathrm{Tr}\,(\mathbf{A})}{\lambda_{\min}(\mathbf{A})} \log \frac{1}{\epsilon}\right). \tag{2}$$

Sketch and Project

Algorithm 2 (Gower & Richtárik 2015)

Parameter: Distribution $\mathcal{D}$ over vectors in $\mathbb{R}^n$

Initialization: Choose $x_0 \in \mathbb{R}^n$

for $t = 0, 1, 2 \ldots$ **do**

Sample random vector s_t from $\mathcal{D}$

Set $x_{t+1} = x_t - \frac{s_t^\top (\mathbf{A}x_t - b)}{s_t^\top \mathbf{A} s_t} s_t$

end for

Sketch and Project

Algorithm 2 (Gower & Richtárik 2015)

Parameter: Distribution $\mathcal{D}$ over vectors in $\mathbb{R}^n$
Initialization: Choose $x_0 \in \mathbb{R}^n$
for $t = 0, 1, 2 \ldots$ **do**
  Sample random vector s_t from $\mathcal{D}$
  Set $x_{t+1} = x_t - \frac{s_t^\top (\mathbf{A}x_t - b)}{s_t^\top \mathbf{A} s_t} s_t$
end for

Theorem (Gower & Richtárik 2015, Richtárik & Takáč 2017)

Let $\mathbf{H} = \frac{ss^\top}{s^\top \mathbf{A} s}$, $\mathbf{W} = \mathbb{E}_{s\sim\mathcal{D}}\left[\mathbf{A}^{1/2}\mathbf{H}\mathbf{A}^{1/2}\right]$. *Then the random iterates of Algorithm 2 satisfy* $\mathbb{E}\left[\|x_t - x_*\|_{\mathbf{A}}^2\right] \leq \epsilon$ *as long as the number of iterations* t *is at least*

$$\mathcal{O}\left(\frac{1}{\lambda_{\min}(\mathbf{W})} \log \frac{1}{\epsilon}\right). \tag{3}$$

RCD with Arbitrary Probabilities

Applying the previous theorem for RCD with arbitrary probabilities gives the following rate:

$$\mathcal{O}\left(\frac{1}{\lambda_{\min}\left(\mathbf{A}\mathrm{Diag}\left(\frac{p_i}{\mathbf{A}_{ii}}\right)\right)}\log\frac{1}{\epsilon}\right). \tag{4}$$

Uniform Probabilities Can Be Optimal

Theorem

Let $n = 2$ and consider RCD with probabilities $p_1 > 0$ and $p_2 > 0$, $p_1 + p_2 = 1$. Then the choice $p_1 = p_2 = \frac{1}{2}$ optimizes the rate of RCD in (4).

Theorem

Let $n \geq 2$ and let $\mathbf{A}$ be diagonal. Then uniform probabilities ($p_i = \frac{1}{n}$ for all i) optimize the rate of RCD in (4).

Importance Sampling Can Be Unimportant

Diagonal and row-squared-norm probabilities can lead to an arbitrarily worse performance than uniform probabilities:

Theorem

For every $n \geq 2$ and $T > 0$, there exists $\mathbf{A}$ such that:

(i) *The rate of RCD with $p_i \sim \mathbf{A}_{ii}$ is T times worse than the rate of RCD with uniform probabilities.*

(ii) *The rate of RCD with $p_i \sim \|\mathbf{A}_{ii}\|^2$ is T times worse than the rate of RCD with uniform probabilities.*

Optimal Probabilities Can Be Bad

We can't adjust probabilities in (4) to obtain a rate that is independent of matrix $\mathbf{A}$:

Theorem

For every $n \geq 2$ and $T > 0$, there exists $\mathbf{A}$ such that the number of iterations (as expressed by formula (4)) of RCD with any choice of probabilities $p_1, \ldots, p_n > 0$ is $\mathcal{O}(T \log(1/\epsilon))$.

Optimal Probabilities Can Be Bad

We can't adjust probabilities in (4) to obtain a rate that is independent of matrix $\mathbf{A}$:

Theorem

For every $n \geq 2$ and $T > 0$, there exists $\mathbf{A}$ such that the number of iterations (as expressed by formula (4)) of RCD with any choice of probabilities $p_1, \ldots, p_n > 0$ is $\mathcal{O}(T \log(1/\epsilon))$.

Lower bound can also be arbitrarily bad:

Theorem

For every $n \geq 2$ and $T > 0$, there exists an $n \times n$ positive definite matrix $\mathbf{A}$ and starting point x_0, such that the number of iterations of RCD with any choice probabilities $p_1, \ldots, p_n > 0$ is $\Omega(T \log(1/\epsilon))$.

Stochastic Spectral Descent (SSD)

- Algorithm 2 obtains the **optimal** rate

$$\mathcal{O}\left(n \log \frac{1}{\epsilon}\right) \tag{5}$$

  when $\mathcal{D}$ is chosen to be the uniform distribution over the eigenvectors of $\mathbf{A}$. We call this method stochastic spectral descent (SSD).
- The same rate is obtained when $\mathcal{D}$ is chosen to be the uniform distribution over $\mathbf{A}$-orthogonal vectors (i.e. vectors $u_1, \ldots, u_n$ such that $u_i^\top \mathbf{A} u_j = 0$ for all $i \neq j$). We call this method stochastic conjugate descent (SconD).
- SSD is not a practical method due to high preprocessing cost: computation of eigenvectors.

Stochastic Spectral Coordinate Descent (SSCD)

Consider eigenvalue decomposition of $\mathbf{A}$:

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \tag{6}$$

eigenvalues: $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, eigenvectors: $u_1 \ldots, u_n$.

Stochastic Spectral Coordinate Descent (SSCD)

Consider eigenvalue decomposition of $\mathbf{A}$:

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \tag{6}$$

eigenvalues: $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, eigenvectors: $u_1 \ldots, u_n$.

Algorithm 3

Parameter: $k \in \{0, \ldots, n-1\}$
Set $C_k = k\lambda_{k+1} + \sum_{i=k+1}^{n} \lambda_i$
Set $\mathcal{D}$ to be the following distrubution:

$$s = \begin{cases} e_i, & \text{with probability } \frac{\mathbf{A}_{ii}}{C_k}, i = 1, \ldots, n \\ u_i, & \text{with probability } \frac{\lambda_{k+1} - \lambda_i}{C_k}, i = 1, \ldots, k \end{cases}$$

Run Algorithm 2 with distribution $\mathcal{D}$

Stochastic Spectral Coordinate Descent (SSCD)

Theorem

The random iterates of Algorithm 3 satisfy $\mathbb{E}\left[\|x_t - x_*\|_{\mathbf{A}}^2\right] \leq \epsilon$ *as long as the number of iterations* t *is at least*

$$\mathcal{O}\left(\frac{C_k}{\lambda_{k+1}} \log \frac{1}{\epsilon}\right). \tag{7}$$

Moreover the rate of convergence improves as k *grows:*

$$\frac{\mathrm{Tr}\left(\mathbf{A}\right)}{\lambda_{\min}(\mathbf{A})} = \frac{C_0}{\lambda_1} \geq \cdots \geq \frac{C_{n-1}}{\lambda_n} = n. \tag{8}$$

Convergence Rate: Unaffected by k if All Eigenvalues are Tightly Clustered

Convergence rate is unaffected by k if all eigenvalues are tightly clustered:

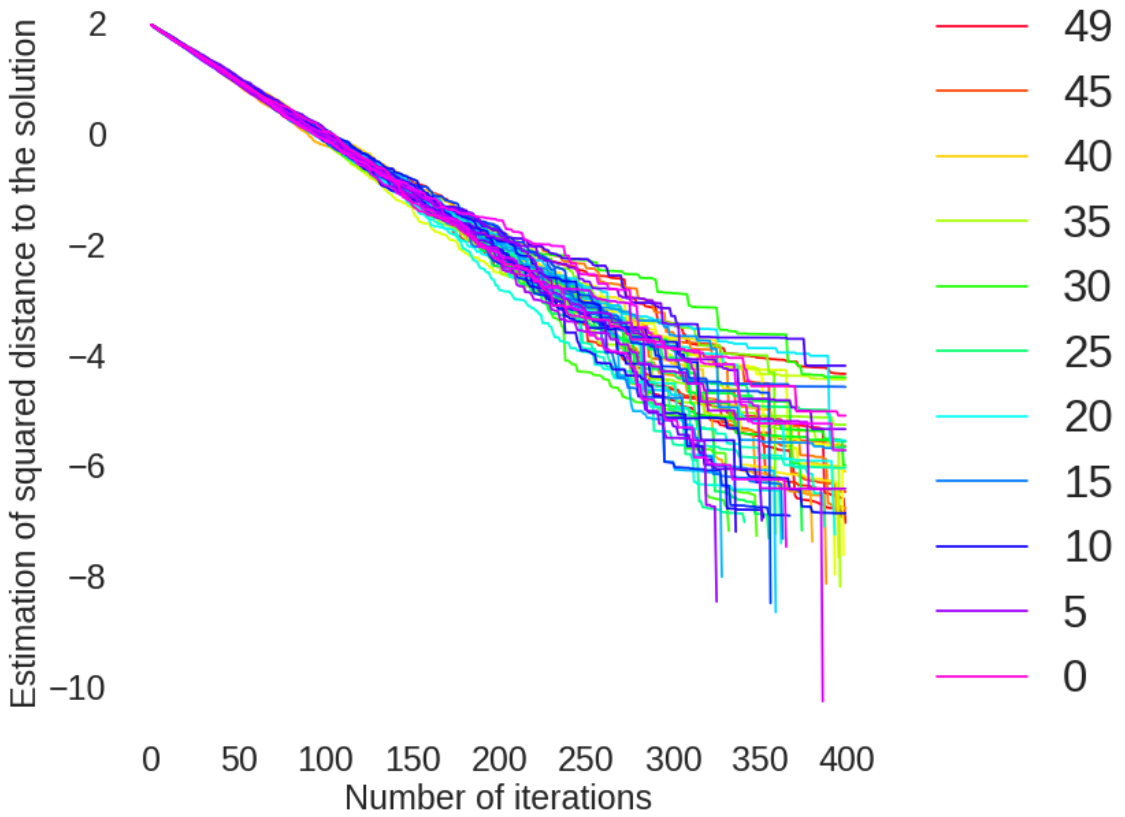

Figure: Eigenvalues were sampled from uniform distribution on $[10; 11]$; $n = 50$

Convergence Rate Improves as k Increases

Convergence rate improves as k increases:

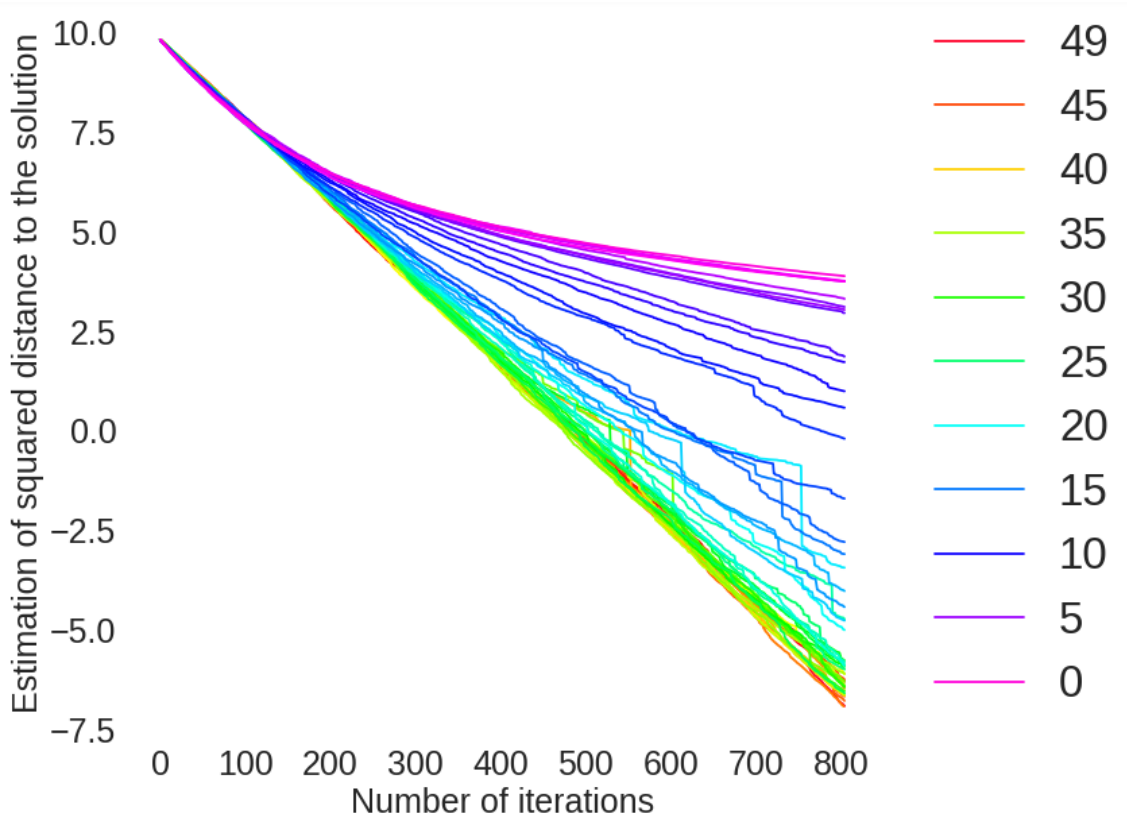

Figure: Eigenvalues were sampled from uniform distribution on $[0; 10^5]$; $n = 50$

Convergence Rate: Phase Transition when k Crosses from One Cluster of Eigenvalues to Another

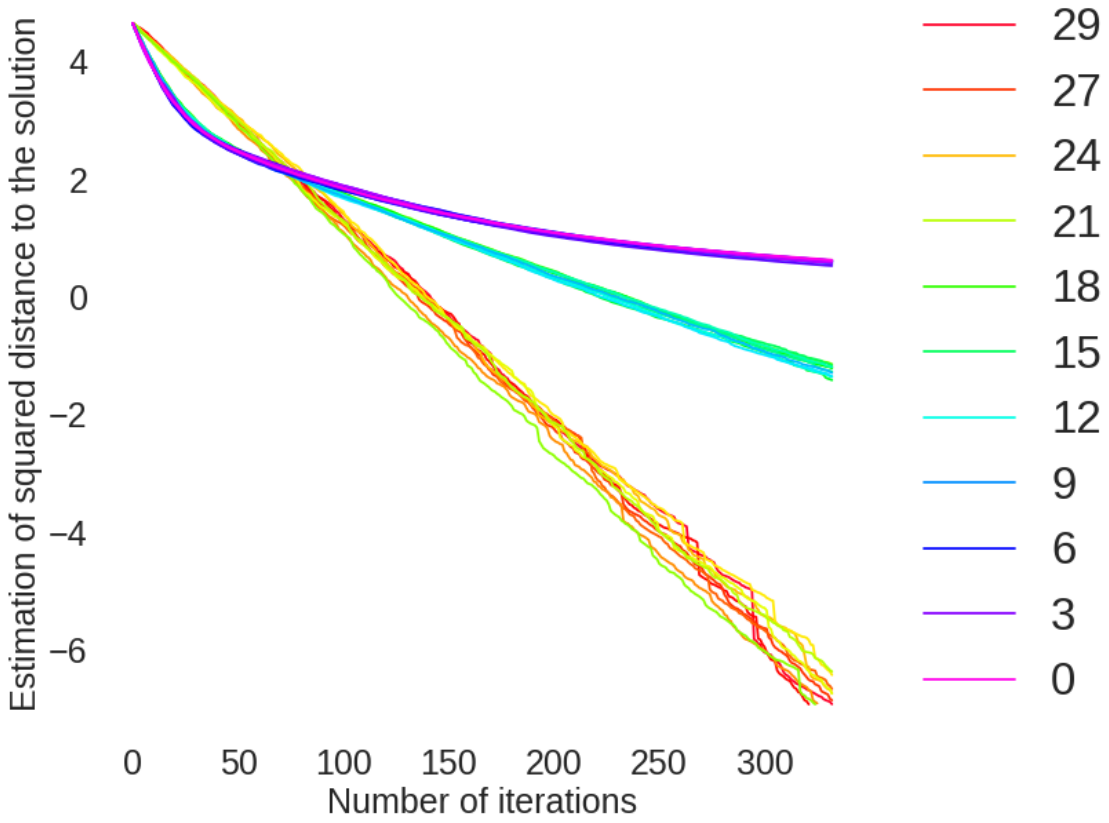

Figure: One third of eigenvalues were sampled from uniform distribution on $[10; 11]$, one third from uniform distribution on $[100; 101]$ and one third from uniform distribution on $[1,000; 1,001]$; $n = 30$

Matrix with 10 Billion Entries

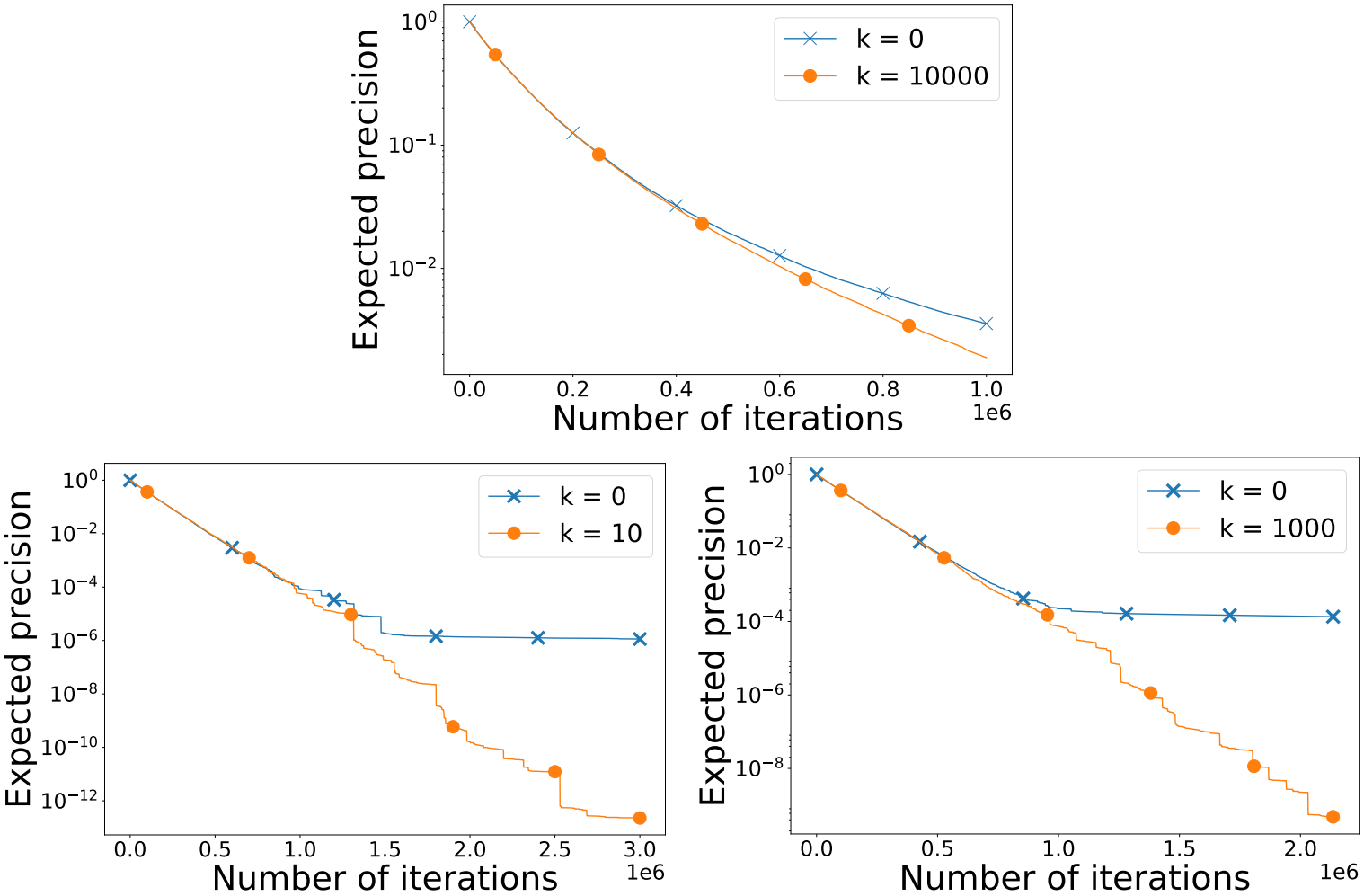

Figure: Top row: spectrum of $\mathbf{A}$ is uniformly distributed on $[1, 100]$; bottom row: spectrum contained in two clusters: $[1, 2]$ and $[100, 200]$; $n = 10^5$

Conclusions

- Negative results that highlight limitations of RCD with importance sampling
- Acceleration of RCD based on the augmentation of the set of coordinate directions by a few spectral directions
- Not mentioned: SSD/SconD with inexact spectral/conjugate directions.