Stochastic Spectral Descent Methods

Дмитрий Ковалев

14 июня 2018 г.

Задача

Рассмотрим задачу квадратичной оптимизации:

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

- $\mathbf{A}$ ($n \times n$ симметричная положительно определенная матрица)
- Единственное решение $x_* = \mathbf{A}^{-1} b$
- $f(x)$ сильно выпуклая квадратичная функция

Рандомизированный покомпонентный спуск (RCD)

Алгоритм 1 RCD

Параметры: вероятности $p_1, \ldots, p_n > 0$
Инициализация: выбрать $x_0 \in \mathbb{R}^n$
for $t = 0, 1, 2 \ldots$ **do**
Выбрать случайный номер $i \in \{1, \ldots, n\}$ с вероятностью p_i
$x_{t+1} \leftarrow x_t - \frac{\mathbf{A}_{:i}^\top x_t - b_i}{\mathbf{A}_{ii}} e_i,$
end for

Сходимость (Leventhal & Lewis 2010)

Пусть выбраны вероятности $p_i \sim \mathbf{A}_{ii}$. Тогда для достижения точности $\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq \epsilon$ алгоритму 1 требуется

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

итераций.

Стохастический спуск

Алгоритм 2 Стохастический спуск (Gower & Richtárik 2015)

Параметр: распределение $\mathcal{D}$ на векторах из $\mathbb{R}^n$

Инициализация: выбрать $x_0 \in \mathbb{R}^n$

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

Выбрать случайный вектор s_t из $\mathcal{D}$

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

end for

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

Для достижения точности $\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq \epsilon$ алгоритму 2 требуется

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

итераций, где $\mathbf{W} := \mathbb{E}_{s\sim\mathcal{D}}[\mathbf{A}^{1/2}\mathbf{H}\mathbf{A}^{1/2}]$, $\mathbf{H} := \frac{ss^\top}{s^\top \mathbf{A} s}$. (Предполагается, что $\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}]$ – обратимая матрица)

RCD с произвольными вероятностями

Сходимость RCD с произвольными вероятностями

Пусть выбраны вероятности $p_1, \ldots, p_n > 0$. Тогда для достижения точности $\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq \epsilon$ рандомизированному покомпонентному спуску требуется

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

итераций.

RCD с произвольными вероятностями

Равномерные вероятности оптимальны в 2D:

Теорема

Рассмотрим $n = 2$ и RCD с вероятностями $p_1, p_2 > 0$. Вероятности $p_1 = p_2 = \frac{1}{2}$ максимизируют скорость сходимости RCD.

«Типичный» выбор вероятностей ($p_i \sim \mathbf{A}_{ii}$, $p_i \sim \|\mathbf{A}_{i:}\|^2$) может оказаться «плохим»:

Теорема

Для любых $n \geq 2$ и $T > 0$ существует матрица $\mathbf{A}$, такая что: (i) Скорость сходимости RCD с вероятностями $p_i \sim \mathbf{A}_{ii}$ в T раз хуже, чем скорость сходимости RCD с равномерными вероятностями. (ii) Скорость сходимости RCD с вероятностями $p_i \sim \|\mathbf{A}_{i:}\|^2$ в T раз хуже, чем скорость сходимости RCD с равномерными вероятностями.

RCD с произвольными вероятностями

Полученная скорость сходимости RCD может быть сколь угодно медленной:

Теорема

Для любых $n \geq 2$ и $T > 0$ существует такая матрица $\mathbf{A}$, что число итераций (по формуле (1)) RCD с любым выбором вероятностей $p_1, \ldots, p_n > 0$ равно $\mathcal{O}(T \log(1/\epsilon))$.

Нижняя оценка на скорость сходимости RCD также может быть сколь угодно плохой:

Теорема

Для любых $n \geq 2$ и $T > 0$ существуют такие $n \times n$ положительно определенная матрица $\mathbf{A}$ и начальная точка x_0, что число итераций RCD с любыми вероятностями $p_1, \ldots, p_n > 0$ равно $\Omega(T \log(1/\epsilon))$.

Стохастический спектральный спуск (SSD)

- Алгоритм 2 (стохастический спуск) достигает оптимальной скорости сходимости

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

в случае, когда распределение $\mathcal{D}$ состоит из собственных векторов матрицы $\mathbf{A}$ с равными вероятностями.

- Аналогичный результат в случае, когда распределение $\mathcal{D}$ состоит из $\mathbf{A}$-ортогональных векторов с равными вероятностями.

Стохастический спектральный покомпонентный спуск (SSCD)

Собственное разложение матрицы A:

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

собственные значения: $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ собственные векторы: u

Предположим, что известны векторы $u_1, \ldots, u_k$ и значения $\lambda_1, \ldots, \lambda_{k+1}$.

Алгоритм 3 SSCD

Параметр: Выбрать $k \in \{0, \ldots, n-1\}$; $C_k = k\lambda_{k+1} + \sum_{i=k+1}^{n} \lambda_i$
Запустить Алгоритм 2 с распределением $\mathcal{D}$:

$$s_t = \begin{cases} e_i & \text{с вероятностью } p_i = \frac{\mathbf{A}_{ii}}{C_k}, \quad i = 1, 2, \ldots, n \\ u_i & \text{с вероятностью } p_{n+i} = \frac{\lambda_{k+1} - \lambda_i}{C_k}, \quad i = 1, 2, \ldots, k. \end{cases}$$

Сходимость SSCD

Теорема

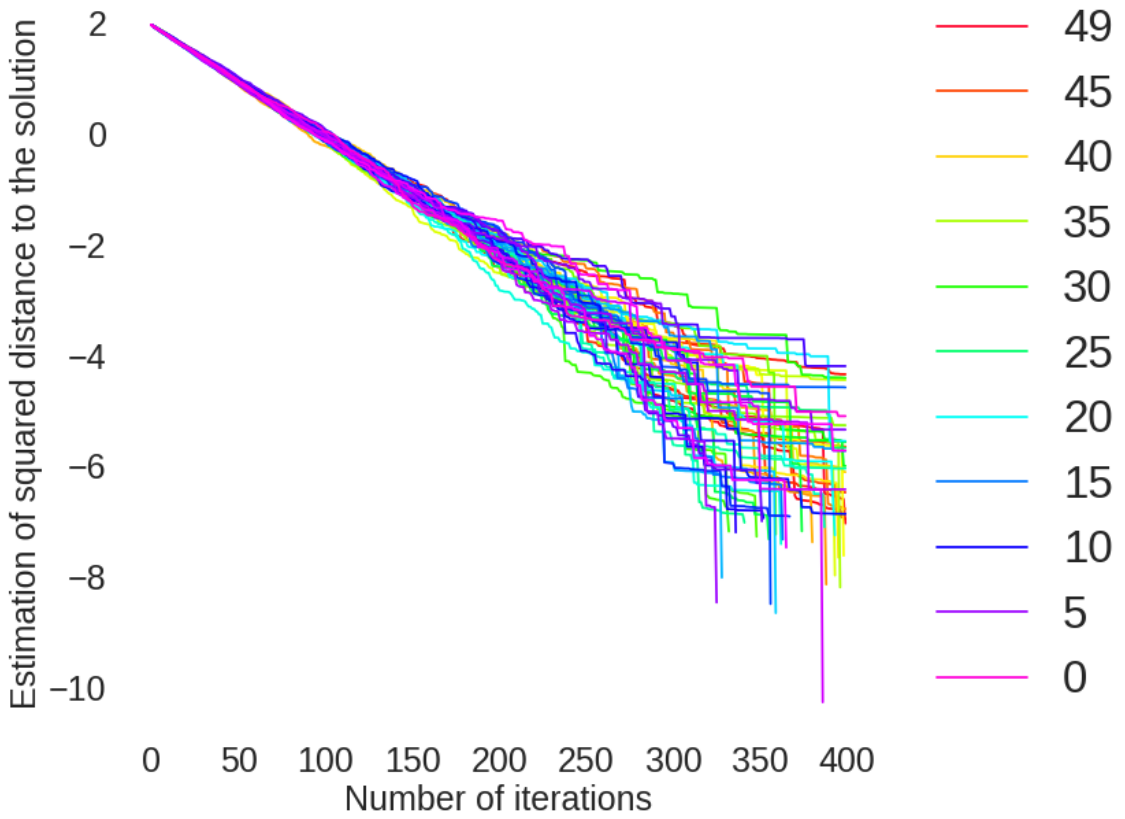
Для любого $n \geq 2$, алгоритм 3 (SSCD) сходится с линейной скоростью

$$\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq \left(1 - \frac{\lambda_{k+1}}{C_k}\right)^t \|x_0 - x_*\|_{\mathbf{A}}^2.$$

Более того, скорость сходимости улучшается с ростом числа k, и интерполируется между скоростью RCD $\lambda_1/\mathrm{Tr}(\mathbf{A})$ для $k = 0$, и оптимальной скоростью $1/n$ для $k = n - 1$:

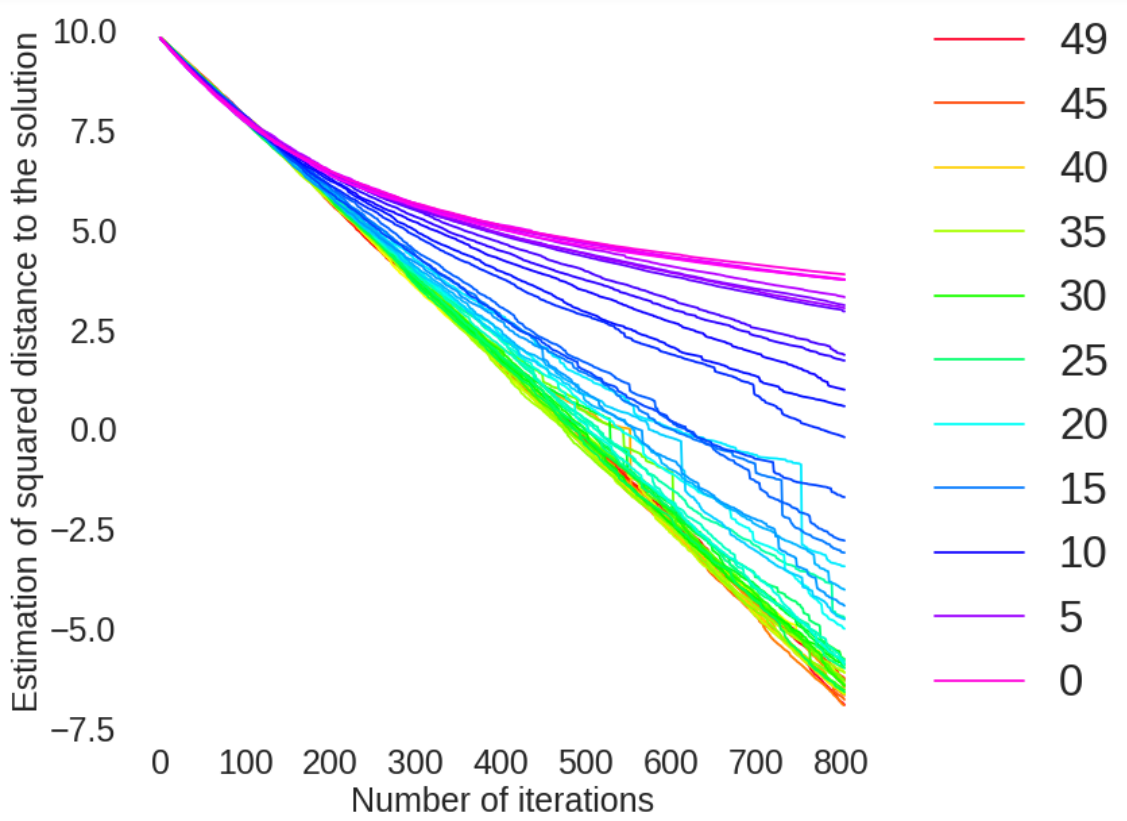
$$\frac{\lambda_1}{\mathrm{Tr}(\mathbf{A})} = \frac{\lambda_1}{C_0} \leq \cdots \leq \frac{\lambda_{k+1}}{C_k} \leq \cdots \leq \frac{\lambda_n}{C_{n-1}} = \frac{1}{n}.$$

Сходимость SSCD: Не зависит от k если собственные значения кластеризованы



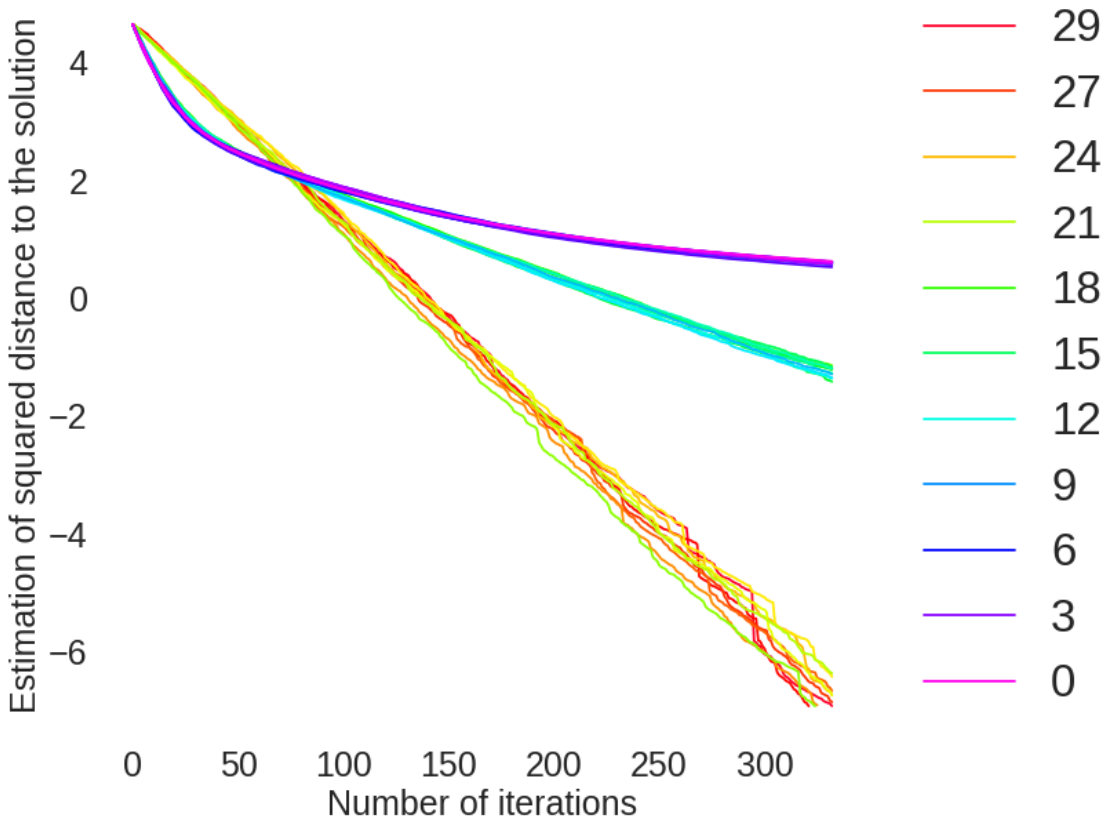
Собственные значения равномерно распределены на $[10; 11]$; $n = 50$

Скорость сходимости SSCD растет с увеличением k



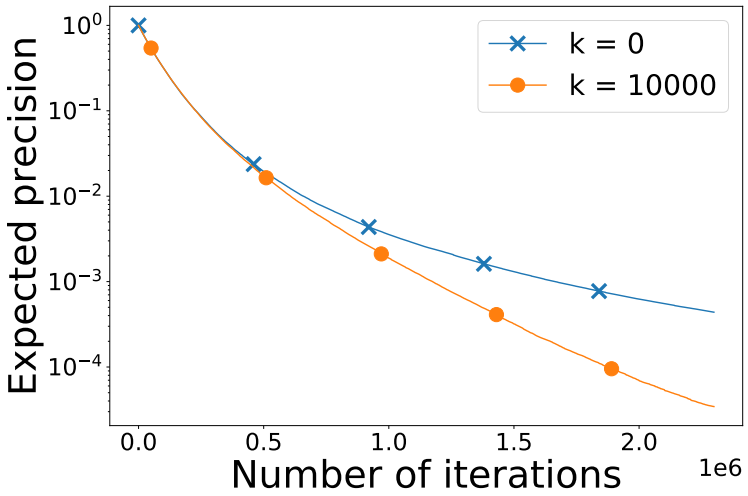
Собственные значения равномерно распределены на $[0; 10^5]$; $n = 50$

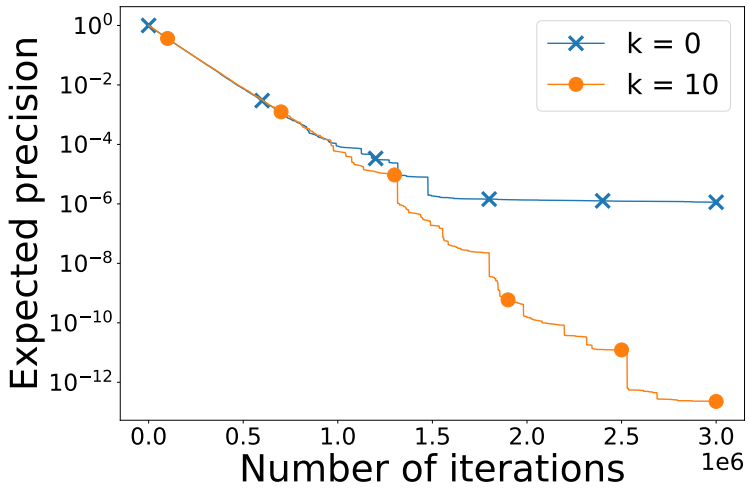
Сходимость SSCD: скачок скорости, когда k переходит между кластерами собственных значений

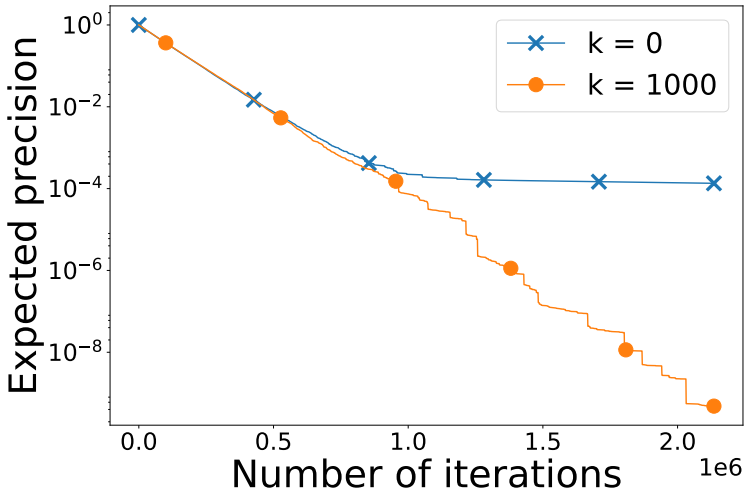


По одной трети собственных значений распределены равномерно на отрезках $[10; 11]$, $[100; 101]$ и $[1,000; 1,001]$ соответственно; $n = 30$

Сходимость SSCD: разреженная матрица, $n = 10^5$







Верхний ряд: спектр **A** равномерно распределен на $[1, 100]$.
Нижний ряд: спектр содержится в двух кластерах: $[1, 2]$ и $[100, 200]$.

Заключение

Некоторые результаты не вошедшие в презентацию

- **оптимальность распределения** в алгоритме 3 (SSCD)
- использование **приближенных** сопряженных и собственных направлений
- **распределенные** варианты методов

Участие в конференциях

- KAUST Research Workshop on Optimization and Big Data. Poster Session.