Министерство науки и высшего образования
Российской Федерации
«Московский физико-технический институт
(государственный университет)»
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Ковалев Дмитрий Александрович

# Стохастический спектральный спуск

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа
(бакалаврская диссертация)

**Научный руководитель:**
д.ф-м.н.
Гасников Александр Владимирович

Москва
2018

**Аннотация**

Современные методы решения оптимизационных задач большой размерности являются вариантами рандомизированного покомпонетного спуска. В данной работе представлен фундаментально новый тип ускорения рандомизированного покомпонетного спуска, основанный на пополнении набора координатных направлений несколькими *спектральными* или *сопряженными* направлениями. С ростом числа дополнительных направлений скорость сходимости метода улучшается и интерполируется между линейной скоростью сходимости рандомизированного покомпонетного спуска и линейной скоростью сходимости, *не зависящей от числа обусловленности.* Также рассматриваются неточные варианты данных методов для случая, когда известны приближения спектральных и сопряженных направлений. Вышеуказанные исследования мотивированы несколькими отрицательными результатами, которые подчеркивают ограничения рандомизированного покомпонентного спуска с выборкой по значимости.

**Abstract**

The state-of-the-art methods for solving optimization problems in big dimensions are variants of randomized coordinate descent (RCD). In this paper we introduce a fundamentally new type of acceleration strategy for RCD based on the augmentation of the set of coordinate directions by a few *spectral* or *conjugate* directions. As we increase the number of extra directions to be sampled from, the rate of the method improves, and interpolates between the linear rate of RCD and a linear rate *independent of the condition number*. We develop and analyze also inexact variants of these methods where the spectral and conjugate directions are allowed to be approximate only. We motivate the above development by proving several negative results which highlight the limitations of RCD with importance sampling.

# Table of contents

# 1 Introduction

An increasing array of learning and training tasks reduce to optimization problem in very large dimensions. The state-of-the-art algorithms in this regime are based on *randomized coordinate descent (RCD)*. Various acceleration strategies were proposed for RCD in the literature in recent years, based on techniques such as Nesterov's momentum [12, 9, 5, 1, 14], heavy ball momentum [16, 11], importance sampling [13, 19], adaptive sampling [4], random permutations [8], greedy rules [15], mini-batching [20], and locality breaking [21]. These techniques enable faster rates in theory and practice.

In this paper we introduce a fundamentally new type of acceleration strategy for RCD which relies on the idea of *enriching* the set of (unit) coordinate directions $\{e_1, e_2, \ldots, e_n\}$ in $\mathbb{R}^n$, which are used in RCD as directions of descent, via the addition of a few *spectral* or *conjugate* directions. The algorithms we develop and analyze in this paper randomize over this enriched larger set of directions.

## 1.1 The problem

For simplicity[1], we focus on quadratic minimization

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

where $\mathbf{A}$ is an $n\times n$ symmetric and positive definite matrix. The optimal solution is unique, and equal to $x_* = \mathbf{A}^{-1}b$.

## 1.2 Randomized coordinate descent

Applied to (1), RCD performs the iteration

$$x_{t+1} = x_t - \frac{\mathbf{A}_{:i}^\top x_t - b_i}{\mathbf{A}_{ii}} e_i, \tag{2}$$

where at each iteration, $i$ is chosen with probability $p_i > 0$. It was shown by Leventhal and Lewis [10] that if the probabilities are proportional to the diagonal elements of $\mathbf{A}$ (i.e., $p_i \sim \mathbf{A}_{ii}$), then the random iterates of RCD satisfy

$$\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq (1-\rho)^t \|x_0 - x_*\|_{\mathbf{A}}^2,$$

where $\rho = \frac{\lambda_{\min}(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})}$ and $\lambda_{\min}(\mathbf{A})$ is the minimal eigenvalue of $\mathbf{A}$. That is, as long as the number of iterations $t$ is at least

$$\mathcal{O}\left(\frac{\mathrm{Tr}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}\log\tfrac{1}{\epsilon}\right), \tag{3}$$

we have $\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq \epsilon$. Note that $\mathrm{Tr}(\mathbf{A})/\lambda_{\min}(\mathbf{A}) \geq n$, and that this can be arbitrarily larger that $n$.

[1] Many of our results can be extended to convex functions of the form $f(x) = \phi(\mathbf{A}x) - b^\top x$, where $\phi$ is a smooth and strongly convex function. However, due to space limitations, and the fact that we already have a lot to say in the special case $\phi(y) = \frac{1}{2}\|y\|^2$, we leave these more general developments to a follow-up paper.

| Method Name | Algorithm | Rate | Reference |
|---|---|---|---|
| stochastic descent (SD) | (4), Algorithm 1 | (5), Lemma 1 | Gower and Richtárik [6] |
| stochastic spectral descent (SSD) | Algorithm 2 | (6), Theorem 2 | NEW |
| stochastic conjugate descent (SconD) | read Section 2.2 | Theorem 2 | NEW |
| randomized coordinate descent (RCD) | (2), Algorithm 3 | (3), (13) | Gower and Richtárik [6] |
| stochastic spectral coordinate descent (SSCD) | Algorithm 4 | (7), Theorem 8 | NEW |
| mini-batch SD (mSD) | Algorithm 5 | Lemma 9 | Richtárik and Takáč [18] |
| mini-batch SSCD (mSSCD) | Algorithm 6 | Theorem 10 | NEW |
| inexact SconD (iSconD) | Algorithm 7 | Theorem 15 | NEW |
| inexact SSD (iSSD) | Algorithm 8 | see Section E.2 | NEW |

Table 1: Algorithms described in this paper.

## 1.3 Stochastic descent

Recently, Gower and Richtárik [6] developed an iterative "sketch and project" framework for solving linear systems and quadratic optimization problems; see also [7] for extensions. In the context of problem (1), and specialized to sketching matrices with a single column, their method takes the form

$$x_{t+1} = x_t - \frac{s_t^\top (\mathbf{A}x_t - b)}{s_t^\top \mathbf{A} s_t} s_t, \tag{4}$$

where $s_t \in \mathbb{R}^n$ is a random vector sampled from some fixed distribution $\mathcal{D}$. In this paper we will refer to this method by the name *stochastic descent (SD)*.

Note that $x_{t+1}$ is obtained from $x_t$ by minimizing $f(x_t + hs_t)$ for $h \in \mathbb{R}$ and setting $x_{t+1} = x_t + hs_t$. Further, note that RCD arises as a special case with $\mathcal{D}$ being a discrete probability distribution over the set $\{e_1, \dots, e_n\}$. However, SD converges for virtually any distribution $\mathcal{D}$, including discrete and continuous distributions. In particular, Gower and Richtárik [6] show that as long as $\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}]$ is invertible, where $\mathbf{H} := \frac{ss^\top}{s^\top \mathbf{A} s}$, then SD converges as

$$\mathcal{O}\left(\frac{1}{\lambda_{\min}(\mathbf{W})} \log \tfrac{1}{\epsilon}\right), \tag{5}$$

where $\mathbf{W} := \mathbb{E}_{s\sim\mathcal{D}}[\mathbf{A}^{1/2}\mathbf{H}\mathbf{A}^{1/2}]$ (see Lemma 1 for a more refined result due to Richtárik and Takáč [18]). Rate of RCD in (3) can be obtained as a special case of (5).

## 1.4 Stochastic spectral descent

The starting point of this paper is the new observation that stochastic descent obtains the rate

$$\mathcal{O}\left(n \log \frac{1}{\epsilon}\right) \tag{6}$$

in the special case when $\mathcal{D}$ is chosen to be the uniform distribution over the eigenvectors of $\mathbf{A}$ (see Theorem 2). For obvious reasons, we refer to this new method as *stochastic spectral descent (SSD)*.

To the best of our knowledge, SSD was not explicitly considered in the literature before. We should note that SSD is fundamentally different from *spectral gradient descent* [3, 2], which refers to a family of gradient descent methods with a special choice of stepsize depending on the spectrum of the Hessian of $f$.

| Result | Theorem |
|---|---|
| Uniform probabilities are optimal for $n = 2$ | 3 |
| Uniform probabilities are optimal for any $n \geq 2$ as long as $\mathbf{A}$ is diagonal | 4 |
| "Importance sampling" $p_i \sim \mathbf{A}_{ii}$ can lead to an arbitrarily worse rate than uniform probabilities | 5 |
| "Importance sampling" $p_i \sim \|\mathbf{A}_{i:}\|^2$ can lead to an arbitrarily worse rate than uniform probabilities | 5 |
| For every $n \geq 2$ and $T > 0$, there is $\mathbf{A}$ such that the rate of RCD with optimal probabilities is $\mathcal{O}(T \log \frac{1}{\epsilon})$ | 6 |
| For every $n \geq 2$ and $T > 0$, there is $\mathbf{A}$ such that the rate of RCD with optimal probabilities is $\Omega(T \log \frac{1}{\epsilon})$ | 7 |

Table 2: Summary of results on importance and optimal sampling in RCD.

The rate (6) does not merely provide an improvement on the rate of RCD given in (3); what is remarkable is that this rate is completely independent of the properties (such as conditioning) of $\mathbf{A}$. Moreover, we show that this method is *optimal* among the class of stochastic descent methods (4) parameterized by the choice of the distribution $\mathcal{D}$ (see Theorem 8). Despite the attractiveness of its rate, SSD is not a practical method. This is because once we have the eigenvectors of $\mathbf{A}$ available, the optimal solution $x_*$ can be assembled directly without the need for an iterative method.

## 1.5 Stochastic conjugate descent

We extend all results discussed above for SSD, including the rate (6), to the more general class of methods we call *stochastic conjugate descent (SconD)*, for which $\mathcal{D}$ is the uniform distribution over vectors $v_1, \ldots, v_n$ which are mutually $\mathbf{A}$ *conjugate*: $v_i^\top \mathbf{A} v_j = 0$ for $i \neq j$ and $v_i^\top \mathbf{A} v_i = 1$.

## 1.6 Optimizing probabilities in RCD

The idea of speeding up RCD via the use of non-uniform probabilities was pioneered by Nesterov [13] in the context of smooth convex minimization, and later built on by many authors [19, 17, 1]. In the case of non-accelerated RCD, and in the context of smooth convex optimization, the most popular choice of probabilities is to set $p_i \sim L_i$, where $L_i$ is the Lipschitz constant of the gradient of the objective corresponding to coordinate $i$ [13, 19]. For problem (1), we have $L_i = \mathbf{A}_{ii}$. Gower and Richtárik [6] showed that the optimal probabilities for (1) can in principle be computed through semidefinite programming (SDP); however, no theoretical properties of the optimal solution of the SDP were given.

As a warm-up, we first ask the following question: how important is importance sampling? More precisely, we investigate RCD with probabilities $p_i \sim \mathbf{A}_{ii}$, and RCD with probabilities $p_i \sim \|\mathbf{A}_{i:}\|^2$, considered as RCD with "importance sampling", and compare these with the baseline RCD with uniform probabilities. Our result (see Theorem 5) contradicts conventional "wisdom". In particular, we show that for every $n$ there is a matrix $\mathbf{A}$ such that diagonal probabilities lead to the best rate. Moreover, the rate of RCD with "importance" can be arbitrarily worse than the rate of RCD with uniform probabilities. The same result applies to probabilities proportional to the square of the norm of the $i$th row of $\mathbf{A}$.

We then switch gears, and motivated by the nature of SSD, we ask the following question: in order to obtain a condition-number-independent rate such as (6), do we *have to* consider new (and hard to compute) descent directions, such as eigenvectors of $\mathbf{A}$, or can a similar effect be obtained using RCD with a better selection of probabilities? We give two negative results to this question (see Theorems 6 and 7). First, we show that

| | general spectrum | $n-k$ largest eigvls are $\gamma$-clustered $c \leq \lambda_i \leq \gamma c$ for $k+1 \leq i \leq n$ | $\alpha$-exp decaying eigvls |
|---|---|---|---|
| RCD ($p_i \sim \mathbf{A}_{ii}$) | $\tilde{\mathcal{O}}\left(\frac{\sum_i \lambda_i}{\lambda_1}\right)$ | $\tilde{\mathcal{O}}\left(\frac{\gamma n c}{\lambda_1}\right)$ | $\tilde{\mathcal{O}}\left(\frac{1}{\alpha^{n-1}}\right)$ |
| SSCD | $\tilde{\mathcal{O}}\left(\frac{(k+1)\lambda_{k+1}+\sum_{i=k+2}^{n}\lambda_i}{\lambda_{k+1}}\right)$ | $\tilde{\mathcal{O}}(\gamma n)$ | $\tilde{\mathcal{O}}\left(\frac{1}{\alpha^{n-k-1}}\right)$ |
| SSD | $\tilde{\mathcal{O}}(n)$ | $\tilde{\mathcal{O}}(n)$ | $\tilde{\mathcal{O}}(n)$ |

Table 3: Comparison of complexities of RCD, SSCD (with parameter $0 \leq k \leq n-1$) and SSD under various regimes on the spectrum of $\mathbf{A}$. The $\tilde{\mathcal{O}}$ notation supresses a $\log \frac{1}{\epsilon}$ term.

for any $n \geq 2$ and any $T > 0$, there is a matrix $\mathbf{A}$ such that the rate of RCD with *any probabilities* (including the optimal probabilities) is $\mathcal{O}(T \log \frac{1}{\epsilon})$. Second, we give a similar but much stronger statement where we reach the same conclusion, but for the *lower bound* as opposed to the upper bound. That is, $\mathcal{O}$ is replaced by $\Omega$.

As a by-product of our investigations into importance sampling, we establish that for $n = 2$, *uniform probabilities* are optimal for all matrices $\mathbf{A}$ (see Theorem 3). For a summary of all these results, see Table 2.

## 1.7 Interpolating between RCD and SSD

RCD and SSD lie on opposite ends of a continuum of stochastic descent methods for solving (1). RCD "minimizes" the work per iteration without any regard for the number of iterations, while SSD minimizes the number of iterations without any regard for the cost per iteration (or pre-processing cost). Indeed, one step of RCD costs $\mathcal{O}(\|\mathbf{A}_{i:}\|_0)$ (the number of nonzero entries in the $i$th row of $\mathbf{A}$), and hence RCD can be implemented very efficiently for sparse $\mathbf{A}$. If uniform probabilities are used, no pre-processing (for computing probabilities) is needed. These advantages are paid for by the rate (3), which can be arbitrarily high. On the other hand, the rate of SSD does not depend on $\mathbf{A}$. This advantage is paid for by a high pre-processing cost: the computation of the eigenvectors. This pre-processing cost makes the method utterly impractical.

One of the main contributions of this paper is the development of a new *parametric family of algorithms that in some sense interpolate between RCD and SSD*.

In particular, we consider the stochastic descent algorithm (4) with $\mathcal{D}$ being a discrete distribution over the search directions $\{e_1, \ldots, e_n\} \cup \{u_1, \ldots, u_k\}$, where $u_i$ is the eigenvectors of $\mathbf{A}$ corresponding to the $i$th smallest eigenvalue of $\mathbf{A}$. We refer to this new method by the name *stochastic spectral coordinate descent (SSCD)*.

We compute the optimal probabilities of this distribution, which turn out to be unique, and show that for $k \geq 1$ they depend on the $k+1$ smallest eigenvalues of $\mathbf{A}$: $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{k+1}$. In particular, we prove (see Theorem 8) that the rate of SSCD with optimal probabilities is

$$\mathcal{O}\left(\frac{(k+1)\lambda_{k+1} + \sum_{i=k+2}^{n} \lambda_i}{\lambda_{k+1}} \log \tfrac{1}{\epsilon}\right). \tag{7}$$

For $k = 0$, SSCD reduces to RCD with $p_i \sim \mathbf{A}_{ii}$, and the rate (7) reduces to (3). For $k = n-1$, SSCD *does not* reduce to SSD. However, the rates match. Indeed, in this case the rate (7) reduces to (6). Moreover, the rate improves monotonically as $k$ increases, from $\mathcal{O}(\frac{\mathrm{Tr}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})} \log \frac{1}{\epsilon})$ (for $k = 0$) to $\mathcal{O}(n \log \frac{1}{\epsilon})$ (for $k = n-1$).

**SSCD removes the effect of the $k$ smallest eigenvalues.** Note that the rate (7) does *not depend* on the $k$ smallest eigenvalues of $\mathbf{A}$. That is, by adding the eigenvectors $u_1, \dots, u_k$ corresponding to the $k$ smallest eigenvalues to the set of descent directions, we have removed the effect of these eigenvalues.

**Clustered eigenvalues.** Assume that the $n-k$ largest eigenvalues are clustered: $c \leq \lambda_i \leq \gamma c$ for some $c > 0$ and $\gamma > 1$, for all $k+1 \leq i \leq n$. In this case, the rate (7) can be estimated as a function of the clustering "tightness" parameter $\gamma$: $\mathcal{O}\left(\gamma n \log \frac{1}{\epsilon}\right)$. See Table 3.

This can be arbitrarily better than the rate of RCD, even for $k = 1$. In other words, there are situations where by enriching the set of directions used by RCD by a single eigenvector only, the resulting method accelerates dramatically. To give a concrete and simplified example to illustrate this, assume that $\lambda_1 = \delta > 0$, while $\lambda_2 = \cdots = \lambda_n = 1$. In this case, RCD has the rate $\mathcal{O}((1 + \frac{n-1}{\delta}) \log \frac{1}{\epsilon})$, while SSCD with $k = 1$ has the rate $\mathcal{O}(n \log \frac{1}{\epsilon})$. So, SSCD is $\frac{1}{\delta}$ times better than RCD, and the difference grows to infinity as $\delta$ approaches zero even for fixed dimension $n$.

**Exponentially decaying eigenvalues.** If the eigenvalues of $\mathbf{A}$ follow an exponential decay with factor $0 < \alpha < 1$, then the rate of RCD is $\mathcal{O}(\frac{1}{\alpha^{n-1}} \log \frac{1}{\epsilon})$, while the rate of SSCD is $\mathcal{O}(\frac{1}{\alpha^{n-k-1}} \log \frac{1}{\epsilon})$. This is an improvement by the factor $\frac{1}{\alpha^k}$, which can be very large even for small $k$ if $\alpha$ is small. See Table 3. For an experimental confirmation of this prediction, see Figure 5.

**Adding a few "largest" eigenvectors does not help.** We show that in contrast with the situation above, adding a few of the "largest" eigenvectors to the coordinate directions of RCD does not help. This is captured formally in the appendix as Theorem 12.

**Mini-batching.** We extend SSCD to a mini-batch setting; we call the new method *mSSCD*. We show that the rate of mSSCD interpolates between the rate of mini-batch RCD and rate of SSD. Moreover, we show that mSSCD is optimal among a certain parametric family of methods, and that its rate improves as $k$ increases. See Theorem 10.

### 1.8 Inexact Directions

Finally, we relax the need to compute exact eigenvectors or $\mathbf{A}$- conjugate vectors, and analyze the behavior of our methods for inexact directions. Moreover, we propose and analyze an inexact variant of SSD which does *not* arise as a special case of SD. See Sections D and E.

## 2 Stochastic Descent

The stochastic descent method was described in (4). We now formalize it as Algorithm 1, and equip it with a stepsize, which will be useful in Section 3.2, where we study mini-batch version of SD.

In order to guarantee convergence of SD, we restrict our attention to the class of *proper* distributions, defined next.

**Algorithm 1** Stochastic Descent (SD)

**Parameters:** Distribution $\mathcal{D}$; Stepsize parameter $\omega > 0$
**Initialize:** Choose $x_0 \in \mathbb{R}^n$
**for** $t = 0, 1, 2, \ldots$ **do**
  Sample search direction $s_t \sim \mathcal{D}$
  Set $x_{t+1} = x_t - \omega \frac{s_t^\top (\mathbf{A}x_t - b)}{s_t^\top \mathbf{A} s_t} s_t$
**end for**

**Assumption 1.** *Distribution $\mathcal{D}$ is* proper *with respect to $\mathbf{A}$. That is, $\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}]$ is invertible, where*

$$\mathbf{H} := \frac{ss^\top}{s^\top \mathbf{A} s}. \tag{8}$$

Next we present the main convergence result for SD.

**Lemma 1** (Convergence of stochastic descent [6, 18])**.** *Let $\mathcal{D}$ be proper with respect to $\mathbf{A}$, and let $0 < \omega < 2$. Stochastic descent (Algorithm 1) converges linearly in expectation. In particular, we have*

$$(1 - \omega(2-\omega)\lambda_{\max}(\mathbf{W}))^t \|x_0 - x_*\|_{\mathbf{A}}^2 \leq \mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \tag{9}$$

*and*

$$\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq (1 - \omega(2-\omega)\lambda_{\min}(\mathbf{W}))^t \|x_0 - x_*\|_{\mathbf{A}}^2, \tag{10}$$

*where*

$$\mathbf{W} := \mathbb{E}_{s\sim\mathcal{D}}[\mathbf{A}^{1/2}\mathbf{H}\mathbf{A}^{1/2}]. \tag{11}$$

*Finally, the statement remains true if we replace $\|x_t - x_*\|_{\mathbf{A}}^2$ by $f(x_t) - f(x_*)$ for all $t$.*

It is easy to observe that the stepsize choice $\omega = 1$ is optimal. This is why we have decided to present the SD method (4) with this choice of stepsize. Moreover, notice that due to linearity of expectation,

$$\begin{aligned}
\mathrm{Tr}(\mathbf{W}) &\overset{(11)}{=} \mathbb{E}[\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{H}\mathbf{A}^{1/2})] \\
&\overset{(8)}{=} \mathbb{E}\left[\mathrm{Tr}\left(\frac{zz^\top}{z^\top z}\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\frac{z^\top z}{z^\top z}\right)\right] \\
&= 1,
\end{aligned}$$

where $z = \mathbf{A}^{1/2}s$. Therefore,

$$0 < \lambda_{\min}(\mathbf{W}) \leq \frac{1}{n} \leq \lambda_{\max}(\mathbf{W}) \leq 1.$$

## 2.1 Stochastic Spectral Descent

Let $\mathbf{A} = \sum_{i=1}^n \lambda_i u_i u_i^\top$ be the eigenvalue decomposition of $\mathbf{A}$. That is, $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ are the eigenvalues of $\mathbf{A}$ and $u_1, \ldots, u_n$ are the corresponding orthonormal eigenvectors. Consider now the SD method with $\mathcal{D}$ being the uniform distribution over the

**Algorithm 2** Stochastic Spectral Descent (SSD)

**Initialize:** $x_0 \in \mathbb{R}^n$; $(u_1, \lambda_1), \dots (u_n, \lambda_n)$: eigenvectors and eigenvalues of $\mathbf{A}$
**for** $t = 0, 1, 2, \dots$ **do**
  Choose $i \in [n]$ uniformly at random
  Set $x_{t+1} = x_t - \left( u_i^\top x_t - \frac{u_i^\top b}{\lambda_i} \right) u_i$
**end for**

set $\{u_1, \dots, u_n\}$, and $\omega = 1$. This gives rise to a new variant of SD which we call *stochastic spectral descent (SSD)*.

For SSD we can establish an unusually strong convergence result, both in terms of speed and tightness.

**Theorem 2** (Convergence of stochastic spectral descent)**.** *Let $\{x_k\}$ be the sequence of random iterates produced by stochastic spectral descent (Algorithm 2). Then*

$$\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] = \left(1 - \frac{1}{n}\right)^t \|x_0 - x_*\|_{\mathbf{A}}^2. \tag{12}$$

The above theorem implies the rate (6) mentioned in the introduction. It means that up to a logarithmic factor, SSD only needs $n$ iterations to converge. Notice that (12) is an *identity*, and hence the rate is not improvable.

## 2.2 Stochastic Conjugate Descent

The same rate as in Theorem 2 holds for the *stochastic conjugate descent* (SconD) method, which arises as a special case of stochastic descent for $\omega = 1$ and $\mathcal{D}$ being a uniform distribution over a set of $\mathbf{A}$-orthogonal (i.e., conjugate) vectors. The proof follows by combining Lemmas 1 and 13.

## 2.3 Randomized Coordinate Descent

RCD (Algorithm 3) arises as a special case of SD with unit stepsize ($\omega = 1$) and distribution $\mathcal{D}$ given by $s_t = e_i$ with probability $p_i > 0$.

**Algorithm 3** Randomized Coordinate Descent (RCD)

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

The rate of RCD (Algorithm 3) can therefore be deduced from Lemma 1. Notice that in view of (8), we have

$$\mathbb{E}[\mathbf{H}] = \sum_{i=1}^n p_i \frac{e_i e_i^\top}{\mathbf{A}_{ii}} = \mathrm{Diag}\left( \frac{p_1}{\mathbf{A}_{11}}, \dots, \frac{p_n}{\mathbf{A}_{nn}} \right).$$

So, as long as all probabilities are positive, Assumption 1 is satisfied. Therefore, Lemma 1 applies and RCD enjoys the rate

$$\mathcal{O}\left(\frac{1}{\lambda_{\min}\left(\mathbf{A}\mathrm{Diag}\left(\frac{p_i}{\mathbf{A}_{ii}}\right)\right)}\log\frac{1}{\epsilon}\right). \tag{13}$$

**Uniform probabilities can be optimal.** We first prove that uniform probabilities are optimal in 2D.

**Theorem 3.** *Let $n = 2$ and consider RCD (Algorithm 3) 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* (13)*.*

Next we claim that uniform probabilities are optimal in any dimension $n$ as long as the matrix $\mathbf{A}$ is diagonal.

**Theorem 4.** *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* (13)*.*

**"Importance" sampling can be unimportant.** In our next result we contradict conventional wisdom about typical choices of "importance sampling" probabilities. In particular, we claim that diagonal and row-squared-norm probabilities can lead to an arbitrarily worse performance than uniform probabilities.

**Theorem 5.** *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 $p_i \sim \|\mathbf{A}_{i:}\|^2$ is $T$ times worse than the rate of RCD with uniform probabilities.*

**Optimal probabilities can be bad.** Finally, we show that there is no hope for adjustment of probabilities in RCD to lead to a rate independent of the data $\mathbf{A}$, as is the case for SSD. Our first result states that such a result can't be obtained from the generic rate (13).

**Theorem 6.** *For every $n \geq 2$ and $T > 0$, there exists $\mathbf{A}$ such that the number of iterations (as expressed by formula* (13)*) of RCD with any choice of probabilities $p_1, \ldots, p_n > 0$ is $\mathcal{O}(T\log(1/\epsilon))$.*

However, that does not mean, by itself, that such a result can't be possibly obtained via a different analysis. Our next result shatters these hopes as we establish a *lower bound* which can be arbitrarily larger than the dimension $n$.

**Theorem 7.** *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 $\Omega(T\log(1/\epsilon))$.*

# 3 Interpolating Between RCD and SSD

Assume now that we have some partial spectral information available. In particular, fix $k \in \{0, 1, \ldots, n-1\}$ and assume we know eigenvectors $u_i$ and eigenvalues $\lambda_i$ for $i = 1, \ldots, k$. We now define a parametric distribution $\mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$ with parameters $\alpha > 0$ and $\beta_1, \ldots, \beta_k \geq 0$ as follows. Sample $s \sim \mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$ arises through the process

$$s = \begin{cases} e_i & \text{with probability } p_i = \frac{\alpha \mathbf{A}_{ii}}{C_k},\ i \in [n], \\ u_i & \text{with probability } p_{n+i} = \frac{\beta_i}{C_k},\ i \in [k], \end{cases} \tag{14}$$

where $C_k := \alpha \mathrm{Tr}(\mathbf{A}) + \sum_{i=1}^k \beta_i$ is a normalizing factor ensuring that the probabilities sum up to 1.

## 3.1 SSCD

Applying the SD method with the distribution $\mathcal{D} = \mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$ gives rise to a new specific method which we call *stochastic spectral coordinate descent (SSCD)*.

**Algorithm 4** Stochastic Spectral Coordinate Descent (SSCD)

**Parameters:** Distribution $\mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$
**Initialize:** $x_0 \in \mathbb{R}^n$
**for** $t = 0, 1, 2, \ldots$ **do**
  Sample $s_t \sim \mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$
  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 8.** *Consider Stochastic Spectral Coordinate Descent (Algorithm 4) for fixed $k \in \{0, 1, \ldots, n-1\}$. The method converges linearly for all positive $\alpha > 0$ and nonnegative $\beta_i$. The best rate is obtained for parameters $\alpha = 1$ and $\beta_i = \lambda_{k+1} - \lambda_i$; and this is the unique choice of parameters leading to the best rate. In this case,*

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

*where*

$$C_k = (k+1)\lambda_{k+1} + \sum_{i=k+2}^{n} \lambda_i.$$

*Moreover, the rate improves as $k$ grows, and we have*

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

If $k = 0$, SSCD reduces to RCD (with diagonal probabilities). Since $\frac{\lambda_1}{C_0} = \frac{\lambda_1}{\mathrm{Tr}(\mathbf{A})}$, we recover the rate of RCD of Leventhal and Lewis [10]. With the choice $k = n-1$ our method does *not* reduce to SSD. However, the rates match. Indeed, $\frac{\lambda_n}{C_{n-1}} = \frac{\lambda_n}{n\lambda_n} = \frac{1}{n}$ (compare with Theorem 2).

**"Largest" eigenvectors do not help.** It is natural to ask whether there is any benefit in considering a few "largest" eigenvectors instead. Unfortunately, for the same parametric family as in Theorem 8, the answer is negative. The optimal parameters suggest that RCD has better rate without these directions. See Theorem 12 in the appendix.

## 3.2 Mini-batch SD

A mini-batch version of SD was developed by Richtárik and Takáč [18]. Here we restate the method as Algorithm 5.

**Algorithm 5** Mini-batch Stochastic Descent (mSD)

**Parameters:** Distribution $\mathcal{D}$; stepsize parameter $\omega > 0$; mini-batch size $\tau \geq 1$
**Initialize:** $x_0 \in \mathbb{R}^n$
**for** $t = 0, 1, 2, \ldots$ **do**
  **for** $i = 1, 2, \ldots, \tau$ **do**
    Sample $s_{ti} \sim \mathcal{D}$
    Set $x_{t+1,i} = x_t - \omega \frac{s_{ti}^\top (\mathbf{A}x_t - b)}{s_{ti}^\top \mathbf{A} s_{ti}} s_{ti}$
  **end for**
  Set $x_{t+1} = \frac{1}{\tau} \sum_{i=1}^{\tau} x_{t+1,i}$
**end for**

**Lemma 9** (Convergence of mSD [18])**.** *Let $\mathcal{D}$ be proper with respect to $\mathbf{A}$, and let $0 < \omega < \frac{2}{\xi(\tau)}$, where $\xi(\tau) := \frac{1}{\tau} + \left(1 - \frac{1}{\tau}\right) \lambda_{\max}(\mathbf{W})$. Then*

$$\mathbb{E}[\|x_t - x_*\|_{\mathbf{A}}^2] \leq (\rho(\omega, \tau))^t \|x_0 - x_*\|_{\mathbf{A}}^2, \tag{15}$$

*where*

$$\rho(\omega, \tau) = 1 - \omega[2 - \omega\xi(\tau)]\lambda_{\min}(\mathbf{W}).$$

*For any fixed $\tau \geq 1$, the optimal stepsize choice is $\omega(\tau) = \frac{1}{\xi(\tau)}$ and the associated optimal rate is*

$$\rho(\omega(\tau), \tau) = 1 - \frac{\lambda_{\min}(\mathbf{W})}{\frac{1}{\tau} + \left(1 - \frac{1}{\tau}\right) \lambda_{\max}(\mathbf{W})}.$$

## 3.3 Mini-batch SSCD

Specializing mSD to the distribution $\mathcal{D} = \mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$ gives rise to a new specific method which we call *mini-batch stochastic spectral coordinate descent (mSSCD)*, and formalize as Algorithm 6.

The rate of mSSCD is governed by the following result.

**Theorem 10.** *Consider mSSCD (Algorithm 6) for fixed $k \in \{0, 1, \ldots, n-1\}$ and optimal stepsize parameter $\omega(\tau) = \frac{1}{\xi(\tau)}$. The method converges linearly for all positive $\alpha > 0$ and nonnegative $\beta_i$. The best rate is obtained for parameters $\alpha = 1$ and $\beta_i = \lambda_{k+1} - \lambda_i$; and this is the unique choice of parameters leading to the best rate. In this case,*

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

**Algorithm 6** Mini-batch Stochastic Spectral Coordinate Descent (mSSCD)

**Parameters:** Distribution $\mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$; relaxation parameter $\omega \in \mathbb{R}$; mini-batch size $\tau \geq 1$

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

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

  **for** $i = 1, 2, \ldots, \tau$ **do**

    Sample $s_{ti} \sim \mathcal{D}(\alpha, \beta_1, \ldots, \beta_k)$

    Set $x_{t+1,i} = x_t - \omega \frac{s_{ti}^\top (\mathbf{A}x_t - b)}{s_{ti}^\top \mathbf{A} s_{ti}} s_{ti}$

  **end for**

  Set $x_{t+1} = \frac{1}{\tau} \sum_{i=1}^{\tau} x_{t+1,i}$

**end for**

*where*

$$F_k := \frac{1}{\tau}\left((k+1)\lambda_{k+1} + \sum_{i=k+2}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\lambda_n.$$

*Moreover, the rate improves as $k$ grows, and we have*

$$\frac{\lambda_1}{\frac{1}{\tau}\mathrm{Tr}(\mathbf{A}) + \left(1 - \frac{1}{\tau}\right)\lambda_n} = \frac{\lambda_1}{F_0} \leq \cdots \leq \frac{\lambda_{k+1}}{F_k}$$

*and*

$$\frac{\lambda_{k+1}}{F_k} \leq \cdots \leq \frac{\lambda_n}{F_{n-1}} = \frac{1}{\frac{n-1}{\tau} + 1}.$$

If $k = 0$, mSSCD reduces to mini-batch RCD (with diagonal probabilities). Since $\frac{\lambda_1}{F_0} = \frac{\lambda_1}{\frac{1}{\tau}\mathrm{Tr}(\mathbf{A}) + \left(1-\frac{1}{\tau}\right)\lambda_n}$, we recover the rate of mini-batch RCD [18]. With the choice $k = n-1$ our method does *not* reduce to mSSD. However, the rates match.

# 4 Experiments

## 4.1 Stochastic spectral coordinate descent (SSCD)

In our first experiment we study how the practical behavior of SSCD (Algorithm 4) depends on the choice of $k$. What we study here does not depend on the dimensionality of the problem ($n$), and hence it suffices to perform the experiments on small dimensional problems ($n = 30$).

In this experiment we consider the regime of *clustered eigenvalues* described in Section 1.7 and summarized in Table 3. In particular, we construct a synthetic matrix $\mathbf{A} \in \mathbb{R}^{30\times 30}$ with the smallest 15 eigenvalues clustered in the interval $(5, 5+\Delta)$ and the largest 15 eigenvalues clustered in the interval $(\theta, \theta+\Delta)$. We vary the *tightness* parameter $\Delta$ and the *separation* parameter $\theta$, and study the performance of SSCD for various choices of $k$. See Figure 3.

Our first finding is a confirmation of the *phase transition* phenomenon predicted by our theory. Recall that the rate of SSCD (see Theorem 8) is

$$\tilde{\mathcal{O}}\left(\frac{(k+1)\lambda_{k+1} + \sum_{i=k+2}^{n} \lambda_i}{\lambda_{k+1}}\right).$$

If $k < 15$, we know $\lambda_i \in (5, 5+\Delta)$ for $i = 1, 2, \ldots, k+1$, and $\lambda_i \in (\theta, \theta+\Delta)$ for $i = k+2, \ldots, n$. Therefore, the rate can be estimated as

$$r_{small} := \tilde{\mathcal{O}}\left(k + 1 + \frac{(n-k-1)(\theta+\Delta)}{5}\right).$$

On the other hand, if $k \geq 15$, we know that $\lambda_i \in (\theta, \theta+\Delta)$ for $i = k+1, \ldots, n$, and hence the rate can be estimated as

$$r_{large} := \tilde{\mathcal{O}}\left(k + 1 + \frac{(n-k-1)(\theta+\Delta)}{\theta}\right).$$

Note that if the separation $\theta$ between the two clusters is large, the rate $r_{large}$ is much better than the rate $r_{small}$. Indeed, in this regime, the rate $r_{large}$ becomes $\tilde{\mathcal{O}}(n)$, while $r_{small}$ can be arbitrarily large.

Going back to Figure 3, notice that this can be observed in the experiments. There is a clear *phase transition* at $k = 15$, as predicted be the above analysis. Methods using $k \in \{0, 6, 12\}$ are relatively slow (although still enjoying a linear rate), and tend to have similar behaviour, especially when $\Delta$ is small. On the other hand, methods using $k \in \{18, 24, 29\}$ are much faster, with a behaviour nearly independent of $\theta$ and $\Delta$. Moreover, as $\theta$ increases, the difference in the rates between the *slow* methods using $k \in \{0, 6, 12\}$ and the *fast* methods using $k \in \{18, 24, 29\}$ grows.

We have performed additional experiments with three clusters; see Figure 4 in the appendix.

## 4.2 Mini-batch SSCD

In Figure 2 we report on the behavior of mSSCD, the mini-batch version of SSCD, for four choices of the mini-batch parameter $\tau$, and several choices of $k$. Mini-batch of size $\tau$ is processed in parallel on $\tau$ processors, and the cost of a single iteration of mSSCD is (roughly) the same for all $\tau$.

For $\tau = 1$, the method reduces to SSCD, considered in previous experiment (but on a different dataset). Since the number of iterations is small, there are no noticeable differences across using different values of $k$. As $\tau$ grows, however, all methods become faster. Mini-batching seems to be more useful as $k$ is larger. Moreover, we can observe that acceleration through mini-batching starts more aggressively for small values op $k$, and its added benefit for increasing values of $k$ is getting smaller and smaller. This means that even for relatively small values of $k$, mini-batching can be expected to lead to substantial speed-ups.

## 4.3 Matrix with 10 billion entries

In Figure 3 we report on an experiment using a synthetic problem with data matrix $\mathbf{A}$ of dimension $n = 10^5$ (i.e., potentially with $10^{10}$ entries). As all experiments were done on a laptop, we worked with sparse matrices with $10^6$ nonzeros only.

In the first row of Figure 3 we consider matrix $\mathbf{A}$ with all eigenvalues distributed uniformly on the interval $[1, 100]$. We observe that SSCD with $k = 10^4$ (just 10% of $n$) requires about an *order of magnitude* less iterations than SSCD with $k = 0$ (=RCD).

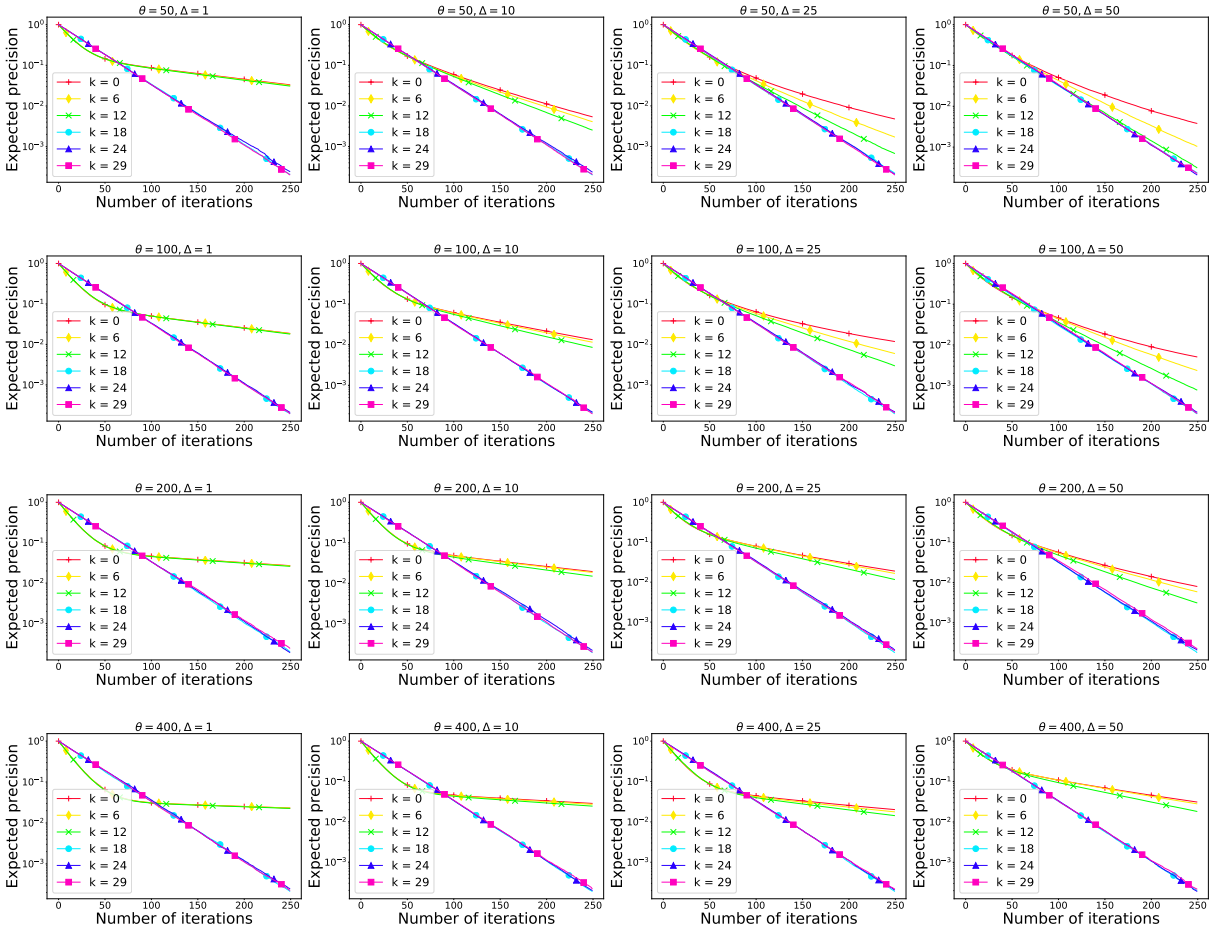


Figure 1: Expected precision $\mathbb{E}\left[\|x_t - x_*\|_{\mathbf{A}}^2 / \|x_0 - x_*\|_{\mathbf{A}}^2\right]$ versus # iterations of SSCD for symmetric positive definite matrices $\mathbf{A}$ of size $30 \times 30$ with different structures of spectra. The spectrum of $\mathbf{A}$ consists of 2 equally sized clusters of eigenvalues; one in the interval $(5, 5+\Delta)$, and the other in the interval $(\theta, \theta+\Delta)$.

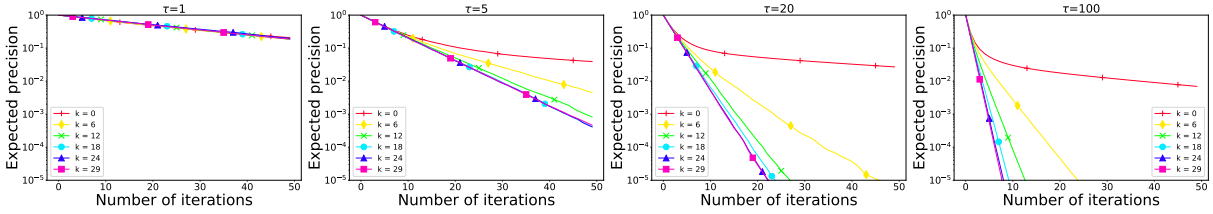


Figure 2: Expected precision $\mathbb{E}\left[\|x_t - x_*\|_{\mathbf{A}}^2 / \|x_0 - x_*\|_{\mathbf{A}}^2\right]$ versus # iterations of mini-batch SSCD for $\mathbf{A} \in \mathbb{R}^{30\times 30}$ and several choices of mini-batch size $\tau$. The spectrum of $\mathbf{A}$ was chosen as a uniform discretization of the interval $[1, 60]$.

In the second row we consider a scenario where $l$ eigenvalues are small, contained in $[1, 2]$, with the rest of the eigenvalues contained in $[100, 200]$. We consider $l = 10$ and $l = 1000$ and study the behaviour of SSCD with $k = l$. We see that for $l = 10$, SSCD performs dramatically better than RCD: it is able to achieve machine precision while RCD struggles to reduce the initial error by a factor larger than $10^6$. For $l = 1000$, SSCD

achieves error $10^{-9}$ while RCD struggles to push the error below $10^{-4}$. These tests show that in terms of # iterations, *SSCD has the capacity to accelerate on RCD by many orders of magnitude.*

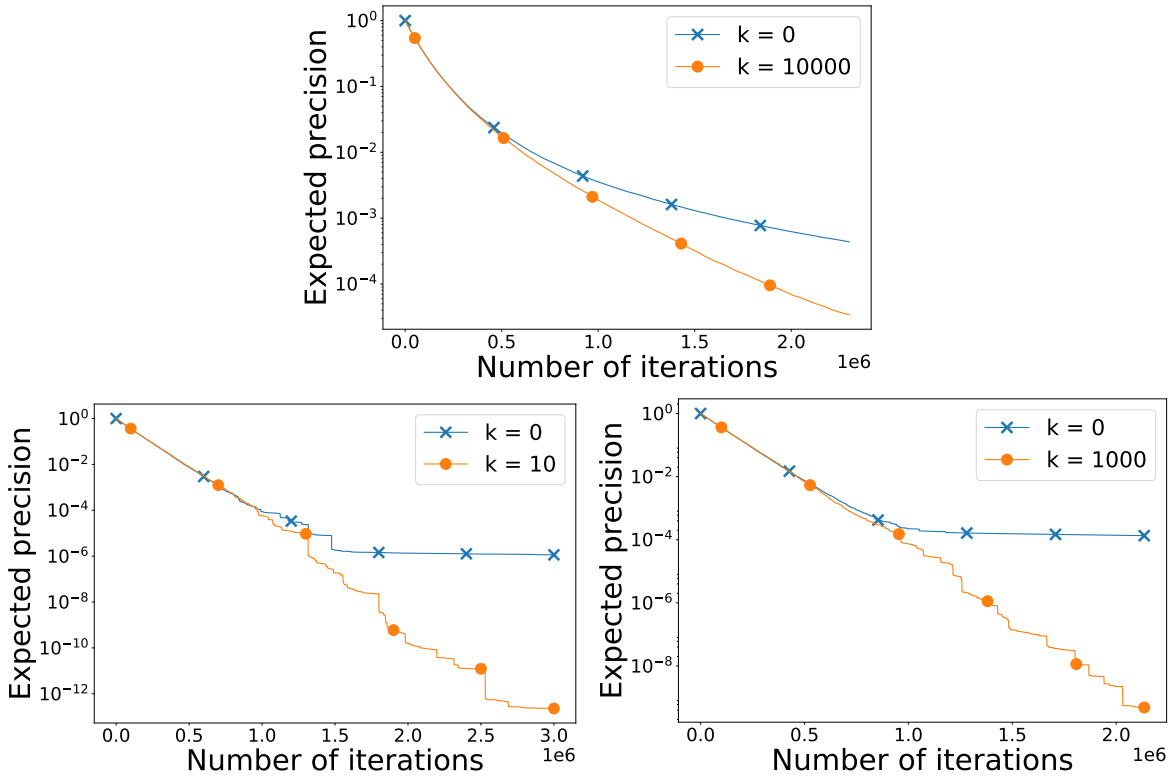


Figure 3: Expected precision $\mathbb{E}\left[\|x_t - x_*\|_{\mathbf{A}}^2 / \|x_0 - x_*\|_{\mathbf{A}}^2\right]$ versus # iterations of SSCD for a matrix $\mathbf{A} \in \mathbb{R}^{10^5 \times 10^5}$. Top row: spectrum of $\mathbf{A}$ is uniformly distributed on $[1, 100]$. Bottom row: spectrum contained in two clusters: $[1, 2]$ and $[100, 200]$.

# 5 Extensions

Our algorithms and convergence results can be extended to eigenvectors and conjugate directions which are only computed *approximately*. Some of this development can be found in the appendix (see Section D). Finally, as mentioned in the introduction, our results can be extended to the more general problem of minimizing $f(x) = \phi(\mathbf{A}x)$, where $\phi$ is smooth and strongly convex.

# References


[1] Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *ICML*, pages 1110–1119, 2016.

[2] Jonathan Barzilai and Borwein Jonathan M. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

[3] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. Spectral projected gradient methods: Review and perspectives. *Journal of Statistical Software*, 60(3): 1–21, 2014.

[4] Dominik Csiba, Zheng Qu, and Peter Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. In *ICML*, pages 674–683, 2015.

[5] Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[6] Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[7] Robert Mansel Gower and Peter Richtárik. Stochastic dual ascent for solving linear systems. *arXiv preprint arXiv:1512.06890*, 2015.

[8] Ching-Pei Lee and Stephen J. Wright. Random permutations fix a worst case for cyclic coordinate descent. *arXiv:1607.08320*, 2016.

[9] Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *FOCS*, 2013.

[10] Dennis Leventhal and Adrian Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35:641–654, 2010.

[11] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

[12] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[13] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001. URL `https://doi.org/10.1137/100802001`. First appeared in 2010 as CORE discussion paper 2010/2.

[14] Yurii Nesterov and Sebastian Stich. Efficiency of accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1): 110–123, 2017.

[15] Julie Nutini, Mark Schmidt, Issam H. Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *ICML*, 2015.

[16] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1 – 17, 1964.

[17] Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling I: algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857, 2016.

[18] Peter Richtárik and Martin Takáč. Stochastic reformulations of linear systems: Algorithms and convergence theory. *arXiv preprint arXiv:1706.01108*, 2017.

[19] Peter Richtárik and Martin Takáč. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243, 2016.

[20] Peter Richtárik and Peter Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1):433–484, 2016.

[21] Stephen Tu, Shivaram Venkataraman, Ashia C. Wilson, Alex Gittens, Michael I. Jordan, and Benjamin Recht. Breaking locality accelerates block Gauss-Seidel. In *ICML*, 2017.

# Appendix

## A Extra Experiments

In this section we report on some additional experiments which shed more light on the behaviour of our methods.

### A.1 Performance on SSCD on A with three clusters eigenvalues

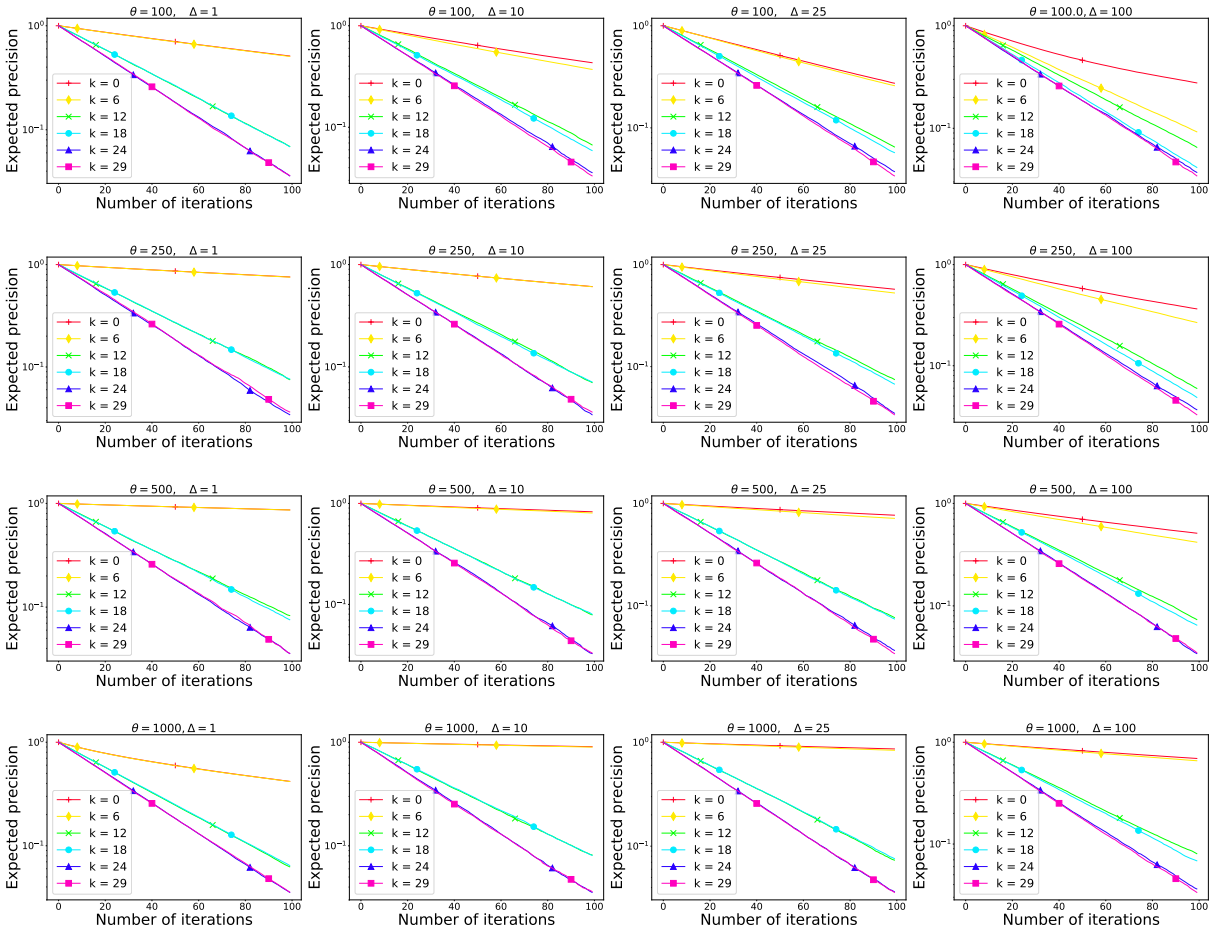


Figure 4: Expected precision $\mathbb{E}\left[\frac{||x_t - x_*||^2_{\mathbf{A}}}{||x_0 - x_*||^2_{\mathbf{A}}}\right]$ versus the number of iterations of SSCD for symmetric positive definite matrices $\mathbf{A}$ of size $30 \times 30$ with different structures of spectrum. The spectrum of $\mathbf{A}$ consists of 3 equally sized clusters of eigenvalues; one in the interval $(10, 10 + \Delta)$, the second in the interval $(\theta, \theta + \Delta)$ and the third in the interval $(2\theta, 2\theta + \Delta)$. We show results for 16 combinations of $\theta$ and $\Delta$: $\Delta \in \{1, 10, 25, 100\}$ and $\theta \in \{100, 250, 500, 1000\}$.

In Figure 4 we report on experiments similar to those performed in Section 4.1, but on data matrix $\mathbf{A} \in \mathbb{R}^{30\times 30}$ whose eigenvalues belong to three clusters, with 10 eigenvalues in each. We can observe that the SSCD methods can be grouped into three categories: slow,

fast, and very fast, depending on whether $k$ corresponds to the smallest 10 eigenvalues, the next cluster of 10 eigenvalues, or the 10 largest eigenvalues. That is, there are *two phase transitions*.

### A.2 Exponentially decaying eigenvalues

We now consider matrix $\mathbf{A} \in \mathbb{R}^{10\times 10}$ with eigenvalues $2^0, 2^1, \ldots, 2^9$. We apply SSCD with increasing values of $k$ (see Figure 5).

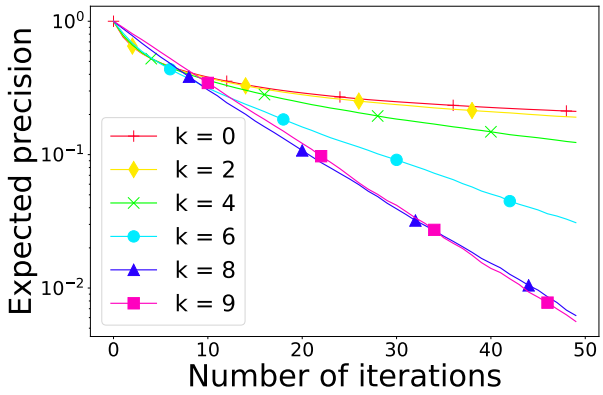


Figure 5: Expected precision $\mathbb{E}\left[\frac{||x_t-x_*||^2_{\mathbf{A}}}{||x_0-x_*||^2_{\mathbf{A}}}\right]$ versus the number of iterations of SSCD for symmetric positive definite matrix $\mathbf{A}$ of size $10 \times 10$.

We can see that the *performance boost accelerates as $k$ increases*. So, while one may not expect much speed-up for very small $k$, there will be substantial speed-up for moderate values of $k$. This is *predicted* by our theory. Indeed, consulting Table 3 (last column), we have $\alpha = 1/2$, and hence for $k = 0$ the theoretical rate is $\tilde{\mathcal{O}}(\frac{1}{\alpha^9})$. For general $k$ we have $\tilde{\mathcal{O}}(\frac{1}{\alpha^{9-k}})$. So, the speedup for value $k > 0$ compared to the baseline case of $k = 0$ (=RCD) is $2^k$, i.e., *exponential*.

# B Proofs

In this section we provide proofs of the statements from the main body of the paper. Table 4 provides a guide on where the proof of the various results can be found.

| Result | Section |
|---|---|
| Lemma 1 | B.1 |
| Theorem 2 | B.2 |
| Theorem 3 | B.3 |
| Theorem 4 | B.4 |
| Theorem 5 | B.5 |
| Theorem 6 | B.6 |
| Theorem 7 | B.7 |
| Theorem 8 | B.8 |
| Lemma 9 | B.9 |
| Theorem 10 | B.10 |

Table 4: Proof of lemmas and theorems stated in the main paper.

## B.1 Proof of Lemma 1

The result follows from Theorem 4.8(i) in [18] with the choice $\mathbf{B} = \mathbf{A}$. Note that since $x_* = \mathbf{A}^{-1}b$ is the unique solution of $\mathbf{A}x = b$, it is equal to the projection of $x_0$ onto the solution space of $\mathbf{A}x = b$, as required by the assumption in Theorem 4.8(i). It only remains to check that Assumption 3.5 (exactness) in [18] holds. In view of Theorem 3.6(iv) in [18], it suffices to check that the nullspace of $\mathbb{E}[\mathbf{H}]$ is trivial. However, this is equivalent to the assumption in Lemma 1 that $\mathbb{E}[\mathbf{H}]$ be invertible.

Finally, observe that

$$\begin{aligned}\tfrac{1}{2}\|x - x_*\|^2_{\mathbf{A}} &= \tfrac{1}{2}(x-x_*)^\top \mathbf{A}(x-x_*) = \tfrac{1}{2}x^\top\mathbf{A}x + \tfrac{1}{2}x_*^\top\mathbf{A}x_* - x^\top\mathbf{A}x_* \\ &= \tfrac{1}{2}x^\top\mathbf{A}x + \tfrac{1}{2}x_*^\top\mathbf{A}x_* - x^\top\mathbf{A}\mathbf{A}^{-1}b \overset{(1)}{=} f(x) + \tfrac{1}{2}x_*^\top\mathbf{A}x_* \\ &= f(x) - f(x_*).\end{aligned}$$

## B.2 Proof of Theorem 2

We will break down the proof into three steps.

1. First, let us show that Algorithm 2 is indeed SSD, as described in (4), i.e., $x_{t+1} = x_t - \frac{s_t^\top(\mathbf{A}x_t - b)}{s_t^\top\mathbf{A}s_t}s_t$. We known that $s_t = u_i$ with probability $1/n$. Since $\mathbf{A}u_i = \lambda_i u_i$, and assuming that at iteration $t$ we have $s_t = u_i$, we get

$$\begin{aligned}x_{t+1} &= x_t - \frac{u_i^\top(\mathbf{A}x_t - b)}{u_i^\top\mathbf{A}u_i}u_i = x_t - \frac{u_i^\top(\mathbf{A}x_t - b)}{\lambda_i}u_i \\ &= x_t - \frac{\lambda_i u_i^\top x_t - u_i^\top b}{\lambda_i}u_i = x_t - \left(u_i^\top x_t - \tfrac{u_i^\top b}{\lambda_i}\right)u_i.\end{aligned}$$

2. We now need to argue that the assumption that $\mathbb{E}[\mathbf{H}]$ is invertible is satisfied.

$$\mathbb{E}[\mathbf{H}] \overset{(8)}{=} \sum_{i=1}^n \frac{1}{n}\frac{u_i u_i^\top}{u_i^\top\mathbf{A}u_i} = \sum_{i=1}^n \frac{1}{n}\frac{u_i u_i^\top}{\lambda_i}. \tag{16}$$

Since $\mathbb{E}[\mathbf{H}]$ has positive eigenvalues $1/(n\lambda_i)$, it is invertible.

3. Applying Lemma 1, we get

$$(1-\lambda_{\max}(\mathbf{W}))^t\mathbb{E}[\|x_0-x_*\|_{\mathbf{A}}^2] \le \mathbb{E}[\|x_t-x_*\|_{\mathbf{A}}^2] \le (1-\lambda_{\min}(\mathbf{A}))^t\mathbb{E}[\|x_0-x_*\|_{\mathbf{A}}^2].$$

It remains to show that $\lambda_{\min}(\mathbf{W}) = \lambda_{\max}(\mathbf{W}) = \frac{1}{n}$. In view of (16), and since $\mathbf{A}^{1/2}u_i = \sqrt{\lambda_i}u_i$, we get

$$\mathbf{W} \overset{(11)}{=} \mathbf{A}^{1/2}\mathbb{E}[\mathbf{H}]\mathbf{A}^{1/2} \overset{(16)}{=} \mathbf{A}^{1/2}\sum_{i=1}^n \frac{1}{n}\frac{u_iu_i^\top}{\lambda_i}\mathbf{A}^{1/2} = \sum_{i=1}^n \frac{1}{n}\frac{\mathbf{A}^{1/2}u_iu_i^\top\mathbf{A}^{1/2}}{\lambda_i} = \frac{1}{n}\mathbf{I}.$$

## B.3 Proof of Theorem 3

Let $\mathbf{A}$ be a $2\times 2$ symmetric positive definite matrix:

$$\mathbf{A} = \begin{pmatrix} a & c \\ c & b \end{pmatrix}.$$

We know that $a, b > 0$, and $ab - c^2 > 0$. Assume that $s_t = e_1 = (1,0)^\top$ with probability $p > 0$ and $s_t = e_2 = (0,1)^\top$ with probability $q > 0$, where $p + q = 1$. Then

$$\mathbb{E}[\mathbf{H}] \overset{(8)}{=} p\frac{e_1e_1^\top}{e_1^\top\mathbf{A}e_1} + q\frac{e_2e_2^\top}{e_2^\top\mathbf{A}e_2} = \begin{pmatrix} \frac{p}{a} & 0 \\ 0 & \frac{q}{b} \end{pmatrix},$$

and therefore,

$$\mathbb{E}[\mathbf{H}]\mathbf{A} = \begin{pmatrix} p & p\frac{c}{a} \\ q\frac{c}{b} & q \end{pmatrix}.$$

Note that $\mathbb{E}[\mathbf{H}]\mathbf{A}$ has the same eigenvalues as $\mathbf{W} = \mathbf{A}^{1/2}\mathbb{E}[\mathbf{H}]\mathbf{A}^{1/2}$. We now find the eigenvalues of $\mathbb{E}[\mathbf{H}]\mathbf{A}$ by finding the zeros of the characteristic polynomial:

$$\det(\mathbb{E}[\mathbf{H}]\mathbf{A} - \lambda\mathbf{I}) = \det\begin{pmatrix} p-\lambda & p\frac{c}{a} \\ q\frac{c}{b} & q-\lambda \end{pmatrix} = \lambda^2 - \lambda + pq\left(1-\frac{c^2}{ab}\right) = 0$$

It can be seen that

$$\lambda_{\min}(\mathbb{E}[\mathbf{H}]\mathbf{A}) = \frac{1}{2} - \frac{1}{2}\sqrt{1-4pq\left(1-\frac{c^2}{ab}\right)} = \frac{1}{2} - \frac{1}{2}\sqrt{1-4p(1-p)\left(1-\frac{c^2}{ab}\right)}.$$

The expression $\lambda_{\min}(\mathbb{E}[\mathbf{H}]\mathbf{A})$ is maximized for $p = \frac{1}{2}$, independently of the values of $a, b$ and $c$.

## B.4 Proof of Theorem 4

Fix $n \ge 2$, and let $\Delta_n^+ := \{p \in \mathbb{R}^n : p > 0, \sum_i p_i = 1\}$ be the (interior of the) probability simplex. Further, let $\mathbf{A} = \mathrm{Diag}(\mathbf{A}_{11}, \mathbf{A}_{22}, \ldots, \mathbf{A}_{nn})$ be a diagonal matrix with positive diagonal entries.

The rate of RCD with any probabilities arises as a special case of Lemma 1. We therefore need to study the smallest eigenvalue of $\mathbf{W}$ (defined in (11)) as a function of $p = (p_1, \ldots, p_n)$. We have

$$\mathbf{H}(p) := \mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}] \overset{(8)}{=} \sum_i \frac{p_i}{\mathbf{A}_{ii}}e_ie_i^\top = \mathrm{Diag}(p_1/\mathbf{A}_{11}, p_2/\mathbf{A}_{22}, \ldots, p_n/\mathbf{A}_{nn}),$$

and hence

$$\mathbf{W} \overset{(11)}{=} \mathbf{W}(p) := \mathbf{A}^{1/2}\mathbf{H}(p)\mathbf{A}^{1/2} = \sum_{i=1}^{n} p_i e_i e_i^\top = \begin{pmatrix} p_1 & 0 & \dots & \\ 0 & p_2 & \dots & \\ \dots & \dots & \ddots & \\ 0 & 0 & \dots & p_n \end{pmatrix}. \tag{17}$$

Note that $\lambda_{\min}(\mathbf{W}(p)) \overset{(17)}{=} \lambda_{\min}(\mathrm{Diag}(p_1, p_2, \dots, p_n)) = \min_i p_i$, and thus

$$\max_{p \in \Delta_n^+} \lambda_{\min}(\mathbf{W}(p)) = \frac{1}{n}.$$

Clearly, the optimal probabilities are uniform: $p_i^* = \frac{1}{n}$ for all $i$.

## B.5 Proof of Theorem 5

We continue from the proof of Theorem 4.

1. Consider probabilities proportional to the diagonal elements: $p_i = \mathbf{A}_{ii}/\mathrm{Tr}(\mathbf{A})$ for all $i$. Choose $\mathbf{A}_{11} := t$, and $\mathbf{A}_{22} = \cdots = \mathbf{A}_{nn} = 1$. Then

$$\lambda_{\min}(\mathbf{W}(p)) \le p_2 = \frac{\mathbf{A}_{22}}{\mathrm{Tr}(\mathbf{A})} = \frac{1}{t+n-1} \longrightarrow 0 \text{ as } t \longrightarrow \infty.$$

2. Consider probabilities proportional to the squared row norms: $p_i = \|\mathbf{A}_{i:}\|^2/\mathrm{Tr}(\mathbf{A}^\top\mathbf{A})$ for all $i$. Choose $\mathbf{A}_{11} := t$, and $\mathbf{A}_{22} = \cdots = \mathbf{A}_{nn} = 1$. Then

$$\lambda_{\min}(\mathbf{W}(p)) \le p_2 = \frac{\mathbf{A}_{22}}{\mathrm{Tr}(\mathbf{A}^\top\mathbf{A})} = \frac{1}{t^2+n-1} \longrightarrow 0 \text{ as } t \longrightarrow \infty.$$

In both cases, $\frac{\lambda_{\min}(\mathbf{W}(p))}{\lambda_{\min}(\mathbf{W}(p^*))}$ can be made arbitrarily small by a suitable choice of $t$.

## B.6 Proof of Theorem 6

The rate of RCD with any probabilities arises as a special case of Lemma 1. We therefore need to study the smallest eigenvalue of $\mathbf{W}$ (defined in (11)). Since we wish to show that the rate can be bad, we will first prove a lemma bounding $\lambda_{\min}(\mathbf{W})$ from above.

**Lemma 11.** *Let $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be the eigenvalues of $\mathbf{A}$. Then*

$$\lambda_{\min}(\mathbf{W}) \le \frac{1}{n}\left(\prod_{k=1}^{n} \frac{\lambda_k}{\mathbf{A}_{kk}}\right)^{1/n}. \tag{18}$$

*Доказательство.* We have

$$\mathbf{W} \overset{(11)}{=} \mathbf{A}^{\frac{1}{2}}\mathbb{E}\left[\mathbf{H}\right]\mathbf{A}^{\frac{1}{2}} \overset{(8)}{=} \mathbf{A}^{\frac{1}{2}}\left(\sum_{k=1}^{n} \frac{p_k e_k e_k^\top}{\mathbf{A}_{kk}}\right)\mathbf{A}^{\frac{1}{2}} = \mathbf{A}^{\frac{1}{2}}\mathrm{Diag}\left(\frac{p_k}{\mathbf{A}_{kk}}\right)\mathbf{A}^{\frac{1}{2}}.$$

From the above we see that the determinant of $\mathbf{W}$ is given by

$$\det(\mathbf{W}) = \det(\mathbf{A}) \prod_{k=1}^{n} \frac{p_k}{\mathbf{A}_{kk}}. \tag{19}$$

On the other hand, we have the trivial bound

$$\det(\mathbf{W}) = \prod_{k=1}^{n} \lambda_k(\mathbf{W}) \geq (\lambda_{\min}(\mathbf{W}))^n. \tag{20}$$

Putting these together, we get an upper bound on $\lambda_{\min}(\mathbf{W})$ in terms of the eigenvalues and diagonal elements of $\mathbf{A}$:

$$\begin{aligned}
\lambda_{\min}(\mathbf{W}) \overset{(20)}{\leq} \sqrt[n]{\det(\mathbf{W})} &\overset{(19)}{=} \sqrt[n]{\det(\mathbf{A})} \cdot \sqrt[n]{\prod_{k=1}^{n} \frac{p_k}{\mathbf{A}_{kk}}} \\
&= \sqrt[n]{\det(\mathbf{A})} \cdot \sqrt[n]{\prod_{k=1}^{n} \frac{1}{\mathbf{A}_{kk}}} \cdot \sqrt[n]{\prod_{k=1}^{n} p_k} \\
&\overset{(*)}{\leq} \sqrt[n]{\det(\mathbf{A})} \cdot \sqrt[n]{\prod_{k=1}^{n} \frac{1}{\mathbf{A}_{kk}}} \cdot \frac{\sum_{k=1}^{n} p_k}{n} \\
&= \frac{\sqrt[n]{\det(\mathbf{A})}}{n} \cdot \sqrt[n]{\prod_{k=1}^{n} \frac{1}{\mathbf{A}_{kk}}} \\
&\overset{(20)}{=} \frac{1}{n} \sqrt[n]{\prod_{k=1}^{n} \frac{\lambda_k}{\mathbf{A}_{kk}}},
\end{aligned}$$

where (*) follows from the arithmetic-geometric mean inequality. □

**The Proof:** Let $\lambda_1, \ldots, \lambda_n$ are any positive real numbers. We now construct matrix $\mathbf{A} = \mathbf{M}\Lambda\mathbf{M}^\top$, where $\Lambda := \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$ and

$$\mathbf{M} := \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 & \cdots & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Clearly, $\mathbf{A}$ is symmetric. Since $\mathbf{M}$ is orthonormal, $\lambda_1, \ldots, \lambda_n$ are, by construction, the eigenvalues of $\mathbf{A}$. Hence, $\mathbf{A}$ is symmetric and positive definite. Further, note that the diagonal entries of $\mathbf{A}$ are related to its eigenvalues as follows:

$$\mathbf{A}_{kk} = \begin{cases} \frac{\lambda_1 + \lambda_2}{2}, & k = 1, 2; \\ \lambda_k, & \text{otherwise.} \end{cases} \tag{21}$$

Applying Lemma 11, we get the bound

$$\begin{aligned}\lambda_{\min}(\mathbf{W}) &\overset{(18)}{\leq} \frac{1}{n}\left(\prod_{k=1}^{n}\frac{\lambda_k}{\mathbf{A}_{kk}}\right)^{1/n} \\ &= \frac{1}{n}\left(\prod_{k=1}^{2}\frac{\lambda_k}{\mathbf{A}_{kk}}\cdot\prod_{k=3}^{n}\frac{\lambda_k}{\mathbf{A}_{kk}}\right)^{1/n} \\ &\overset{(21)}{=} \frac{1}{n}\left(\prod_{k=1}^{2}\frac{\lambda_k}{\mathbf{A}_{kk}}\right)^{1/n} \\ &\overset{(21)}{=} \frac{1}{n}\left(\frac{4\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^2}\right)^{1/n}.\end{aligned}$$

Let $c>0$ be such that $\lambda_1 = c\lambda_2$. Then $\frac{4\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^2} = \frac{4c}{(1+c)^2}$. If choose $c$ small enough so that $\frac{4c}{(1+c)^2} \leq \left(\frac{n}{T}\right)^n$, then $\lambda_{\min}(\mathbf{W}) \leq \frac{1}{T}$. The statement of the theorem follows.

## B.7 Proof of Theorem 7

Let $\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigenvalue decomposition of $\mathbf{W}$, where $\mathbf{U} = [u_1,\ldots,u_n]$ are the eigenvectors, $\lambda_1(\mathbf{W}) \leq \ldots \leq \lambda_n(\mathbf{W})$ are the eigenvalues and $\mathbf{\Lambda} = \mathrm{Diag}\left(\lambda_1(\mathbf{W}),\ldots,\lambda_n(\mathbf{W})\right)$. From Theorem 4.3 of [18] we get

$$\mathbb{E}\left[\mathbf{U}^\top\mathbf{A}^{1/2}(x_t - x_*)\right] = (\mathbf{I}-\mathbf{\Lambda})^t\mathbf{U}^\top\mathbf{A}^{1/2}(x_0-x_*). \tag{22}$$

Now we use Jensen's inequality and get

$$\mathbb{E}\left[\|x_t - x_*\|_{\mathbf{A}}^2\right] = \mathbb{E}\left[\left\|\mathbf{U}^\top\mathbf{A}^{1/2}(x_t-x_*)\right\|_2^2\right] \geq \left\|\mathbb{E}\left[\mathbf{U}^\top\mathbf{A}^{1/2}(x_t-x_*)\right]\right\|_2^2 \overset{(22)}{=} \left\|(\mathbf{I}-\mathbf{\Lambda})^t\mathbf{U}^\top\mathbf{A}^{1/2}(x_0-x_*)\right\|_2^2 \tag{23}$$

$$= \sum_{i=1}^{n}(1-\lambda_i(\mathbf{W}))^{2t}\left(u_i^\top\mathbf{A}^{1/2}(x_0-x_*)\right)^2 \geq (1-\lambda_1(\mathbf{W}))^{2t}\left(u_1^\top\mathbf{A}^{1/2}(x_0-x_*)\right)^2. \tag{24}$$

Now we take an example of matrix $\mathbf{A}$, for which we set $\lambda_{\min}(\mathbf{W}) \leq \frac{1}{T}$ for arbitrary $T>0$, like we did in Section B.6. We also choose $x_0 = x_* + \mathbf{A}^{-1/2}u_1$. For this choice of $\mathbf{A}$ and $x_0$ we get $\|x_0-x_*\|_{\mathbf{A}}^2 = \|u_1\|_2^2$ and

$$\mathbb{E}\left[\|x_t-x_*\|_{\mathbf{A}}^2\right] \geq (1-\lambda_1(\mathbf{W}))^{2t}\|u_1\|_2^2 \geq \left(1-\frac{1}{T}\right)^{2t}\|u_1\|_2^2 = \left(1-\frac{1}{T}\right)^{2t}\|x_0-x_*\|_{\mathbf{A}}^2. \tag{25}$$

## B.8 Proof of Theorem 8

We divide the proof into several steps.

1. Let us first show that SSCD converges with a linear rate for any choice of $\alpha>0$ and nonnegative $\{\beta_i\}$. Since SSCD arises as a special case of SD, it suffices to apply Lemma 1. In order to apply this lemma, we need to argue that $\mathcal{D} = \mathcal{D}(\alpha,\beta_1,\ldots,\beta_n)$ is a proper distribution. Indeed,

$$\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}] \overset{(8)}{=} \sum_{i=1}^{n} p_i \frac{e_i e_i^\top}{e_i^\top \mathbf{A} e_i} + \sum_{i=1}^{k} p_{n+i} \frac{u_i u_i^\top}{u_i^\top \mathbf{A} u_i}$$
$$= \frac{1}{C_k}\left(\alpha \mathbf{I} + \sum_{i=1}^{k} u_i u_i^\top \frac{\beta_i}{\lambda_i}\right) \tag{26}$$
$$\succeq \frac{\alpha}{C_k}\mathbf{I} \quad \succ \quad 0.$$

2. For the specific choice of parameters $\alpha = 1$ and $\beta_i = \lambda_{k+1} - \lambda_i$ we have

$$\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}] = \frac{1}{C_k}\left(\mathbf{I} + \sum_{i=1}^{k} u_i u_i^\top \frac{\lambda_{k+1} - \lambda_i}{\lambda_i}\right),$$

and $C_k = (k+1)\lambda_{k+1} + \sum\limits_{i=k+2}^{m} \lambda_i$. Therefore,

$$\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}] = \frac{1}{C_k}\left(\sum_{i=1}^{k} \lambda_{k+1} u_i u_i^\top + \sum_{i=k+1}^{n} \lambda_i u_i u_i^\top\right).$$

The minimal eigenvalue of this matrix, which has the same spectrum as $\mathbf{W}$, is

$$\lambda_{\min}(\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}]) = \frac{\lambda_{k+1}}{C_k} = \frac{\lambda_{k+1}}{(k+1)\lambda_{k+1} + \sum\limits_{i=k+2}^{n} \lambda_i}.$$

The main statement follows by applying Lemma 1.

3. We now show that the rate improves as $k$ increases. Indeed,

$$k + \frac{1}{\lambda_{k+1}} \sum_{i=k+1}^{m} \lambda_i = k + 1 + \frac{1}{\lambda_{k+1}} \sum_{i=k+2}^{m} \lambda_i \geq k + 1 + \frac{1}{\lambda_{k+2}} \sum_{i=k+2}^{m} \lambda_i.$$

By taking reciprocals, we get

$$\frac{\lambda_{k+2}}{(k+1)\lambda_{k+2} + \sum\limits_{i=k+2}^{m} \lambda_i} \geq \frac{\lambda_{k+1}}{k\lambda_{k+1} + \sum\limits_{i=k+1}^{m} \lambda_i}.$$

4. It remains to establish optimality of the specific parameter choice $\alpha = 1$ and $\beta_i = \lambda_{k+1} - \lambda_i$. Continuing from (26), we get

$$\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}] \overset{(26)}{=} \frac{1}{C_k}\left(\sum_{i=1}^{n} u_i u_i^\top \alpha\lambda_i + \sum_{i=1}^{k} u_i u_i^\top \beta_i\right) = \frac{1}{C_k}\left(\sum_{i=1}^{k} (\alpha\lambda_i + \beta_i) u_i u_i^\top + \sum_{i=k+1}^{n} \alpha\lambda_i u_i u_i^\top\right). \tag{27}$$

The eigenvalues of $\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}]$ are $\{\frac{\alpha\lambda_i+\beta_i}{C_k}\}_{i=1}^{k} \cup \{\frac{\alpha\lambda_i}{C_k}\}_{i=k+1}^{n}$. Let $\gamma$ be the smallest eigenvalue, i.e., $\gamma := \lambda_{\min}(\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}]) = \frac{\theta}{C_k}$, and $\Omega$ be the largest eigenvalue, i.e., $\Omega := \lambda_{\max}(\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}]) = \frac{\Delta}{C_k}$, where $\theta$ and $\Delta$ are appropriate constants. There are now two options.

(a) $\gamma = \frac{\alpha\lambda_{k+1}}{C_k}$. Then $\alpha\lambda_i + \beta_i \geq \alpha\lambda_{k+1}$ for $i \in \{1, \ldots, k\}$. In this case we obtain:

$$C_k = \alpha \mathrm{Tr}\left(\mathbf{A}\right) + \sum_{i=1}^{k} \beta_i = \sum_{i=1}^{k} (\alpha\lambda_i + \beta_i) + \alpha \sum_{i=k+1}^{n} \lambda_i \geq \alpha \left( k\lambda_{k+1} + \sum_{i=k+1}^{n} \lambda_i \right) \tag{28}$$

and therefore

$$\gamma \leq \frac{\lambda_{k+1}}{k\lambda_{k+1} + \sum\limits_{i=k+1}^{n} \lambda_i}. \tag{29}$$

(b) $\gamma = \frac{\alpha\lambda_j + \beta_j}{C_k} = \frac{\theta}{C_k}$ for some $j \in \{1, \ldots, k\}$. Then

$$C_k = \alpha \mathrm{Tr}\left(\mathbf{A}\right) + \sum_{i=1}^{k} \beta_i = \sum_{i=1}^{k} (\alpha\lambda_i + \beta_i) + \alpha \sum_{i=k+1}^{n} \lambda_i \geq k\theta + \alpha \sum_{i=k+1}^{n} \lambda_i \tag{30}$$

whence

$$\gamma \leq \frac{\theta}{k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i}. \tag{31}$$

Note that the function $f(\theta) = \frac{\theta}{k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i}$ increases monotonically:

$$f'(\theta) = \frac{1}{k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i} - \frac{k\theta}{(k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i)^2} = \frac{\alpha \sum\limits_{i=k+1}^{n} \lambda_i}{(k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i)^2} > 0. \tag{32}$$

From this and inequality $\alpha\lambda_{k+1} \geq \theta$ we get

$$\gamma \leq \frac{\alpha\lambda_{k+1}}{\alpha(k\lambda_{k+1} + \sum\limits_{i=k+1}^{n} \lambda_i)} = \frac{\lambda_{k+1}}{k\lambda_{k+1} + \sum\limits_{i=k+1}^{n} \lambda_i}. \tag{33}$$

In both possible cases we have shown that

$$\lambda_{\min}(\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}]) \leq \frac{\lambda_{k+1}}{k\lambda_{k+1} + \sum\limits_{i=k+1}^{n} \lambda_i}.$$

So, it is the optimal rate in this family of methods. Optimal distribution is unique and it is:

$$s \sim \mathcal{D} \quad \Leftrightarrow \quad s = \begin{cases} e_i & \text{with probability } p_i = \frac{\mathbf{A}_{ii}}{C_k} \quad i = 1, 2, \ldots, n \\ u_i & \text{with probability } p_{n+i} = \frac{\lambda_{k+1} - \lambda_i}{C_k} \quad i = 1, 2, \ldots, k, \end{cases} \tag{34}$$

where $C_k = k\lambda_{k+1} + \sum\limits_{i=k+1}^{n} \lambda_i$.

## B.9 Proof of Lemma 9

The steps are analogous to the proof of Lemma 1.

## B.10 Proof of Theorem 10

Let $C_k = (k+1)\lambda_{k+1} + \sum\limits_{i=k+2}^{n} \lambda_i$ $\gamma = \frac{\theta}{C_k}$ — the minimal eigenvalue of the matrix $\mathbf{W}$ and $\Omega = \frac{\Delta}{C_k}$ — the maximal eigenvalue of the matrix $\mathbf{W}$. The optimal rate of the method [18] is

$$r(\tau) = \frac{\gamma}{\frac{1}{\tau} + \left(1 - \frac{1}{\tau}\right)\Omega} = \frac{\theta}{\frac{1}{\tau}C_k + \left(1 - \frac{1}{\tau}\right)\Delta}. \tag{35}$$

From the Section B.8 we have

$$\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{AH}] = \frac{1}{C_k}\left(\sum_{i=1}^{k} \lambda_{k+1} u_i u_i^\top + \sum_{i=k+1}^{n} \lambda_i u_i u_i^\top\right).$$

There are two options.

1. $\gamma = \frac{\alpha\lambda_{k+1}}{C_k}$. Then $\alpha\lambda_i + \beta_i \geq \alpha\lambda_{k+1}$ for $i \in \{1, \ldots, k\}$ and $\Delta \geqslant \alpha\lambda_n$. In this case we obtain:

$$C_k = \alpha \mathrm{Tr}\left(\mathbf{A}\right) + \sum_{i=1}^{k} \beta_i = \sum_{i=1}^{k} (\alpha\lambda_i + \beta_i) + \alpha \sum_{i=k+1}^{n} \lambda_i \geq \alpha\left(k\lambda_{k+1} + \sum_{i=k+1}^{n} \lambda_i\right) \tag{36}$$

and therefore

$$r(\tau) \leq \frac{\alpha\lambda_{k+1}}{\frac{\alpha}{\tau}\left(k\lambda_{k+1} + \sum\limits_{i=k+1}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\alpha\lambda_n} = \frac{\lambda_{k+1}}{\frac{1}{\tau}\left(k\lambda_{k+1} + \sum\limits_{i=k+1}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\lambda_n}. \tag{37}$$

2. $\gamma = \frac{\alpha\lambda_j + \beta_j}{C_k} = \frac{\theta}{C_k}$ for some $j \in \{1, \ldots, k\}$. Then

$$C_k = \alpha \mathrm{Tr}\left(\mathbf{A}\right) + \sum_{i=1}^{k} \beta_i = \sum_{i=1}^{k} (\alpha\lambda_i + \beta_i) + \alpha \sum_{i=k+1}^{n} \lambda_i \geq k\theta + \alpha \sum_{i=k+1}^{n} \lambda_i, \quad \Delta \geq \alpha\lambda_n \tag{38}$$

whence

$$r(\tau) \leqslant \frac{\theta}{\frac{1}{\tau}\left(k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\alpha\lambda_n}. \tag{39}$$

Note that the function $f(\theta) = \frac{\theta}{\frac{1}{\tau}\left(k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\alpha\lambda_n}$ increases monotonically:

$$\begin{aligned} f'(\theta) &= \frac{1}{\frac{1}{\tau}\left(k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\alpha\lambda_n} - \frac{\frac{k}{\tau}\theta}{\left(\frac{1}{\tau}\left(k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\alpha\lambda_n\right)^2} \\ &= \frac{\frac{\alpha}{\tau} \sum\limits_{i=k+1}^{n} \lambda_i + \left(1 - \frac{1}{\tau}\right)\alpha\lambda_n}{\left(\frac{1}{\tau}\left(k\theta + \alpha \sum\limits_{i=k+1}^{n} \lambda_i\right) + \left(1 - \frac{1}{\tau}\right)\alpha\lambda_n\right)^2} > 0. \end{aligned} \tag{40}$$

From this and inequality $\alpha\lambda_{k+1} \geq \theta$ we get

$$r(\tau) \leq \frac{\alpha\lambda_{k+1}}{\frac{1}{\tau}\left(\alpha k\lambda_{k+1} + \alpha\sum\limits_{i=k+1}^{n}\lambda_i\right) + \left(1-\frac{1}{\tau}\right)\alpha\lambda_n} = \frac{\lambda_{k+1}}{\frac{1}{\tau}\left(k\lambda_{k+1} + \sum\limits_{i=k+1}^{n}\lambda_i\right) + \left(1-\frac{1}{\tau}\right)\lambda_n}. \tag{41}$$

For both possible cases we shown that $r(\tau) \leq \frac{\lambda_{k+1}}{\frac{1}{\tau}\left(k\lambda_{k+1}+\sum\limits_{i=k+1}^{n}\lambda_i\right)+\left(1-\frac{1}{\tau}\right)\lambda_n}$. So, it is the optimal rate in this family of methods. Note that $\alpha$ could be any positive number. Optimal distribution is unique and it is:

$$s \sim \mathcal{D} \quad \Leftrightarrow \quad s = \begin{cases} e_i & \text{with probability } p_i = \frac{\mathbf{A}_{ii}}{C_k} \quad i = 1, 2, \ldots, n \\ u_i & \text{with probability } p_{n+i} = \frac{\lambda_{k+1}-\lambda_i}{C_k} \quad i = 1, 2, \ldots, k, \end{cases} \tag{42}$$

where $C_k = k\lambda_{k+1} + \sum\limits_{i=k+1}^{n}\lambda_i$. For $k = 0$ we obtain mRCD, for $k = n-1$ we get the optimal rate $\frac{\frac{1}{n}}{\frac{1}{\tau}+(1-\frac{1}{\tau})\frac{1}{n}}$ and rate increases when $k$ increases.

# C Results mentioned informally in the paper

## C.1 Adding "largest" eigenvectors does not help

In Section 3.1 describing the SSCD method we have argued, without supplying any detail, that it does not make sense to consider replacing the $k$ "smallest" eigenvectors with a few "largest" eigenvectors. Here we make this statement precise, and prove it.

Fix $k \in \{0, 1, \ldots, n-1\}$ and consider running stochastic descent with the distribution $\mathcal{D}$ defined via

$$s \sim \mathcal{D} \quad \Leftrightarrow \quad s = \begin{cases} e_i & \text{with probability } p_i = \frac{\alpha\mathbf{A}_{ii}}{C_k} \quad i = 1, 2, \ldots, n \\ u_i & \text{with probability } p_{n-k+i} = \frac{\beta_i}{C_k} \quad i = k+1, k+2, \ldots, n, \end{cases} \tag{43}$$

where $C_k = \alpha\mathrm{Tr}\left(\mathbf{A}\right) + \sum\limits_{i=k+1}^{n}\beta_i$ and for $\beta_i \geq 0$ for $i \in \{1, 2, \ldots, k\}$.

That is, we consider "enriching" RCD with a collection of a $n-k$ eigenvectors corresponding to the $n-k$ largest eigenvectors of $\mathbf{A}$. We have the following negative result, which loosely speaking says that it is not worth enriching RCD with such vectors.

**Theorem 12.** *The optimal parameters of the above method are $k = n$ or $\beta_i = 0$ for all $i = k+1, \ldots, n$.*

*Доказательство.* We follow similar steps as in the proof of Theorem 8. In this setting we have

$$\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}] = \frac{1}{C_k}\left(\alpha\mathbf{I} + \sum_{i=k+1}^{n}\frac{\beta_i}{\lambda_i}u_iu_i^\top\right),$$

whence

$$\mathbf{A}\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}] = \frac{1}{C_k}\left(\alpha\mathbf{A} + \sum_{i=k+1}^{n}\beta_iu_iu_i^\top\right) = \frac{1}{C_k}\left(\sum_{i=1}^{k}\alpha\lambda_iu_iu_i^\top + \sum_{i=k+1}^{n}(\beta_i+\alpha\lambda_i)u_iu_i^\top\right)$$

and

$$\lambda_{\min}\left(\mathbf{A}\mathbb{E}_{s\sim\mathcal{D}}[\mathbf{H}]\right) = \frac{\alpha\lambda_1}{C_k} \leq \frac{\alpha\lambda_1}{\alpha \mathrm{Tr}\left(\mathbf{A}\right)} = \frac{\lambda_1}{\mathrm{Tr}\left(\mathbf{A}\right)}.$$

It means that the best rate in this family of methods is obtained when $k = n$ or $\beta_i = 0$ for all $i = k+1, \ldots, n$. □

So, to use spectral information about $n-k$ last eigenvectors we should use more complicated distributions (for instance, one may need to replace $\alpha$ by $\alpha_i$).

## C.2 Stochastic Conjugate Descent

The lemma below was referred to in Section 2.2. As explained in that section, this lemma can be used to argue that stochastic conjugate descent achieves the same rate as SSD: $\mathcal{O}(n \log \frac{1}{\epsilon})$.

**Lemma 13.** *Let $\{v_1 \ldots v_n\}$ be an $\mathbf{A}$-orthonormal system:*

$$v_i^\top \mathbf{A} v_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

*If distribution $\mathcal{D}$ consists of vectors $v_i$ chosen with uniform probabilities, then $\lambda_{\min}(\mathbf{W}) = \frac{1}{n}$*

*Доказательство.* That is,

$$\mathbf{W} = \mathbf{A}^{1/2}\mathbb{E}[\mathbf{H}]\mathbf{A}^{1/2} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{A}^{1/2} v_i v_i^\top \mathbf{A}^{1/2}}{v_i^\top \mathbf{A} v_i} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}^{1/2} v_i v_i^\top \mathbf{A}^{1/2}. \tag{44}$$

Making a substitution $u_i = \mathbf{A}^{1/2} s_i$, we get

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n} u_i u_i^\top = \frac{1}{n}\mathbf{I}, \tag{45}$$

because $\{u_1 \ldots u_n\}$ is orthonormal system. □

# D Inexact Stochastic Conjuagate Descent

In Section 2.2 we stated, that we can achieve an optimal rate of stochastic descent by using uniform distribution over a set of $n$ $\mathbf{A}$-conjugate directions. In this section we consider the case when $\mathbf{A}$-conjugate directions are computed approximately.

More formally, we consider a system of vectors $v_1, \ldots, v_n$, which satisfies $\left|v_i^\top \mathbf{A} v_j\right| \leq \varepsilon$ for $i \neq j$ and $v_i^\top \mathbf{A} v_i = 1$ for some parameter $\varepsilon > 0$. Further we'll call such vectors $\varepsilon$-approximate $\mathbf{A}$-conjugate vectors.

Now we formalize the idea of using approximate $\mathbf{A}$-conjugate directions in Stochastic Conjugate Descent, which leads to Algorithm 7.

**Algorithm 7** Inexact Stochastic Conjugate Descent (iSconD)

**Initialize:** $x_0 \in \mathbb{R}^n$; $v_1, \ldots, v_n$: $\varepsilon$-approximate $\mathbf{A}$-conjugate directions
**for** $t = 0, 1, 2, \ldots$ **do**
  Choose $i \in [n]$ uniformly at random
  Set $x_{t+1} = x_t - v_i^\top (\mathbf{A}x_t - b)\, v_i$
**end for**

For this algorithm we are going to obtain rate $\mathcal{O}(n \log \frac{1}{\epsilon})$, the optimal rate for stochastic descent.

## D.1 Lemma

**Lemma 14.** *Let* $\mathbf{S} = [v_1, \ldots, v_n]$*, where* $v_1, \ldots, v_n$ *are* $\varepsilon$*-approximate* $\mathbf{A}$*-conjugate vectors. If* $\varepsilon$ *satisfies*

$$\varepsilon < \frac{1}{n-1} \tag{46}$$

*then* $\tilde{\mathbf{I}} := \mathbf{S}^\top \mathbf{A}\mathbf{S}$ *is positive definite matrix and*

$$\lambda_{\min}(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}) \geq 1 - \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)} \tag{47}$$

$$\lambda_{\max}(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}) \leq 1 + \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)} \tag{48}$$

*Доказательство.* For unit vector $x$ we can write

$$x^\top \tilde{\mathbf{I}} x = \sum_{i,l} x_i x_l \tilde{\mathbf{I}}_{il} = 1 + \sum_{i,l:i\neq l} x_i x_l \tilde{\mathbf{I}}_{il} \geq 1 - \varepsilon \sum_{i,l:i\neq l} \frac{1}{2}(x_i^2 + x_l^2) = 1 - \varepsilon(n-1).$$

Under condition (46) we get $x^\top \tilde{\mathbf{I}} x > 0$ for any $x$, which proves the first part of lemma.

Since $\mathbf{S}^\top\mathbf{A}\mathbf{S}$ is positive definite, vectors $\mathbf{A}^{1/2}v_1, \ldots, \mathbf{A}^{1/2}v_n$ are linearly independent. Any unit vector $x$ may be represented as $x = \mathbf{A}^{1/2}\mathbf{S}\alpha$ with normalization condition:

$$1 = x^\top x = \alpha^\top \tilde{\mathbf{I}} \alpha = \alpha^\top \alpha + \sum_{i,l:i\neq l} \tilde{\mathbf{I}}_{il}\alpha_i\alpha_l, \tag{49}$$

or

$$\alpha^\top \alpha = 1 - \sum_{i,l:i\neq l} \tilde{\mathbf{I}}_{il}\alpha_i\alpha_l. \tag{50}$$

Now we can analyse spectrum of matrix $\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}$.

$$x^\top \mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}x = \alpha^\top \mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{S}^\top\mathbf{A}\mathbf{S}\alpha = \alpha^\top \tilde{\mathbf{I}}^2\alpha = \left\|\tilde{\mathbf{I}}\alpha\right\|_2^2 = \sum_{i=1}^n \left(\sum_{l=1}^n \tilde{\mathbf{I}}_{il}\alpha_l\right)^2 =$$

$$= \sum_{i=1}^n \left(\alpha_i + \sum_{l:l\neq i} \tilde{\mathbf{I}}_{il}\alpha_l\right)^2 = \sum_{i=1}^n \left(\alpha_i^2 + 2\alpha_i \sum_{l:l\neq i} \tilde{\mathbf{I}}_{il}\alpha_l + \left(\sum_{l:l\neq i} \tilde{\mathbf{I}}_{il}\alpha_l\right)^2\right).$$

Using (50) we get

$$x^\top \mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}x = 1 + \underbrace{\sum_{i,l:l\neq i} \tilde{\mathbf{I}}_{il}\alpha_i\alpha_l}_{R_1} + \underbrace{\sum_{i=1}^n \left(\sum_{l:l\neq i} \tilde{\mathbf{I}}_{il}\alpha_l\right)^2}_{R_2} = 1 + R_1 + R_2 \tag{51}$$

To estimate $|R_1|$ and $|R_2|$ we need to estimate $\alpha^\top\alpha$ using (50):

$$\alpha^\top\alpha \leq 1 + \varepsilon \sum_{i,l:i\neq l} \frac{\alpha_i^2 + \alpha_l^2}{2} = 1 + \varepsilon(n-1)\alpha^\top\alpha,$$

which under condition (46) implies that $\alpha^\top\alpha \leq \frac{1}{1-\varepsilon(n-1)}$. Now we can estimate $|R_1|$ and $|R_2|$.

$$R_1 \leq \varepsilon \sum_{i,l:i\neq l} \frac{\alpha_i^2 + \alpha_l^2}{2} = \varepsilon(n-1)\alpha^\top\alpha \leq \frac{\varepsilon(n-1)}{1-\varepsilon(n-1)} \tag{52}$$

$$R_2 \leq \sum_{i=1}^{n}(n-1)\sum_{l:l\neq i}\alpha_l^2\varepsilon^2 = \varepsilon^2(n-1)^2\alpha^\top\alpha \leq \frac{\varepsilon^2(n-1)^2}{1-\varepsilon(n-1)} \tag{53}$$

Finally from (51), (52) and (53) we get

$$\lambda_{\min}(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}) \geq 1 - \frac{\varepsilon(n-1)+\varepsilon^2(n-1)^2}{1-\varepsilon(n-1)} = 1 - \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)} \tag{54}$$

$$\lambda_{\max}(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}) \leq 1 + \frac{\varepsilon(n-1)+\varepsilon^2(n-1)^2}{1-\varepsilon(n-1)} = 1 + \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)} \tag{55}$$

□

**Corollary 14.1.** *If $\varepsilon < \frac{\sqrt{2}-1}{(n-1)}$ then $\lambda_{\min}(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}) > 0$ and condition number of $\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}$ has the following bound:*

$$\frac{\lambda_{\max}(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2})}{\lambda_{\min}(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2})} < \frac{1+\varepsilon^2(n-1)^2}{1-2\varepsilon(n-1)-\varepsilon^2(n-1)^2} \tag{56}$$

## D.2 Rate of convergence

The following theorem gives the rate of convergence of iSconD.

**Theorem 15.** *Let $\mathbf{S} = [v_1 \dots v_n]$, where $\{v_1 \dots v_n\}$ is $\varepsilon$-approximate $\mathbf{A}$-conjugate system. If $\varepsilon \leq \frac{1}{3(n-1)}$ then $\lambda_{\min}(\mathbf{W}) > \frac{1}{3n}$, which means that the rate of iSconD is $\mathcal{O}(n\log\frac{1}{\epsilon})$.*

*Доказательство.* As in Lemma 13, we can show that $\mathbf{W} = \frac{1}{n}\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}$, where $\mathbf{S} = [v_1 \dots v_n]$. Using bound (47) and Corollary 14.1, we get

$$\lambda_{\min}(\mathbf{W}) > \frac{1}{n}\left(1 - \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)}\right) \tag{57}$$

for small enough $\varepsilon$ (see Corollary 14.1). For $\varepsilon = \frac{1}{3(n-1)}$ we get $\lambda_{\min}(\mathbf{W}) > \frac{1}{3n}$. □

## D.3 Experiment

Figure 6 illustrates the theoretical results about iSonD. For this experiment we generate random orthogonal matrix $\mathbf{V}$ and random symmetric positive definite matrix $\tilde{\mathbf{I}}$, which satisfies $\tilde{\mathbf{I}}_{ii} = 1$, $\left|\tilde{\mathbf{I}}_{ij}\right| \leq \varepsilon$ for $i \neq j$. Columns of matrix $\mathbf{A}^{-1/2}\mathbf{V}\tilde{\mathbf{I}}^{1/2}$ were taken as approximate $\mathbf{A}$-conjugate vectors.

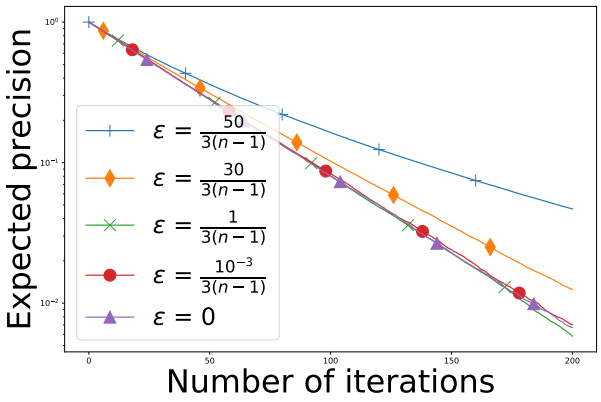


Figure 6: Expected precision $\mathbb{E}\left[\frac{\|x_t - x_*\|_{\mathbf{A}}^2}{\|x_0 - x_*\|_{\mathbf{A}}^2}\right]$ vs. the number of iterations of iSconD with different choices of parameter $\varepsilon$.

## D.4 Approximate solution without iterative methods

Note that the problem (1) is equivalent to the following problem of finding $x$ such that

$$\mathbf{A}x = b. \tag{58}$$

Let $\mathbf{S} = [v_1 \dots v_n]$ be a set of $\mathbf{A}$-conjugate vectors, i.e., $\mathbf{S}^\top \mathbf{A}\mathbf{S} = \mathbf{I}$. We can now find the solution to the linear system (58). Since $\mathbf{S}^\top b = \mathbf{S}^\top \mathbf{A}x = \mathbf{S}^\top \mathbf{A}\mathbf{S}\mathbf{S}^{-1}x = \mathbf{S}^{-1}x$, we conclude that

$$x = \mathbf{S}\mathbf{S}^\top b. \tag{59}$$

We will now show that unlike in the exact case, using formula (59) with $\varepsilon$-approximate $\mathbf{A}$-conjugate vectors does *not* lead to a precise solution of our problem.

**Lemma 16.** *Let $\mathbf{S} = [v_1 \dots v_n]$ be an $\varepsilon$-$\mathbf{A}$-orthonormal system. Let $x_* = \mathbf{A}^{-1}b$ be the solution of the linear system (58). Let $\hat{x}$ be an estimate of the solution, calculated with formula (59) using $\varepsilon$-approximate $\mathbf{A}$-conjugate vectors: $\hat{x} = \mathbf{S}\mathbf{S}^\top b$. If $\varepsilon < 1/(n-1)$, then*

$$\|\hat{x} - x_*\|_{\mathbf{A}} \leq \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)}\|x_*\|_{\mathbf{A}} \tag{60}$$

*Доказательство.* Note that $\mathbf{A}^{1/2}\hat{x} = \mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}\mathbf{A}^{1/2}x_* = \hat{\mathbf{I}}\mathbf{A}^{1/2}x_*$, where $\hat{\mathbf{I}} = \mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^\top\mathbf{A}^{1/2}$. From Lemma 14 we now get that

$$\left|\lambda_i(\hat{\mathbf{I}} - \mathbf{I})\right| \leq \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)}, \tag{61}$$

and hence

$$\left\|\hat{\mathbf{I}} - \mathbf{I}\right\|_2 \leq \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)}. \tag{62}$$

Therefore,

$$\|\hat{x} - x_*\|_{\mathbf{A}} = \left\|\mathbf{A}^{1/2}(\hat{x} - x_*)\right\|_2 = \left\|(\hat{\mathbf{I}} - \mathbf{I})\mathbf{A}^{1/2}x_*\right\|_2 \leq \left\|\hat{\mathbf{I}} - \mathbf{I}\right\|_2 \left\|\mathbf{A}^{1/2}x_*\right\|_2 \leq \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)}\|x_*\|_{\mathbf{A}}.$$

□

If we choose $\varepsilon = \frac{1}{3(n-1)}$, like we did in Theorem 15, we get the following precision:

$$\|\hat{x} - x_*\|_{\mathbf{A}} \leq \frac{2}{3} \|x_*\|_{\mathbf{A}}, \tag{63}$$

which is rather poor. However if we use Algorithm 7, we can get approximate solution with any precision (after enough iterations).

# E Inexact SSD: a method that is not a special case of stochastic descent

In Section 2.1 we defined Stochastic Spectral Descent (Algorithm 2). We now design a new method which will "try" to use the same iterations, but with *inexact* eigenvectors of $\mathbf{A}$. We call $w$ an inexact eigenvector of $\mathbf{A}$ if

$$\mathbf{A}w = \lambda w + \varepsilon \tag{64}$$

for some $\varepsilon$ and $\lambda > 0$ (inexact eigenvalue). Clearly, *any* vector can be written in the form (64). This idea leads to Algorithm 8.

**Algorithm 8** Inexact Stochastic Spectral Descent (iSSD)

**Initialize:** $x_0 \in \mathbb{R}^n$; $(w_1, \lambda_1), \dots (w_n, \lambda_n)$: inexact eigenvectors and eigenvalues of $\mathbf{A}$
**for** $t = 0, 1, 2, \dots$ **do**
  Choose $i \in [n]$ uniformly at random
  Set $x_{t+1} = x_t - \left( w_i^\top x_t - \frac{w_i^\top b}{\lambda_i} \right) w_i$
**end for**

Note that the above method is *not equivalent* to applying stochastic descent $\mathcal{D}$ being the uniform distribution over the inexact eigenvectors. This is because in arriving at SSD, we have used some properties of the eigenvectors and eigenvalues to simplify the calculation of the stepsize. The same simplifications do *not* apply for inexact eigenvectors. Nevertheless, we can formally run SSD, as presented in Algorithm 2, and replace the exact eigenvectors and eigenvalues by inexact versions thereof, thus capitalizing on the fast computation of stepsize which positively affects the cost of one iteration of the method. This leads to Algorithm 8.

Hence, in order to analyze the above method, we need to develop a completely new approach. We will show that Algorithm 8 converges only to a neighbourhood of the optimal solution.

## E.1 Lemmas

Further we will use the following notation: $\mathbf{S} = [w_1 \dots w_n]$ – inexact eigenvectors matrix, $\Lambda = \mathrm{Diag}(\lambda_1 \dots \lambda_n)$ – inexact eigenvalues matrix, $\mathbf{E} = [\varepsilon_1 \dots \varepsilon_n]$ – error matrix, $\tilde{\mathbf{A}} = \mathbf{S}\Lambda\mathbf{S}^\top$ – estimation of matrix $\mathbf{A}$. We also assume, that inexact eigenvectors are $\varepsilon$-approximate orthonormal for $\varepsilon < \frac{1}{n-1}$, i.e. $w_i^\top w_i = 1$, $\left| w_i^\top w_j \right| \leq \varepsilon$ for $i \neq j$.

The following lemma gives an answer to the question: how precise is $\tilde{\mathbf{A}}$ as an estimate of matrix $\mathbf{A}$?

**Lemma 17.** $\tilde{\mathbf{A}} = \hat{\mathbf{I}}\mathbf{A} - \mathbf{S}\mathbf{E}^\top$, *where matrix* $\hat{\mathbf{I}} = \mathbf{S}\mathbf{S}^\top$ *satisfies*

$$\left\|\hat{\mathbf{I}} - \mathbf{I}\right\|_2 \leq \varepsilon(n-1)\frac{1+\varepsilon(n-1)}{1-\varepsilon(n-1)}. \tag{65}$$

*Доказательство.* Indeed, the definition of inexact eigenvectors can be written in matrix form as $\mathbf{A}\mathbf{S} = \mathbf{S}\Lambda + \mathbf{E}$, from which follows that $\hat{\mathbf{I}}\mathbf{A} = \mathbf{S}\mathbf{S}^\top\mathbf{A} = \mathbf{S}\Lambda\mathbf{S}^\top + \mathbf{S}\mathbf{E}^\top$. Equality (65) follows immediately from Lemma 14. □

The next lemma gives a general recursion capturing one step of iSSD, shedding light on the convergence of the method.

**Lemma 18.** *Sequence of* $\{x_t\}$ *generated by inexact SSD satisfies equality*

$$\begin{aligned}\mathbb{E}\left\|x_{t+1} - x_*\right\|_{\mathbf{A}}^2 &= \left(1 - \frac{1}{n}\right)\mathbb{E}\left\|x_t - x_*\right\|_{\mathbf{A}}^2 + \frac{1}{n}\mathbb{E}\left[(x_t - x_*)^\top\Gamma(x_t - x_*)\right] \\ &+ \frac{1}{n}\left(\mathbb{E}\left\|x_t\right\|_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top}^2 + x_*^\top\mathbf{E}\Lambda^{-2}\mathbf{C}\mathbf{E}^\top x_*\right) - \frac{2}{n}\mathbb{E}\left[(x_t - x_*)^\top\mathbf{S}\mathbf{C}\Lambda^{-1}\mathbf{E}^\top x_*\right],\end{aligned}$$

*where* $\Gamma = (\mathbf{I} - \hat{\mathbf{I}})\mathbf{A} - \mathbf{S}\mathbf{E}^\top - \mathbf{E}\Lambda^{-1}\mathbf{E}^\top + \mathbf{S}\mathbf{C}\mathbf{S}^\top$ *and*

$$\mathbf{C} = \mathrm{Diag}\left(w_1^\top\varepsilon_1 \ldots w_n^\top\varepsilon_n\right). \tag{66}$$

*Доказательство.*

$$\begin{aligned}\left\|x_{t+1} - x_*\right\|_{\mathbf{A}}^2 &= \left\|x_t - x_* - \omega w_t w_t^\top(x_t - x_*) + \omega\frac{\varepsilon_t^\top x_*}{\lambda_t}w_t\right\|_{\mathbf{A}}^2 \\ &= \left\|x_t - x_*\right\|_{\mathbf{A}}^2 + \omega^2 w_t^\top\mathbf{A}w_t\left(w_t^\top(x_t - x_*) - \frac{\varepsilon_t^\top x_*}{\lambda_t}\right)^2 \\ &\quad + 2\omega(x_t - x_*)^\top\mathbf{A}w_t\left(\frac{\varepsilon_t^\top x_*}{\lambda_t} - w_t^\top(x_t - x_*)\right) \\ &= \left\|x_t - x_*\right\|_{\mathbf{A}}^2 + \omega^2(\lambda_t + w_t^\top\varepsilon_t)\left(w_t^\top(x_t - x_*) - \frac{\varepsilon_t^\top x_*}{\lambda_t}\right)^2 \\ &\quad + 2\omega(x_t - x_*)^\top(\lambda_t w_t + \varepsilon_t)\left(\frac{\varepsilon_t^\top x_*}{\lambda_t} - w_t^\top(x_t - x_*)\right) \\ &= \left\|x_t - x_*\right\|_{\mathbf{A}}^2 - \omega(2-\omega)(x_t - x_*)^\top\lambda_t w_t w_t^\top(x_t - x_*) + \omega^2\frac{x_*^\top\varepsilon_t\varepsilon_t^\top x_*}{\lambda_t} \\ &\quad + 2\omega\frac{(x_t - x_*)^\top\varepsilon_t\varepsilon_t^\top x_*}{\lambda_t} + \omega^2 w_t^\top\varepsilon_t\left(w_t^\top(x_t - x_*) - \frac{\varepsilon_t^\top x_*}{\lambda_t}\right)^2 \\ &\quad + 2(\omega - \omega^2)(x_t - x_*)^\top w_t\varepsilon_t^\top x_* - 2\omega(x_t - x_*)^\top w_t\varepsilon_t^\top(x_t - x_*) \\ &= \left\|x_t - x_*\right\|_{\mathbf{A}}^2 - \omega(2-\omega)(x_t - x_*)^\top\lambda_t w_t w_t^\top(x_t - x_*) + \left\|x_*(\omega - 1) + x_t\right\|_{\frac{\varepsilon_t\varepsilon_t^\top}{\lambda_t}} \\ &\quad - \left\|x_t - x_*\right\|_{\frac{\varepsilon_t\varepsilon_t^\top}{\lambda_t}} + 2\omega(x_t - x_*)^\top w_t\varepsilon_t^\top(x_*(2-\omega) - x_t) + \omega^2 w_t^\top\varepsilon_t\left(w_t^\top(x_t - x_*) - \frac{\varepsilon_t^\top}{\lambda_t}\right.\end{aligned}$$

Now we can take conditional expectation $\mathbb{E}[\,\cdot \mid x_t]$.

$$\begin{aligned}\mathbb{E}[\left\|x_{t+1} - x_*\right\|_{\mathbf{A}}^2 \mid x_t] = \left\|x_t - x_*\right\|_{\mathbf{A}}^2 - \frac{\omega(2-\omega)}{n}\left\|x_t - x_*\right\|_{\tilde{\mathbf{A}}}^2 + \frac{1}{n}\left\|x_*(\omega - 1) + x_t\right\|_{\Sigma}^2 - \frac{1}{n}\left\|x_t - x_*\right\|_{\Sigma}^2 - \\ - \frac{2\omega}{n}(x_t - x_*)^\top\mathbf{S}\mathbf{E}^\top(x_t - (2-\omega)x_*) + \frac{\omega^2}{n}\sum_{i=1}^n w_i^\top\varepsilon_i\left(w_i^\top(x_t - x_*) - \frac{\varepsilon_i^\top x_*}{\lambda_i}\right)^2,\end{aligned}$$

where $\Sigma = \mathbf{E}\Lambda^{-1}\mathbf{E}^\top$.

Now we set $\omega = 1$ and use Lemma 17.

$$\mathbb{E}[\|x_{t+1} - x_*\|_{\mathbf{A}}^2 \mid x_t] = \|x_t - x_*\|_{\mathbf{A}}^2 - \frac{1}{n}\|x_t - x_*\|_{\hat{\mathbf{A}}}^2 + \frac{1}{n}\|x_t\|_\Sigma^2 - \frac{1}{n}\|x_t - x_*\|_\Sigma^2 -$$
$$-\frac{2}{n}(x_t - x_*)^\top \mathbf{S}\mathbf{E}^\top (x_t - x_*) + \frac{1}{n}\sum_{i=1}^n w_i^\top \varepsilon_i \left(w_i^\top (x_t - x_*) - \frac{\varepsilon_i^\top x_*}{\lambda_i}\right)^2 =$$
$$= \|x_t - x_*\|_{\mathbf{A}}^2 \left(1 - \frac{1}{n}\right) + \frac{1}{n}(x_t - x_*)^\top \left((\mathbf{I} - \hat{\mathbf{I}})\mathbf{A} + \mathbf{S}\mathbf{E}^\top - 2\mathbf{S}\mathbf{E}^\top\right)(x_t - x_*) + \frac{1}{n}\|x_t\|_\Sigma^2 - \frac{1}{n}\|x_t - x_*\|_\Sigma^2 +$$
$$+\frac{1}{n}\sum_{i=1}^n w_i^\top \varepsilon_i \left(w_i^\top (x_t - x_*) - \frac{\varepsilon_i^\top x_*}{\lambda_i}\right)^2 =$$
$$= \left(1 - \frac{1}{n}\right)\|x_t - x_*\|_{\mathbf{A}}^2 + \frac{1}{n}(x_t - x_*)^\top \left((\mathbf{I} - \hat{\mathbf{I}})\mathbf{A} - \mathbf{S}\mathbf{E}^\top - \Sigma\right)(x_t - x_*) + \frac{1}{n}\|x_t\|_\Sigma^2 +$$
$$+\frac{1}{n}\sum_{i=1}^n w_i^\top \varepsilon_i \left(w_i^\top (x_t - x_*) - \frac{\varepsilon_i^\top x_*}{\lambda_i}\right)^2 =$$
$$= \left(1 - \frac{1}{n}\right)\|x_t - x_*\|_{\mathbf{A}}^2 + \frac{1}{n}(x_t - x_*)^\top \left((\mathbf{I} - \hat{\mathbf{I}})\mathbf{A} - \mathbf{S}\mathbf{E}^\top - \Sigma\right)(x_t - x_*) + \frac{1}{n}\|x_t\|_\Sigma^2 +$$
$$+\frac{1}{n}\|x_t - x_*\|_{\mathbf{S}\mathbf{C}\mathbf{S}^\top}^2 + \frac{1}{n}x_*^\top \left(\mathbf{E}\Lambda^{-2}\mathbf{C}\mathbf{E}^\top\right)x_* - \frac{2}{n}(x_t - x_*)^\top \mathbf{S}\mathbf{C}\Lambda^{-1}\mathbf{E}^\top x_*,$$

where $\mathbf{C} = \mathrm{Diag}\left(w_1^\top \varepsilon_1 \dots w_n^\top \varepsilon_n\right)$. We get

$$\mathbb{E}[\|x_{t+1} - x_*\|_{\mathbf{A}}^2 \mid x_t] = \left(1 - \frac{1}{n}\right)\|x_t - x_*\|_{\mathbf{A}}^2 + \frac{1}{n}(x_t - x_*)^\top \Gamma (x_t - x_*)$$
$$+\frac{1}{n}\left(\|x_t\|_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top}^2 + x_*^\top \mathbf{E}\Lambda^{-2}\mathbf{C}\mathbf{E}^\top x_* - 2(x_t - x_*)^\top \mathbf{S}\mathbf{C}\Lambda^{-1}\mathbf{E}^\top x_*\right),$$

where $\Gamma = (\mathbf{I} - \hat{\mathbf{I}})\mathbf{A} - \mathbf{S}\mathbf{E}^\top - \mathbf{E}\Lambda^{-1}\mathbf{E}^\top + \mathbf{S}\mathbf{C}\mathbf{S}^\top$. □

The following lemma describes which inexact eigenvalues are optimal for a fixed set of inexact eigenvectors.

**Lemma 19.** *Let $w_i$ be fixed. Then the choice*

$$\lambda_i = w_i^\top \mathbf{A} w_i \tag{67}$$

*minimizes $\|\varepsilon_i\|_2$ in $\lambda$, where $\epsilon_i := \|\mathbf{A}w_i - \lambda w_i\|_2$. Moreover, for this choice of $\lambda_i$ we get $w_i^\top \varepsilon_i = 0$.*

*Доказательство.* Minimizing $\|\mathbf{A}w_i - \lambda w_i\|_2^2$ in $\lambda$ gives (67). For this choice of $\lambda_i$ we get $w_i^\top \varepsilon_i = w_i^\top \mathbf{A} w_i - \lambda_i w_i^\top w_i = w_i^\top \mathbf{A} w_i - w_i^\top \mathbf{A} w_i = 0$. □

## E.2 Convergence

Choosing eigenvalues as defined in (67), and in view of (66), we see that $\mathbf{C} = 0$. From this and Lemma 18 we get

$$\mathbb{E}\|x_{t+1} - x_*\|_{\mathbf{A}}^2 = \left(1 - \frac{1}{n}\right)\mathbb{E}\|x_t - x_*\|_{\mathbf{A}}^2 + \frac{1}{n}\mathbb{E}\left[(x_t - x_*)^\top \Gamma (x_t - x_*)\right] + \frac{1}{n}\mathbb{E}\|x_t\|_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top}^2, \tag{68}$$

where $\Gamma = (\mathbf{I} - \hat{\mathbf{I}})\mathbf{A} - \mathbf{S}\mathbf{E}^\top - \mathbf{E}\Lambda^{-1}\mathbf{E}^\top$. From the Cauchy–Schwarz inequality we get

$$\frac{1}{n}\mathbb{E}\left\|x_t\right\|^2_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top} = \frac{1}{n}\mathbb{E}\left\|x_t - x_* + x_*\right\|^2_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top} \leq \frac{2}{n}\mathbb{E}\left\|x_t - x_*\right\|^2_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top} + \frac{2}{n}\mathbb{E}\left\|x_*\right\|^2_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top}, \tag{69}$$

which leads to

$$\mathbb{E}\left\|x_{t+1} - x_*\right\|^2_{\mathbf{A}} \leq \left(1 - \frac{1}{n}\right)\mathbb{E}\left\|x_t - x_*\right\|^2_{\mathbf{A}} + \frac{1}{n}\mathbb{E}\left[(x_t - x_*)^\top \mathbf{Q}(x_t - x_*)\right] + \frac{2}{n}\left\|x_*\right\|^2_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top}, \tag{70}$$

where $\mathbf{Q} = (\mathbf{I} - \hat{\mathbf{I}})\mathbf{A} - \mathbf{S}\mathbf{E}^\top + \mathbf{E}\Lambda^{-1}\mathbf{E}^\top$. Inequality (70) implies that

$$\mathbb{E}\left\|x_{t+1} - x_*\right\|^2_{\mathbf{A}} \leq \mathbb{E}\left\|x_t - x_*\right\|^2_{\mathbf{A}} + \frac{q-1}{n}\mathbb{E}\left\|x_t - x_*\right\|^2_{\mathbf{A}} + \frac{r_0}{n}, \tag{71}$$

where $q = \max \frac{z^\top \mathbf{Q} z}{z^\top \mathbf{A} z}$, $r_0 = 2\left\|x_*\right\|^2_{\mathbf{E}\Lambda^{-1}\mathbf{E}^\top}$.

If the errors $\varepsilon_1, \ldots, \varepsilon_n$ and $\varepsilon$ are small enough, we can make $q$ and $r_0$ arbitrarily small for fixed $x_*$. From (71) we can see that $\mathbb{E}\left\|x_{t+1} - x_*\right\|^2_{\mathbf{A}}$ is going to decrease as long as

$$\mathbb{E}\left\|x_t - x_*\right\|^2_{\mathbf{A}} \geq \frac{r_0}{1-q}. \tag{72}$$

Hence, for small enough $\varepsilon_1, \ldots, \varepsilon_n$ and parameter $\varepsilon$, iSSD will converge to a neighborhood of the optimal solution, with limited precision (72).