Introduction to Randomized Methods in Convex Optimization

(Lectures delivered at the Erwin Schrödinger Institute, Vienna)

Peter Richtárik

King Abdullah University of Science and Technology

February 22, 2019

Contents

Introduction to Randomized Methods in Convex Optimization
Peter Richtárik

1. Stochastic Reformulations of Linear Systems

Solving Very Large Linear Systems

In this lecture we are concerned with the problem of solving a **linear system.** In particular, consider the problem

$$\text{solve} \quad \mathbf{A}x = b, \qquad (1)$$

where $0 \neq \mathbf{A} \in \mathbb{R}^{m \times n}$, and m **is very large.**

Let $\mathbf{A}_{i:}$ denote the ith row of $\mathbf{A}$, and $\mathbf{A}_{:j}$ denote the jth column of $\mathbf{A}$. Let $b = (b_1, \dots, b_m)$. Problem (1) can also be written more explicitly as a system of m **linear equations:**

$$\begin{aligned} \mathbf{A}_{1:}x &= b_1 \\ \mathbf{A}_{2:}x &= b_2 \\ &\vdots \\ \mathbf{A}_{m:}x &= b_m. \end{aligned}$$

The ith equation in the system has the form

$$\sum_{j=1}^{m} \mathbf{A}_{ij} x_j = b_j.$$

Consistency

We shall assume throughout the lecture that:

Assumption 1

Linear system (1) *is* **consistent.** *In other words, it has a solution:*

$$\mathcal{L} \overset{def}{=} \{x \;:\; \mathbf{A}x = b\} \neq \emptyset.$$

Introduction

- We will present a fundamental and flexible way of **reformulating each consistent linear system** into a **stochastic problem.**
- Stochasticity is introduced in a controlled way, into an otherwise deterministic problem, as a decomposition tool which can be leveraged to design efficient, granular and scalable **randomized algorithms.**
- **Two parameters:**
  - **Distribution** $\mathcal{D}$ describing an ensemble of random matrices $\mathbf{S} \in \mathbb{R}^{m \times q}$.
  - **Symmetric positive definite matrix B** $\in \mathbb{R}^{n \times n}$**.**
- Presented approach and underlying theory support virtually all thinkable distributions $\mathcal{D}$. The choice of the distribution should ideally depend on the problem itself, as it will affect the complexity of the associated algorithms.
- In this specific setup (=linear systems), we can study many popular stochastic methods used in optimization and machine learning in a **unified way.** You will thus get strong foundations in the field.

Positive Definite Matrices, Inner Products and Norms

Positive Definite Matrices

Definition 1

Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a symmetric matrix.

(i) We say that $\mathbf{M}$ is **positive semidefinite** if

$$x^\top \mathbf{M} x \geq 0 \qquad \forall x \in \mathbb{R}^n.$$

We write this concisely as $\mathbf{M} \succeq 0$.

(ii) We say that $\mathbf{M}$ is **positive definite** if

$$x^\top \mathbf{M} x > 0 \qquad \forall 0 \neq x \in \mathbb{R}^n.$$

We write this concisely as $\mathbf{M} \succ 0$.

Inner Products and Norms

Inner Product in $\mathbb{R}^n$

Given a symmetric positive definite matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$, we equip the space $\mathbb{R}^n$ with the **Euclidean inner product** defined by

$$\langle x, y \rangle_{\mathbf{B}} \overset{\text{def}}{=} x^\top \mathbf{B} y = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \mathbf{B}_{ij} y_j, \qquad x, y \in \mathbb{R}^n.$$

Norm in $\mathbb{R}^n$

We also define the **induced norm:** $\|x\|_{\mathbf{B}} \overset{\text{def}}{=} \sqrt{\langle x, x \rangle_{\mathbf{B}}}$.

Remark: We also use the short-hand notation $\| \cdot \|$ to mean $\| \cdot \|_{\mathbf{I}}$, where $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. We shall sometimes refer to the quantity $\|x\|_{\mathbf{M}}$ with matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ being merely positive definite.

Stochastic Reformulations

Four Reformulations

We reformulate (1) into 4 seemingly different, but equivalent **stochastic problems:**

1. **Stochastic optimization problem** (2)
2. **Stochastic linear system** (4)
3. **Stochastic fixed point problem** (5)
4. **Probabilistic intersection problem** (6)

Reformulation 1: Stochastic Optimization Problem

Consider the **stochastic optimization problem**

$$\text{minimize} \quad f(x) \overset{\text{def}}{=} \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[f_{\mathbf{S}}(x)\right], \tag{2}$$

where

$$f_{\mathbf{S}}(x) \overset{\text{def}}{=} \frac{1}{2}\|\mathbf{A}x - b\|_{\mathbf{H}}^2 = \frac{1}{2}(\mathbf{A}x - b)^\top \mathbf{H}(\mathbf{A}x - b). \tag{3}$$

When solving the problem, we do not have (or do not wish to exercise, as it may be prohibitively expensive) explicit access to f, its gradient or Hessian. Rather, we can repeatedly sample $\mathbf{S} \sim \mathcal{D}$ and receive unbiased samples of these quantities at points of interest. That is, we may obtain local information about the **stochastic function $f_{\mathbf{S}}(x)$,** such as the **stochastic gradient $\nabla f_{\mathbf{S}}(x)$,** or the **stochastic Hessian $\nabla^2 f_{\mathbf{S}}(x)$.**

Reformulation 2: Stochastic Linear System

Consider the following **stochastic linear system:**

$$\text{solve} \quad \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}[\mathbf{H}]\,\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}[\mathbf{H}]\,b. \tag{4}$$

- The system arises by pre-multiplying the system (1) on both sides from the left by matrix $\mathbf{P} = \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}[\mathbf{H}]$.
- The **preconditioner** $\mathbf{P}$ is not assumed to be known explicitly.
- Instead, when solving the problem, we are able to sample $\mathbf{S} \sim \mathcal{D}$, obtaining an unbiased estimate of the preconditioner (not necessarily explicitly), $\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}$, for which we coin the name **stochastic preconditioner**. This gives us access to a random sample of system (4):
$$\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}b.$$
- This information can be obtained by repeatedly querying the stochastic sampling $\mathbf{S} \sim \mathcal{D}$ and utilized by an iterative algorithm.

Reformulation 3: Stochastic Fixed Point Problem

Let $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)$ denote the projection of x onto $\mathcal{L}_{\mathbf{S}} \overset{\text{def}}{=} \{x \;:\; \mathbf{S}^\top\mathbf{A}x = \mathbf{S}^\top b\}$, in the norm $\|x\|_{\mathbf{B}} \overset{\text{def}}{=} \sqrt{x^\top\mathbf{B}x}$.

Consider the **stochastic fixed point problem**

$$\text{solve} \quad x = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right]. \tag{5}$$

That is, we seek to find a **fixed point** of the mapping

$$x \to \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right].$$

When solving the problem, we do not have an explicit access to the average projection map. Instead, we are able to repeatedly sample $\mathbf{S} \sim \mathcal{D}$, and use the stochastic projection map $x \to \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)$.

Reformulation 4: Probabilistic Intersection Problem

Note that $\mathcal{L} \subseteq \mathcal{L}_{\mathbf{S}}$ for all $\mathbf{S}$. We would wish to design $\mathcal{D}$ in such a way that a suitably chosen notion of an intersection of the sets $\mathcal{L}_{\mathbf{S}}$ is equal to $\mathcal{L}$. The correct notion is what we call **probabilistic intersection,** denoted $\cap_{\mathbf{S}\sim\mathcal{D}}\mathcal{L}_{\mathbf{S}}$, and defined as the set of points x which belong to $\mathcal{L}_{\mathbf{S}}$ with probability one.

This leads to the problem:

$$\text{find} \quad x \in \cap_{\mathbf{S}\sim\mathcal{D}}\mathcal{L}_{\mathbf{S}} \overset{\text{def}}{=} \{x \ : \ \mathrm{Prob}(x \in \mathcal{L}_{\mathbf{S}}) = 1\}. \tag{6}$$

As before, we typically do not have an explicit access to the probabilistic intersection when designing an algorithm. Instead, we can repeatedly sample $\mathbf{S} \sim \mathcal{D}$, and utilize the knowledge of $\mathcal{L}_{\mathbf{S}}$ to drive the iterative process. If $\mathcal{D}$ is a discrete distribution, probabilistic intersection reduces to standard intersection.

Reformulations: Remarks

- All of the above formulations have a common feature: they all involve an expectation over $\mathbf{S} \sim \mathcal{D}$, and we either do not assume this expectation is known explicitly, or even if it is, we prefer, due to efficiency or other considerations, to sample from unbiased estimates of the objects (e.g., stochastic gradient $\nabla f_{\mathbf{S}}$, stochastic preconditioner $\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}$, stochastic projection map $x \to \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)$, random set $\mathcal{L}_{\mathbf{S}}$) appearing in the formulation.
- As we shall see later, all these stochastic formulations are equivalent. In particular, the following sets are identical: the set of minimizers of the stochastic optimization problem (2), the solution set of the preconditioned system (4), the set of fixed points of the stochastic fixed point problem (5), and the probabilistic intersection (6).
- Further, we give necessary and sufficient conditions for this set to be equal to $\mathcal{L}$. Distributions $\mathcal{D}$ satisfying these conditions always exist, independently of any assumptions on the system beyond consistency. The simplest, but also the least useful choice of a distribution is to pick $\mathbf{S} = \mathbf{I}$ (the $m \times m$ identity matrix), with probability one. In this case, all of our reformulations become trivial.

Three Algorithms

Besides proposing a family of stochastic reformulations of (1), we also propose several stochastic algorithms for solving them:

- **Basic Method:** Algorithm 1
- **Parallel Method:** Algorithm 2
- **Accelerated Method:** Algorithm 3

Each method can be interpreted naturally from the viewpoint of each of the reformulations.

Introduction to Randomized Methods in Convex Optimization
Peter Richtárik

2. The Basic Method

Basic Method

We shall now discuss some of the interpretations of the **basic method,** which performs updates of the form

$$x_{k+1} \overset{\text{def}}{=} \underbrace{x_k - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top(\mathbf{A}x_k - b)}_{\phi_\omega(x_k,\mathbf{S}_k)}, \qquad (7)$$

where $\mathbf{S}_k \sim \mathcal{D}$ is sampled afresh in each iteration, and † denotes the **Moore-Penrose pseudoinverse.**

Algorithm 1 Basic Method

```
Parameters: distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$; stepsize/relaxation parameter $\omega \in \mathbb{R}$
Choose $x_0 \in \mathbb{R}^n$                                  ▷ Initialization
for $k = 0, 1, 2, \ldots$ do
    Draw a fresh sample $\mathbf{S}_k \sim \mathcal{D}$
    Set $x_{k+1} = x_k - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top(\mathbf{A}x_k - b)$
```

Interpretations of the Basic Method

Stochastic Gradient Descent

Algorithm 1 can be seen as **stochastic gradient descent,** with fixed stepsize, applied to (2).

In iteration k of the method, we sample $\mathbf{S}_k \sim \mathcal{D}$, and compute $\nabla f_{\mathbf{S}_k}(x_k)$, which is an unbiased stochastic approximation of $\nabla f(x_k)$. We then perform the step

$$x_{k+1} = x_k - \omega \nabla f_{\mathbf{S}_k}(x_k), \tag{8}$$

where $\omega > 0$ is a stepsize.

Stochastic Newton Method

The method can also be seen as a **stochastic Newton method.**

At iteration k we sample $\mathbf{S}_k \sim \mathcal{D}$, and instead of applying the inverted Hessian of $f_{\mathbf{S}_k}$ to the stochastic gradient (this is not possible as the Hessian is not necessarily invertible), we apply the **B**-pseudoinverse. That is, we perform the step

$$x_{k+1} = x_k - \omega (\nabla^2 f_{\mathbf{S}_k}(x_k))^{\dagger_{\mathbf{B}}} \nabla f_{\mathbf{S}_k}(x_k), \tag{9}$$

where $\omega > 0$ is a stepsize, and the **B**-pseudoinverse of a matrix **M** is defined as $\mathbf{M}^{\dagger_{\mathbf{B}}} \overset{\text{def}}{=} \mathbf{B}^{-1}\mathbf{M}^\top(\mathbf{M}\mathbf{B}^{-1}\mathbf{M}^\top)^\dagger$.

Remark: One may wonder, why are methods (8) and (9) equivalent? Certainly, in general, stochastic gradient descent and stochastic Newton methods are not equivalent. It turns out that the stochastic gradient is always an eigenvector of the **B**-pseudoinverse Hessian, with eigenvalue 1 (see Lemma 11).

Stochastic Proximal Point Method

The method can also be seen as a **stochastic proximal point method.**

At iteration k we sample $\mathbf{S}_k \sim \mathcal{D}$, and perform the step

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f_{\mathbf{S}_k}(x) + \frac{1-\omega}{2\omega} \|x - x_k\|_{\mathbf{B}}^2 \right\}. \tag{10}$$

Remarks:

(i) The *proximal point method* is obtained from (10) by replacing $f_{\mathbf{S}_k}$ with f.

(ii) Unlike in the case of all other methods, here are limited to choose stepsize $0 < \omega < 1$

Stochastic Fixed Point Method

From the perspective of the stochastic fixed point problem (5), Algorithm 1 can be interpreted as a **stochastic fixed point method, with relaxation.**

We first reformulate the problem into an equivalent form using relaxation, which is done to improve the contraction properties of the map. We pick a relaxation parameter $\omega > 0$, and instead consider the equivalent fixed point problem

$$x = \mathrm{E}_{\mathbf{S} \sim \mathcal{D}} \left[\omega \Pi_{\mathcal{L}_{\mathbf{S}}}^{\mathbf{B}}(x) + (1-\omega)x \right].$$

Now, at iteration k, we sample $\mathbf{S}_k \sim \mathcal{D}$, which enables us to obtain an unbiased estimate of the new fixed point mapping, and then simply perform one step of a fixed point method on this mapping:

$$x_{k+1} = \omega \Pi_{\mathcal{L}_{\mathbf{S}_k}}^{\mathbf{B}}(x_k) + (1-\omega)x_k. \tag{11}$$

Stochastic Projection Method

Algorithm 1 can also be seen as a **stochastic projection method** applied to the probabilistic intersection problem (6).

By sampling $\mathbf{S}_k \sim \mathcal{D}$, we are one of the sets defining the intersection, namely $\mathcal{L}_{\mathbf{S}_k}$. We then project the last iterate onto this set, in the **B**-norm, followed by a relaxation step with relaxation parameter $\omega > 0$. That is, we perform the update

$$x_{k+1} = x_k + \omega(\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}_k}}(x_k) - x_k). \tag{12}$$

This is a randomized variant of an alternating projection method. Note that the representation of $\mathcal{L}$ as a probabilistic intersection of sets is not given to us. Rather, we construct it with the hope to obtain faster convergence.

Filling in Some Technical Details

Moore-Penrose Pseudoinverse - I

Let $\mathbf{M} \in \mathbb{R}^{n \times n}$. If $\mathbf{M}$ is invertible, then there exists a matrix, denoted by $\mathbf{M}^{-1} \in \mathbb{R}^{n \times n}$, called the **inverse matrix,** with the properties:

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}, \qquad \mathbf{M}^{-1}\mathbf{M} = \mathbf{I}.$$

Not every square matrix has an inverse.

There is a generalization of the concept of the inverse, called **(Moore-Penrose) pseudoinverse.** The nice thing about it is that every matrix, even rectangular matrices, have a unique pseudoinverse.

Exercise 1

Use one of the properties of the pseudoinverse listed on the next slide to show that the pseudoinverse of a real number $\alpha \in \mathbb{R}$ is given by:

$$\alpha^{\dagger} = \begin{cases} \frac{1}{\alpha}, & \textit{if} \quad \alpha \neq 0, \\ 0, & \textit{if} \quad \alpha = 0. \end{cases} \tag{13}$$

Moore-Penrose Pseudoinverse - II

Fact 2

Every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has a unique pseudoinverse $\mathbf{A}^{\dagger} \in \mathbb{R}^{n \times m}$. Among others, this matrix satisfies the following properties:

(i) $\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}$

(ii) $\mathbf{A}^{\top} = \mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\top}$

(iii) $\mathbf{A}^{\top} = \mathbf{A}^{\top}\mathbf{A}\mathbf{A}^{\dagger}$

(iv) $\mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}$

(v) $(\mathbf{A}^{\dagger})^{\top} = (\mathbf{A}^{\top})^{\dagger}$

Exercise 2

Use the above fact to show that: i) the pseudoinverse of a symmetric matrix is symmetric, ii) the pseudoinverse of a positive semidefinite matrix is positive semidefinite, iii) if $\mathbf{A}$ is invertible, then $\mathbf{A}^{\dagger} = \mathbf{A}^{-1}$.

Assumption on $\mathcal{D}$

Without the following assumption, the reformulations would not make sense (i.e., the expectations would not be defined/finite):

Assumption 2 (Finite mean)

The random matrix

$$\mathbf{H} = \mathbf{H}_{\mathbf{S}} \overset{def}{=} \mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top \tag{14}$$

has a mean. That is, the following matrix has finite entries:

$$\mathrm{E}[\mathbf{H}] = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}[\mathbf{H}] = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top\right]$$

Remark:

(i) $\mathbf{H} = \mathbf{H}_{\mathbf{S}}$ is a random matrix because it depends on the random matrix $\mathbf{S}$. However, in order to simplify notation, we will drop the subscript highlighting this dependency and will simply write $\mathbf{H}$.

(ii) By $\mathrm{E}[\mathbf{H}]$ we simply mean the matrix whose (i,j) entry is the mean of the (i,j) entry of $\mathbf{H}$:

$$(\mathrm{E}[\mathbf{H}])_{ij} = \mathrm{E}[\mathbf{H}_{ij}].$$

Assumption on $\mathcal{D}$: Exercises

Exercise 3

(i) *Show that the matrix* $\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}$ *is symmetric and positive semidefinite.*

(ii) *It is known (see Exercise 2) that the pseudoinverse of a symmetric and positive semidefinite matrix is again symmetric and positive semidefinite. Show that* $\mathbf{H}$ *is symmetric and positive semidefinite.*

(iii) *Show that* $\mathrm{E}[\mathbf{H}]$ *is symmetric and positive semidefinite.*

Assumption on $\mathcal{D}$: Examples

Let $e_1, e_2, \ldots, e_m$ be standard basis vectors (aka coordinate vectors) in $\mathbb{R}^m$. That is, e_i is the vector whose all entries are zeros, except for the ith entry, which is equal to 1.

Example 3 (Uniform sampling unit of basis vectors)

Let $\mathcal{D}$ be the uniform distribution over $\{e_i\}$. That is, for all $i = 1, 2, \ldots, m$ we let

$$\mathbf{S} = e_i \quad \text{with probability} \quad 1/m.$$

We can then compute:

$$\mathrm{E}\left[\mathbf{H}\right] = \sum_{i=1}^m \tfrac{1}{m} e_i (\mathbf{A}_{i:}\mathbf{B}^{-1}\mathbf{A}_{i:}^\top)^\dagger e_i^\top = \tfrac{1}{m}\mathrm{Diag}\left(\alpha_1, \ldots, \alpha_m\right),$$

where

$$\alpha_i \overset{\text{def}}{=} (\mathbf{A}_{i:}\mathbf{B}^{-1}\mathbf{A}_{i:}^\top)^\dagger \overset{(13)}{=} 1/\|\mathbf{A}_{i:}^\top\|^2_{\mathbf{B}^{-1}}, \qquad i = 1, 2, \ldots, m,$$

and $\mathrm{Diag}\left(\alpha\right)$ is the diagonal matrix with vector α on the diagonal.

Note that if $\mathbf{A}$ has nonzero rows, then $\mathrm{E}\left[\mathbf{H}\right] \succ 0$.

Is f well defined?

We may wonder: does the expectation in (2) exist? That is, is f well defined? The next result says that all is fine.

Lemma 4

Let x_* be any solution of the linear system $\mathbf{A}x = b$ (that is, let $x_* \in \mathcal{L}$). Then

$$f_{\mathbf{S}}(x) = \frac{1}{2}(x - x_*)^\top \mathbf{A}^\top \mathbf{H}\mathbf{A}(x - x_*). \tag{15}$$

Moreover,

$$f(x) = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[f_{\mathbf{S}}(x)\right] = \frac{1}{2}(x - x_*)^\top \mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{H}\right]\mathbf{A}(x - x_*), \tag{16}$$

and hence $f(x)$ is finite for all $x \in \mathbb{R}^n$. Thus, f is well defined.

Proof of Lemma 4

Step 1: $x_* \in \mathcal{L}$ implies $\mathbf{A}x_* = b$. Plugging this into (3) gives (15).

Step 2: It remains to establish (16). We will use two facts:

Fact 5

For any $\mathbf{X} \in \mathbb{R}^{n\times n}$ and $h \in \mathbb{R}^n$, we have[1] $h^\top \mathbf{X} h = \mathrm{Trace}\left(\mathbf{X}hh^\top\right)$.

Fact 6

Fix any $\mathbf{M} \in \mathbb{R}^{n\times n}$. The map $\mathbf{X} \mapsto \mathrm{Trace}\left(\mathbf{X}\mathbf{M}\right)$ is linear.

Now back to the proof. Let $h = \mathbf{A}(x - x_*)$. Utilizing the above two facts, we get

$$\begin{aligned} f(x) &\overset{(2)}{=} \mathrm{E}\left[f_{\mathbf{S}}(x)\right] \overset{(15)}{=} \tfrac{1}{2}\mathrm{E}\left[h^\top \mathbf{H} h\right] \overset{(\text{Fact 16})}{=} \tfrac{1}{2}\mathrm{E}\left[\mathrm{Trace}\left(\mathbf{H}hh^\top\right)\right] \\ &\overset{(\text{Fact 17})}{=} \tfrac{1}{2}\mathrm{Trace}\left(\mathrm{E}\left[\mathbf{H}\right] hh^\top\right) \overset{(\text{Fact 16})}{=} \tfrac{1}{2}h^\top \mathrm{E}\left[\mathbf{H}\right] h, \end{aligned}$$

which gives (16). Note that when applying Fact 17, we have also used linearity of expectation.

[1] Recall that **trace** of a matrix, denoted $\mathrm{Trace}\left(\cdot\right)$, is the sum of its diagonal elements.

Introduction to Randomized Methods in Convex Optimization
Peter Richtárik

3. Equivalence and Exactness

Projection and Pseudoinverse

Projection Operators and Pseudoinverse Matrices - I

Definition 7

The **B-pseudoinverse** of a matrix $\mathbf{M}$, is defined as

$$\mathbf{M}^{\dagger_{\mathbf{B}}} \overset{\text{def}}{=} \mathbf{B}^{-1}\mathbf{M}^{\top}(\mathbf{M}\mathbf{B}^{-1}\mathbf{M}^{\top})^{\dagger}, \tag{17}$$

where $\dagger$ denotes the standard pseudoinverse.

Exercise 4

Show that

(i) $\mathbf{A}^{\dagger}\mathbf{A}$ *is a symmetric matrix*

(ii) $\mathbf{A}^{\top}(\mathbf{A}\mathbf{A}^{\top})^{\dagger} = \mathbf{A}^{\dagger}$

(iii) *The* $\mathbf{I}$*-pseudoinverse is the standard Moore-Penrose pseudoinverse.*

Projection Operators and Pseudoinverse Matrices - II

Lemma 8

The projection onto $\mathcal{L} = \{x \,:\, \mathbf{A}x = b\}$ is given by

$$\Pi_{\mathcal{L}}^{\mathbf{B}}(x) = x - \mathbf{B}^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top)^\dagger(\mathbf{A}x - b) \overset{(17)}{=} x - \mathbf{A}^{\dagger_\mathbf{B}}(\mathbf{A}x - b). \quad (18)$$

Proof.

Do it yourself. □

Exercise 5

Show that **B***-pseudoinverse satisfies*

$$\mathbf{A}^{\dagger_\mathbf{B}} b = \Pi_{\mathcal{L}}^{\mathbf{B}}(0) = \arg\min_x \{\|x\|_\mathbf{B} \,:\, \mathbf{A}x = b\}.$$

Equivalence of Algorithms

Gradient and Hessian of $f_{\mathbf{S}}(x)$ - I

In order to keep the expressions as brief as possible throughout, it will be useful to define

$$\mathbf{Z} \overset{\text{def}}{=} \mathbf{A}^\top \mathbf{H} \mathbf{A} \overset{(14)}{=} \mathbf{A}^\top \mathbf{S}(\mathbf{S}^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top \mathbf{A}. \qquad (19)$$

Lemma 9

$\mathbf{B}^{-1}\mathbf{Z}$ *is the projection, in the* $\mathbf{B}$*-norm, onto* $\mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}\right)$*. In particular,*

$$(\mathbf{B}^{-1}\mathbf{Z})^2 = \mathbf{B}^{-1}\mathbf{Z} \qquad \textit{and} \qquad \mathbf{Z}\mathbf{B}^{-1}\mathbf{Z} = \mathbf{Z}. \qquad (20)$$

Recall from (3) that $f_{\mathbf{S}}(x) \overset{\text{def}}{=} \frac{1}{2}\|\mathbf{A}x - b\|_{\mathbf{H}}^2 = \frac{1}{2}(\mathbf{A}x - b)^\top \mathbf{H}(\mathbf{A}x - b)$. By combining this with (19), this can be also written in the compact form

$$f_{\mathbf{S}}(x) = \frac{1}{2}(x - x_*)^\top \mathbf{Z}(x - x_*), \qquad (21)$$

where x_* is any point in $\mathcal{L}$.

Gradient and Hessian of $f_{\mathbf{S}}(x)$ - II

Lemma 10

For each $x, h \in \mathbb{R}^n$ *we have the expansion*

$$f_{\mathbf{S}}(x+h) = f_{\mathbf{S}}(x) + \langle \nabla f_{\mathbf{S}}(x), h\rangle_{\mathbf{B}} + \tfrac{1}{2}\left\langle (\nabla^2 f_{\mathbf{S}})h, h\right\rangle_{\mathbf{B}},$$

where

$$\nabla f_{\mathbf{S}}(x) \overset{\text{def}}{=} \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}(\mathbf{A}x - b) \qquad \textit{and} \qquad \nabla^2 f_{\mathbf{S}} \overset{\text{def}}{=} \mathbf{B}^{-1}\mathbf{Z} \qquad (22)$$

are the gradient and Hessian of $f_{\mathbf{S}}$ *with respect to the* $\mathbf{B}$*-inner product, respectively.*[2]

In view of (19) and (22), the gradient can also be written as

$$\nabla f_{\mathbf{S}}(x) = \mathbf{B}^{-1}\mathbf{Z}(x - x_*), \qquad x \in \mathbb{R}^n,\ x_* \in \mathcal{L}. \qquad (23)$$

[2] If $\mathbf{B} = \mathbf{I}$, then $\langle\cdot,\cdot\rangle_{\mathbf{B}}$ is the standard Euclidean inner product, and we recover formulas for the standard gradient and Hessian. Note that $\mathbf{B}^{-1}\mathbf{Z}$ is both self-adjoint and positive semidefinite with respect to the $\mathbf{B}$-inner product. Indeed, for all $x, y \in \mathbb{R}^n$ we have $\left\langle \mathbf{B}^{-1}\mathbf{Z}x, y\right\rangle_{\mathbf{B}} = \langle \mathbf{Z}x, y\rangle_{\mathbf{I}} = \langle x, \mathbf{Z}y\rangle_{\mathbf{I}} = \left\langle x, \mathbf{B}^{-1}\mathbf{Z}y\right\rangle_{\mathbf{B}}$, and $\left\langle \mathbf{B}^{-1}\mathbf{Z}x, x\right\rangle_{\mathbf{B}} = \langle \mathbf{Z}x, x\rangle_{\mathbf{I}} \geq 0$.

Useful Identities Involving $f_{\mathbf{S}}(x)$

Lemma 11

For all $x \in \mathbb{R}^n$, we have

$$\begin{aligned} \nabla f_{\mathbf{S}}(x) &= (\nabla^2 f_{\mathbf{S}})\nabla f_{\mathbf{S}}(x) = (\nabla^2 f_{\mathbf{S}})^{\dagger_{\mathbf{B}}}\nabla f_{\mathbf{S}}(x) \\ &= x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) = \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}(\mathbf{A}x - b). \end{aligned} \tag{24}$$

Moreover,

$$f_{\mathbf{S}}(x) = \frac{1}{2}\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}}. \tag{25}$$

Finally, if $\mathcal{L}_{\mathbf{S}}$ is the set of minimizers of $f_{\mathbf{S}}$, then $\mathcal{L} \subseteq \mathcal{L}_{\mathbf{S}}$, and

(i) $\mathcal{L}_{\mathbf{S}} = \{x \ : \ f_{\mathbf{S}}(x) = 0\} = \{x \ : \ \nabla f_{\mathbf{S}}(x) = 0\}$

(ii) $\mathcal{L}_{\mathbf{S}} = x_* + \mathrm{Null}\left(\mathbf{B}^{-1}\mathbf{Z}\right)$ *for all $x_* \in \mathcal{L}$*

(iii) $\mathcal{L}_{\mathbf{S}} = \{x \ : \ \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}b\}$ *(see (4))*

(iv) $\mathcal{L}_{\mathbf{S}} = \{x \ : \ \mathbf{S}^\top\mathbf{A}x = \mathbf{S}^\top b\}$ *(see (6))*

Some Consequences of Lemma 11

- The identity $(\nabla^2 f_{\mathbf{S}})\nabla f_{\mathbf{S}}(x) = \nabla f_{\mathbf{S}}(x)$ means that the stochastic gradients of $f_{\mathbf{S}}$ are eigenvectors of the stochastic Hessian $\nabla^2 f_{\mathbf{S}}$, corresponding to eigenvalue one.
- The identity $(\nabla^2 f_{\mathbf{S}})^{\dagger \mathbf{B}}\nabla f_{\mathbf{S}}(x) = \nabla f_{\mathbf{S}}(x)$ means that the stochastic gradients of $f_{\mathbf{S}}$ are eigenvectors of the **B**-pseudoinverse of the stochastic Hessian $\nabla^2 f_{\mathbf{S}}$, corresponding to eigenvalue one.
- Function f can be represented in multiple ways:

$$f(x) = \frac{1}{2}\mathrm{E}\left[\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\|^2_{\mathbf{B}}\right] = \frac{1}{2}\mathrm{E}\left[\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}}\right]. \tag{26}$$

- The gradient and Hessian of f (with respect to the **B**-inner product) are given by

$$\nabla f(x) = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*), \quad \text{and} \quad \nabla^2 f = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right], \tag{27}$$

respectively, where x_* is any point in $\mathcal{L}$.

Equivalence of Algorithms

Theorem 12

Algorithm 1 (Basic Method) can be equivalently written as stochastic gradient descent (8)*, stochastic Newton method* (9)*, stochastic fixed point method* (11)*, and stochastic projection method* (12)*.*

Proof.

This follows from identities (24) in Lemma 11. □

Proof of Lemma 11 - I

Pick any $x_* \in \mathcal{L}$. First, we have

$$\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) \overset{(18)}{=} x - \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}(\mathbf{A}x - b) \overset{(22)}{=} x - \nabla f_{\mathbf{S}}(x).$$

To establish (24), it now only remains to consider the two expressions involving the Hessian. We have

$$\nabla^2 f_{\mathbf{S}} \nabla f_{\mathbf{S}}(x) \overset{(22)+(23)}{=} \mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}(x - x_*) \overset{(20)}{=} \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \overset{(23)}{=} \nabla f_{\mathbf{S}}(x),$$

and

$$\begin{aligned}
(\nabla^2 f_{\mathbf{S}})^{\dagger_{\mathbf{B}}} \nabla f_{\mathbf{S}}(x) &\overset{(17)}{=} \mathbf{B}^{-1}(\nabla^2 f_{\mathbf{S}})^\top \left((\nabla^2 f_{\mathbf{S}})\mathbf{B}^{-1}(\nabla^2 f_{\mathbf{S}})^\top\right)^\dagger \nabla f_{\mathbf{S}}(x) \\
&\overset{(22)}{=} \mathbf{B}^{-1}(\mathbf{B}^{-1}\mathbf{Z})^\top \left((\mathbf{B}^{-1}\mathbf{Z})\mathbf{B}^{-1}(\mathbf{B}^{-1}\mathbf{Z})^\top\right)^\dagger \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&= \mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)^\dagger \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&\overset{(20)}{=} \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)\left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)^\dagger \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)\mathbf{B}(x - x_*) \\
&= \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&\overset{(23)}{=} \nabla f_{\mathbf{S}}(x).
\end{aligned}$$

Proof of Lemma 11 - II

Identity (25) follows from

$$\frac{1}{2}\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}} \overset{(23)}{=} \frac{1}{2}(x-x_*)^\top \mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}(x-x_*) \overset{(20)}{=} \frac{1}{2}(x-x_*)^\top \mathbf{Z}(x-x_*) \overset{(21)}{=} f_{\mathbf{S}}(x).$$

If $x \in \mathcal{L}$, then by picking $x_* = x$ in (23), we see that $x \in \mathcal{L}_{\mathbf{S}}$.

It remains to show that the sets defined in (i)–(iv) are identical.

- Equivalence between (i) and (ii) follows from (23).
- Now consider (ii) and (iii). Any $x_* \in \mathcal{L}$ belongs to the set defined in (iii), which follows immediately by substituting $b = \mathbf{A}x_*$. The rest follows after observing the nullspaces are identical.
- In order to show that (iii) and (iv) are equivalent, it suffices to compute $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)$ and observe that $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) = x$ if and only if x belongs to the set defined in (iii).

Equivalence of 4 Stochastic Reformulations

Equivalence of the Stochastic Formulations

The below theorem says that the solution sets of the fours stochastic problems (2), (4), (5), and (6) are identical. **In this sense, the four stochastic problems are equivalent.**

Theorem 13 (Equivalence of stochastic formulations)

Let $x_* \in \mathcal{L}$. The following sets are identical:

(i) $\mathcal{X} = \arg\min f(x) = \{x \ : \ f(x) = 0\} = \{x \ : \ \nabla f(x) = 0\}$ → (2)

(ii) $\mathcal{X} = \{x \ : \ \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right]\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right] b\} = x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right)$ → (4)

(iii) $\mathcal{X} = \{x \ : \ \mathrm{E}\left[\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right] = x\}$ → (5)

(iv) $\mathcal{X} = \{x \ : \ \mathrm{Prob}(x \in \mathcal{L}_{\mathbf{S}}) = 1\}$ → (6)

Moreover, $\mathcal{X}$ does not depend on $\mathbf{B}$.

Proof of Theorem 13 - Part I

As f is convex, nonnegative and achieving the value of zero (since $\mathcal{L} \neq \emptyset$), the sets in (i) are all identical. We shall now show that the sets defined in (ii)–(iv) are equal to that defined in (i).

(i) ↔ *(ii)*: Using the formula for the gradient from (27), we see that

$$\begin{aligned}\{x \ : \ \nabla f(x) = 0\} &= \{x \ : \ \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*) = 0\} \\ &= \{x \ : \ \mathrm{E}\left[\mathbf{Z}\right](x - x_*) = 0\} \\ &= x_* + \{h \ : \ \mathrm{E}\left[\mathbf{Z}\right] h = 0\} \\ &= x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right),\end{aligned}$$

which shows that (i) and (ii) are the same.

(i) ↔ *(iii)*: Equivalence of (i) and (iii) follows by taking expectations in (24) to obtain

$$\nabla f(x) = \mathrm{E}\left[\nabla f_{\mathbf{S}}(x)\right] \overset{(24)}{=} \mathrm{E}\left[x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right].$$

Proof of Theorem 13 - Part II

$(i) \leftrightarrow (iv)$**:** It remains to establish equivalence between (i) and (iv). Let

$$\mathcal{X} = \{x \;:\; f(x) = 0\} \overset{(26)}{=} \left\{x \;:\; \mathrm{E}\left[\left\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right\|^2_{\mathbf{B}}\right] = 0\right\} \qquad (28)$$

and let $\mathcal{X}'$ be the set from (iv).

We need to show that $\mathcal{X}' = \mathcal{X}$. For easier reference, let

$$\xi_{\mathbf{S}}(x) \overset{\text{def}}{=} \left\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right\|^2_{\mathbf{B}}.$$

Note that the following three probabilistic events are identical:

$$[x \in \mathcal{L}_{\mathbf{S}}] = \left[x = \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right] = [\xi_{\mathbf{S}}(x) = 0]. \qquad (29)$$

We first show that $\mathcal{X}' \subseteq \mathcal{X}$.

In view of (29), if $x \in \mathcal{X}'$, then the random variable $\xi_{\mathbf{S}}(x)$ is equal to zero with probability 1, which implies $\mathrm{E}\left[\xi_{\mathbf{S}}(x)\right] = 0$, whence $x \in \mathcal{X}$.

Proof of Theorem 13 - Part III

Let us now sow that $\mathcal{X} \subseteq \mathcal{X}'$.

Let $1_{[\xi_{\mathbf{S}}(x) \geq t]}$ be the indicator function of the event $[\xi_{\mathbf{S}}(x) \geq t]$. Note that since $\xi_{\mathbf{S}}(x)$ is a nonnegative random variable, for all $t \in \mathbb{R}$ we have the inequality

$$\xi_{\mathbf{S}}(x) \geq t 1_{\xi_{\mathbf{S}}(x) \geq t}. \qquad (30)$$

Now take $x \in \mathcal{X}$ and consider $t > 0$. By taking expectations in (30), we obtain

$$0 = \mathrm{E}\left[\xi_{\mathbf{S}}(x)\right] \geq \mathrm{E}\left[t 1_{\xi_{\mathbf{S}}(x) \geq t}\right] = t\mathrm{E}\left[1_{\xi_{\mathbf{S}}(x) \geq t}\right] = t\mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t),$$

which implies that $\mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t) = 0$. Now choose $t_i = 1/i$ for $i = 1, 2, \ldots$ and note that the event $[\xi_{\mathbf{S}}(x) > 0]$ can be written as

$$[\xi_{\mathbf{S}}(x) > 0] = \bigcup_{i=1}^{\infty} [\xi_{\mathbf{S}}(x) \geq t_i].$$

Proof of Theorem 13 - Part IV

Therefore, by the union bound,

$$\mathrm{Prob}(\xi_{\mathbf{S}}(x) > 0) \leq \sum_{i=1}^{\infty} \mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t_i) = 0,$$

which immediately implies that $\mathrm{Prob}(\xi_{\mathbf{S}}(x) = 0) = 1$. From (29) we conclude that $x \in \mathcal{X}'$.

Independence on B. Since characterization (iv) of $\mathcal{X}$ does not depend on **B**, we conclude that $\mathcal{X}$ does not depend on **B**.

Exactness of the Reformulations

Rangespace and Nullspace of a Matrix - I

Let $\mathbf{M} \in \mathbb{R}^{m\times n}$.

Definition 14 (Rangespace of a matrix)

By $\mathrm{Range}\left(\mathbf{M}\right)$ we mean the **rangespace of matrix M.** This is the linear subspace of $\mathbb{R}^m$ generated by the columns of $\mathbf{M}$:

$$\mathrm{Range}\left(\mathbf{M}\right) \overset{\text{def}}{=} \{\mathbf{M}x \ : \ x \in \mathbb{R}^n\} = \left\{\textstyle\sum_j \mathbf{M}_{:j} x_j, \quad x \in \mathbb{R}^n\right\}.$$

Definition 15 (Nullspace of a matrix)

By $\mathrm{Null}\left(\mathbf{M}\right)$ we mean the **nullspace of matrix M.** This is the linear subspace of $\mathbb{R}^n$ formed by the vectors orthogonal (under standard Euclidean inner product) to all rows of $\mathbf{M}$:

$$\mathrm{Null}\left(\mathbf{M}\right) \overset{\text{def}}{=} \{x \in \mathbb{R}^n \ : \ \mathbf{M}x = 0\} = \left\{x \in \mathbb{R}^n \ : \ \langle \mathbf{M}_{i:}^\top, x\rangle = 0 \quad \forall i\right\}.$$

Rangespace and Nullspace of a Matrix - II

Definition 16 (Orthogonal complement)

Let X be a subspace of a vector space. The **orthogonal complement** of X is the linear subspace $X^\perp \overset{\text{def}}{=} \{y \ : \ \langle y, x\rangle = 0 \ \forall x \in X\}$.

Here we collect some useful identities involving rangespaces and nullspaces of a matrix:

Fact 17

For any $\mathbf{M} \in \mathbb{R}^{m\times n}$, *we have*

(i) $\mathrm{Range}\left(\mathbf{M}^\top\right) = \mathrm{Null}\left(\mathbf{M}\right)^\perp$

(ii) $\mathrm{Range}\left(\mathbf{M}^\top\right)^\perp = \mathrm{Null}\left(\mathbf{M}\right)$

(iii) *If* $\mathbf{G} \succ 0$, *then* $\mathrm{Null}\left(\mathbf{M}^\top \mathbf{G}\mathbf{M}\right) = \mathrm{Null}\left(\mathbf{M}\right)$

Exactness

Key Question: When are the stochastic formulations (2), (4), (5), (6) equivalent to the linear system (1)? That is, when is their set of solutions $\mathcal{X}$ identical to solution set of the linear system $\mathcal{L}$?

This leads to the concept of **exactness:**

Assumption 3 (Exactness)

Stochastic reformulations (2), (4), (5), (6) *of problem* (1) *are* exact. *That is,* $\mathcal{X} = \mathcal{L}$*.*

In what follows, we will

- Give **sufficient,** and **necessary & sufficient** conditions for exactness.
- Use this assumption to **prove convergence** of the algorithms to a specific point in $\mathcal{L}$.

Necessary and Sufficient Conditions for Exactness

Theorem 18 ($\Leftrightarrow$ Conditions for exactness)

The following statements are equivalent:

(i) *Assumption 3 (exactness) holds*

(ii) $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{A}\right)$

(iii) $\mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\right) = \mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right)$

(iv) $\mathrm{Range}\left(\mathbf{A}\right) \cap \mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right) = \{0\}$

Proof of Theorem 18 - I

(i) ↔ *(ii)*: Choose any $x_* \in \mathcal{L}$. We know that $\mathcal{L} = x_* + \mathrm{Null}\left(\mathbf{A}\right)$. On the other hand, Theorem 13 says that $\mathcal{X} = x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right)$.

(ii) ↔ *(iii)*: If (ii) holds, then

$$\mathrm{Null}\left(\mathbf{A}\right) = \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\right),$$

and (iii) follows. If (iii) holds, then

$$\mathrm{Null}\left(\mathbf{A}\right) = \mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right),$$

proving (ii).

Proof of Theorem 18 - II

(ii) ↔ *(iv)*: First, note that

$$\mathrm{E}\left[\mathbf{Z}\right] = \mathbf{A}^\top (\mathrm{E}\left[\mathbf{H}\right])^{1/2} (\mathrm{E}\left[\mathbf{H}\right])^{1/2} \mathbf{A}.$$

In view of Fact 17, for any matrix $\mathbf{M}$ we have $\mathrm{Null}\left(\mathbf{M}^\top \mathbf{M}\right) = \mathrm{Null}\left(\mathbf{M}\right)$. Therefore,

$$\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\mathbf{A}\right).$$

Moreover, we know that

(a) $\mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\mathbf{A}\right) = \mathrm{Null}\left(\mathbf{A}\right)$ if and only if $\mathrm{Range}\left(\mathbf{A}\right) \cap \mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\right) = \{0\}$, and

(b) $\mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\right) = \mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right)$ (see Fact 17).

It remains to combine these observations.

Sufficient Conditions for Exactness

We now list some sufficient conditions for exactness.

Lemma 19 (Sufficient conditions for exactness)

Any of these conditions implies that Assumption 3 is satisfied:

(i) $\mathrm{E}\left[\mathbf{H}\right] \succ 0$

(ii) $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right) \subseteq \mathrm{Null}\left(\mathbf{A}^\top\right)$

Proof.

If (i) holds, then $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right]\mathbf{A}\right) = \mathrm{Null}\left(\mathbf{A}\right)$, where the last equality follows from Fact 17. Exactness now follows by applying Theorem 18.

On the other hand, in view of Fact 17, (ii) implies statement (iv) in Theorem 18, and hence exactness follows. □

Condition Number

Spectral Decomposition of the Hessian of f

Recall that the **Hessian of f** is given by

$$\nabla^2 f = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\nabla^2 f_{\mathbf{S}}\right] = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]. \tag{31}$$

Lemma 20

Matrices $\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]$ and $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$ have the same eigenvalues.

Proof.

It is known that for any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$, the matrices $\mathbf{XY}$ and $\mathbf{YX}$ have the same eigenvalues. It only remains to apply this to $\mathbf{X} = \mathbf{B}^{-1/2}$ and $\mathbf{Y} = \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]$. □

The above result allows us to study spectral properties of the Hessian $\nabla^2 f$ through the **eigenvalue decomposition** of the symmetric positive definite matrix $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$.

Eigenvalues of the Hessian of f

Let

$$\mathbf{W} \overset{\text{def}}{=} \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \tag{32}$$

be the **eigenvalue decomposition of W,** where

$$\mathbf{U} = [u_1, \ldots, u_n] \in \mathbb{R}^{n\times n}$$

is an orthonormal matrix composed of **eigenvectors** (i.e., we have $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$), and

$$\Lambda = \mathrm{Diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right)$$

is a diagonal matrix of **eigenvalues.** Assume without loss of generality that the eigenvalues are ordered from largest to smallest:

$$\lambda_{\max} \overset{\text{def}}{=} \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \overset{\text{def}}{=} \lambda_{\min}.$$

All Eigenvalues of W are Between 0 and 1

Lemma 21

$0 \leq \lambda_i \leq 1$ *for all* i*.*

Proof.

Since $\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}$ is symmetric positive semidefinite, so is its expectation $\mathbf{W}$, implying that $\lambda_i \geq 0$ for all i.

Further, note that $\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}$ is a projection matrix. Indeed, it is the projection (in the standard $\mathbf{I}$-norm) onto $\mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{S}\right)$. Therefore, its eigenvalues are all zeros or ones. Since the map $\mathbf{X} \mapsto \lambda_{\max}(\mathbf{X})$ is convex, by **Jensen's inequality** we get

$$\lambda_{\max}(\mathbf{W}) = \lambda_{\max}\left(\mathrm{E}\left[\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}\right]\right) \leq \mathrm{E}\left[\lambda_{\max}(\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2})\right] \leq 1.$$

□

Smallest Nonzero Eigenvalue

Lemma 22

If Assumption 3 (exactness) holds, then $\lambda_{\max} > 0$*.*

Proof.

Assume, by contradiction, that $\lambda_i = 0$ for all i. Then from Theorem 18 and the fact that $\mathrm{Null}\,(\mathbf{W}) = \mathrm{Range}\,(u_i \;:\; \lambda_i = 0)$ we conclude that $\mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right) = \mathbb{R}^n$, which in turn implies that $\mathrm{Null}\,(\mathbf{A}) = \mathbb{R}^n$. This can only happen if $\mathbf{A} = 0$, which is contradicts with our assumption on $\mathbf{A}$. □

Now, let j be the largest index for which $\lambda_j > 0$. This identifies the **smallest nonzero eigenvalue of W,** which we shall denote as

$$\lambda_{\min}^+ = \lambda_j.$$

If all eigenvalues $\{\lambda_i\}$ are positive, then $j = n$.

Condition Number

Definition 23

The **condition number** associated with the four stochastic reformulations is the quantity[3]

$$\zeta(\mathbf{A},\mathbf{B},\mathcal{D}) = \zeta \overset{\text{def}}{=} \|\mathbf{W}\|\|\mathbf{W}^\dagger\| = \frac{\lambda_{\max}}{\lambda_{\min}^+}. \tag{33}$$

Remark:

- As we shall see, convergence rate of the Basic Method is described by ζ.
- As one varies the parameters defining the reformulation (i.e., $\mathcal{D}$ and $\mathbf{B}$), ζ changes. As a general rule of thumb, simple distributions will lead to reformulations with a small condition number. For instance, choosing $\mathbf{S} = \mathbf{I}$ with probability one gives $\zeta = 1$. However, in such a case each step of the Basic Method is very expensive. One needs to strike the right balance.

[3] $\|\mathbf{X}\|$ denotes the **spectral norm** of $\mathbf{X}$. In general, $\|\mathbf{X}\| = \left(\lambda_{\max}(\mathbf{X}^\top\mathbf{X})\right)^{1/2}$. If $\mathbf{X}$ is symmetric positive semidefinite, then $\|\mathbf{X}\|^2 = \lambda_{\max}(\mathbf{X}^\top\mathbf{X}) = \lambda_{\max}(\mathbf{X}^2) = (\lambda_{\max}(\mathbf{X}))^2$. Therefore, $\|\mathbf{X}\| = \lambda_{\max}(\mathbf{X})$.

Introduction to Randomized Methods in Convex Optimization
Peter Richtárik

4. Convergence Analysis of the Basic Method

Covariance Matrix and Total Variance of a Random Vector

Definition 24 (Covariance matrix)

If $x \in \mathbb{R}^n$ is a random vector, then the matrix

$$\mathsf{Var}\,(x) \overset{\text{def}}{=} \mathrm{E}\left[(x - \mathrm{E}\,[x])(x - \mathrm{E}\,[x])^\top\right]$$

is called the **covariance matrix** of x.

Definition 25 (Total Variance)

If $x \in \mathbb{R}^n$ is a random vector, then the value

$$\mathsf{TVar}\,(x) \overset{\text{def}}{=} \mathrm{E}\left[(x - \mathrm{E}\,[x])^\top(x - \mathrm{E}\,[x])\right] = \mathrm{E}\left[\|x - \mathrm{E}\,[x]\,\|^2\right]$$

is called the **total variance** of x.

Exercise 6

Let $x \in \mathbb{R}^n$ be a random vector. Show that:

(i) *The total variance is the trace of the covariance matrix:* $\mathsf{TVar}(x) = \mathsf{Tr}\,(\mathsf{Var}\,(x))$

(ii) $\mathsf{TVar}\,(\mathbf{U}^\top\mathbf{B}^{1/2}x) = \mathrm{E}\left[\|x - \mathrm{E}\,[x]\,\|_{\mathbf{B}}^2\right].$

Strong vs Weak Convergence

Definition 26 (Strong and Weak Convergence)

We say that a sequence of random vectors $\{x_k\}$ converges to x_*

- **weakly** if $\|\mathrm{E}\,[x_k - x_*]\,\|_{\mathbf{B}}^2 \to 0$ as $k \to \infty$
- **strongly** if $\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \to 0$ as $k \to \infty$ (aka $L2$ **convergence**)

The following lemma explains why **strong convergence** is a stronger convergence concept than **weak convergence.**

Lemma 27

For any random vector $x_k \in \mathbb{R}^n$ and any $x_* \in \mathbb{R}^n$ we have the identity

$$\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] = \|\mathrm{E}\,[x_k - x_*]\|_{\mathbf{B}}^2 + \underbrace{\mathrm{E}\left[\|x_k - \mathrm{E}\,[x_k]\|_{\mathbf{B}}^2\right]}_{\mathsf{TVar}\,(\mathbf{U}^\top\mathbf{B}^{1/2}x_k)}.$$

As a consequence, **strong convergence implies**

- weak convergence,
- convergence of $\mathsf{TVar}\,(\mathbf{U}^\top\mathbf{B}^{1/2}x_k)$ to zero.

Proof of Lemma 27

Let $\mu = \mathrm{E}\left[x_k\right]$. Then

$$\begin{aligned}
\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] &= \mathrm{E}\left[\|x_k - \mu + \mu - x_*\|_{\mathbf{B}}^2\right] \\
&= \mathrm{E}\left[\|x_k - \mu\|_{\mathbf{B}}^2 + \|\mu - x_*\|_{\mathbf{B}}^2 + 2\langle x_k - \mu, \mu - x_*\rangle_{\mathbf{B}}\right] \\
&= \mathrm{E}\left[\|x_k - \mu\|_{\mathbf{B}}^2\right] + \|\mu - x_*\|_{\mathbf{B}}^2 + 2\langle \underbrace{\mathrm{E}\left[x_k - \mu\right]}_{0}, \mu - x_*\rangle_{\mathbf{B}} \\
&= \mathrm{E}\left[\|x_k - \mu\|_{\mathbf{B}}^2\right] + \|\mu - x_*\|_{\mathbf{B}}^2.
\end{aligned}$$

In the first step we have expanded the square and in the second step we have used linearity of expectation.

Weak Convergence

Weak Convergence

Theorem 28 (Weak Convergence 1)

Choose any $x_0 \in \mathbb{R}^n$ and let $\{x_k\}$ be the random iterates produced by Algorithm 1. Let $x_* \in \mathcal{L}$ be chosen arbitrarily. Then

$$\mathrm{E}\left[x_{k+1} - x_*\right] = \left(\mathbf{I} - \omega \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]\right) \mathrm{E}\left[x_k - x_*\right]. \qquad (34)$$

Moreover, by transforming the error via the linear mapping $h \to \mathbf{U}^\top \mathbf{B}^{1/2} h$, this can be written in the form

$$\mathrm{E}\left[\mathbf{U}^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = (\mathbf{I} - \omega\Lambda)^k \mathbf{U}^\top \mathbf{B}^{1/2}(x_0 - x_*), \qquad (35)$$

which is separable in the coordinates of the transformed error:

$$\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = (1 - \omega\lambda_i)^k u_i^\top \mathbf{B}^{1/2}(x_0 - x_*), \qquad i = 1, 2, \ldots, n. \qquad (36)$$

Finally,

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 = \sum_{i=1}^{n} (1 - \omega\lambda_i)^{2k} \left(u_i^\top \mathbf{B}^{1/2}(x_0 - x_*)\right)^2. \qquad (37)$$

Weak Convergence

Theorem 29 (Convergence 2)

Let $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Then for all $i = 1, 2, \ldots, n$,

$$\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = \begin{cases} 0 & \textit{if} \quad \lambda_i = 0, \\ (1 - \omega\lambda_i)^k u_i^\top \mathbf{B}^{1/2}(x_0 - x_*) & \textit{if} \quad \lambda_i > 0. \end{cases} \qquad (38)$$

Moreover,

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \le \rho^k(\omega) \|x_0 - x_*\|_{\mathbf{B}}^2, \qquad (39)$$

where the rate is given by

$$\rho(\omega) \overset{def}{=} \max_{i:\lambda_i > 0} (1 - \omega\lambda_i)^2. \qquad (40)$$

Necessary and Sufficient Conditions for Convergence

Corollary 30 (Necessary and sufficient conditions)

Let Assumption 3 (exactness) hold. Choose any $x_0 \in \mathbb{R}^n$ *and let* $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$.

If $\{x_k\}$ *are the random iterates produced by Algorithm 1, then the following statements are equivalent:*

(i) $|1 - \omega\lambda_i| < 1$ *for all* i *for which* $\lambda_i > 0$

(ii) $0 < \omega < 2/\lambda_{\max}$

(iii) $\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] \to 0$ *for all* i

(iv) $\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \to 0$

Proof of Theorems 28 and 29 - I

We first start with a lemma.

Lemma 31

Let Assumption 3 (exactness) hold. Consider arbitrary $x \in \mathbb{R}^n$ *and let* $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$. *If* $\lambda_i = 0$, *then* $u_i^\top \mathbf{B}^{1/2}(x - x_*) = 0$.

Proof.

From (18) we see that $x - x_* = \mathbf{B}^{-1}\mathbf{A}^\top w$ for some $w \in \mathbb{R}^m$. Therefore, $u_i^\top \mathbf{B}^{1/2}(x - x_*) = u_i^\top \mathbf{B}^{-1/2}\mathbf{A}^\top w$. By Theorem 18, we have $\mathrm{Range}\,(u_i \;:\; \lambda_i = 0) = \mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right)$, from which it follows that $u_i^\top \mathbf{B}^{-1/2}\mathbf{A}^\top = 0$. □

Proof of Theorem 28: Algorithm 1 can be written in the form

$$e_{k+1} = (\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_k)e_k, \tag{41}$$

where $e_k = x_k - x_*$. Multiplying both sides of this equation by $\mathbf{B}^{1/2}$ from the left, and taking expectation conditional on e_k, we obtain

$$\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1} \mid e_k\right] = (\mathbf{I} - \omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathbf{B}^{1/2}e_k.$$

Proof of Theorems 28 and 29 - II

Taking expectations on both sides and using the tower property, we get

$$\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1} \mid e_k\right]\right] = (\mathbf{I} - \omega \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathrm{E}\left[\mathbf{B}^{1/2}e_k\right].$$

We now replace $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$ by its eigenvalue decomposition $\mathbf{U}\Lambda\mathbf{U}^\top$ (see (32)), multiply both sides of the last equality by $\mathbf{U}^\top$ from the left, and use linearity of expectation to obtain

$$\mathrm{E}\left[\mathbf{U}^\top\mathbf{B}^{1/2}e_{k+1}\right] = (\mathbf{I} - \omega\Lambda)\mathrm{E}\left[\mathbf{U}^\top\mathbf{B}^{1/2}e_k\right].$$

Unrolling the recurrence, we get (35). When this is written coordinate-by-coordinate, (36) follows. Identity (37) follows immediately by equating standard Euclidean norms of both sides of (35).

Proof of Theorem 29: If $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$, then from Lemma 31 we see that $\lambda_i = 0$ implies $u_i^\top\mathbf{B}^{1/2}(x_0 - x_*) = 0$. Using this in (36) gives (38).

Proof of Theorems 28 and 29 - III

Finally, inequality (39) follows from

$$\begin{aligned}
\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 &\overset{(37)}{=} \sum_{i=1}^n (1-\omega\lambda_i)^{2k}\left(u_i^\top\mathbf{B}^{1/2}(x_0 - x_*)\right)^2 \\
&= \sum_{i:\lambda_i>0} (1-\omega\lambda_i)^{2k}\left(u_i^\top\mathbf{B}^{1/2}(x_0 - x_*)\right)^2 \\
&\overset{(40)}{\leq} \rho^k(\omega)\sum_{i:\lambda_i>0}\left(u_i^\top\mathbf{B}^{1/2}(x_0 - x_*)\right)^2 \\
&= \rho^k(\omega)\sum_{i:\lambda_i>0}\left(u_i^\top\mathbf{B}^{1/2}(x_0 - x_*)\right)^2 + \rho^k(\omega)\sum_{i:\lambda_i=0}\left(u_i^\top\mathbf{B}^{1/2}(x_0 - x_*)\right)^2 \\
&= \rho^k(\omega)\sum_i\left(u_i^\top\mathbf{B}^{1/2}(x_0 - x_*)\right)^2 \\
&= \rho^k(\omega)\sum_i (x_0 - x_*)^\top\mathbf{B}^{1/2}u_i u_i^\top\mathbf{B}^{1/2}(x_0 - x_*) \\
&= \rho^k(\omega)(x_0 - x_*)^\top\mathbf{B}^{1/2}\left(\sum_i u_i u_i^\top\right)\mathbf{B}^{1/2}(x_0 - x_*) = \rho^k(\omega)\|x_0 - x_*\|_{\mathbf{B}}^2.
\end{aligned}$$

The last identity follows from the fact that $\sum_i u_i u_i^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

Optimal Stepsize Choice for Weak Convergence

Convergence Rate as a Function of ω

We now consider the problem of choosing the stepsize (relaxation) parameter ω.

In view of (39) and (40), the optimal relaxation parameter is the one solving the following optimization problem:

$$\min_{\omega \in \mathbb{R}} \left\{ \rho(\omega) = \max_{i:\lambda_i > 0} (1 - \omega \lambda_i)^2 \right\}. \tag{42}$$

We solve the above problem in the next result (Theorem 32).

Optimal Stepsize

Theorem 32 (Stepsize Choice)

Let $\omega^* \overset{def}{=} 2/(\lambda_{\min}^+ + \lambda_{\max})$. *Then the objective of* (42) *is given by*

$$\rho(\omega) = \begin{cases} (1-\omega\lambda_{\max})^2 & \textit{if} \quad \omega \le 0 \\ (1-\omega\lambda_{\min}^+)^2 & \textit{if} \quad 0 \le \omega \le \omega^* \\ (1-\omega\lambda_{\max})^2 & \textit{if} \quad \omega \ge \omega^* \end{cases}. \tag{43}$$

Moreover, ρ *is decreasing on* $(-\infty, \omega^*]$ *and increasing on* $[\omega^*, +\infty)$*, and hence the optimal solution of* (42) *is* ω^**. Further, we have:*

(i) *If we choose* $\omega = 1$ *(no over-relaxation), then*

$$\rho(1) = (1-\lambda_{\min}^+)^2. \tag{44}$$

(ii) *If we choose* $\omega = 1/\lambda_{\max}$ *(over-relaxation), then*

$$\rho(1/\lambda_{\max}) = \left(1 - \tfrac{\lambda_{\min}^+}{\lambda_{\max}}\right)^2 \overset{(33)}{=} \left(1 - \tfrac{1}{\zeta}\right)^2. \tag{45}$$

(iii) *If we choose* $\omega = \omega^*$ *(optimal over-relaxation), the optimal rate is*

$$\rho(\omega^*) = \left(1 - \tfrac{2\lambda_{\min}^+}{\lambda_{\min}^+ + \lambda_{\max}}\right)^2 \overset{(33)}{=} \left(1 - \tfrac{2}{\zeta+1}\right)^2. \tag{46}$$

Proof of Theorem 32

Recall that $\lambda_{\max} \le 1$. Letting

$$\rho_i(\omega) = (1-\omega\lambda_i)^2,$$

it can be shown that

$$\rho(\omega) = \max\{\rho_j(\omega), \rho_n(\omega)\},$$

where j is such that $\lambda_j = \lambda_{\min}^+$. Note that $\rho_j(\omega) = \rho_n(\omega)$ for $\omega \in \{0, \omega^*\}$. From this we deduce that $\rho_j \ge \rho_n$ on $(-\infty, 0]$, $\rho_j \le \rho_n$ on $[0, \omega^*]$, and $\rho_j \ge \rho_n$ on $[\omega^*, +\infty)$, obtaining (43). We see that ρ is decreasing on $(-\infty, \omega^*]$, and increasing on $[\omega^*, +\infty)$.

The remaining results follow directly by plugging specific values of ω into (43).

Strong Convergence

Decrease of Distance is Proportional to $f_{\mathbf{S}}$

Lemma 33 (Decrease of Distance)

Choose $x_0 \in \mathbb{R}^n$ and let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by Algorithm 1, with an arbitrary relaxation parameter $\omega \in \mathbb{R}$. Let $x_* \in \mathcal{L}$.

Then we have the identities $\|x_{k+1} - x_k\|_{\mathbf{B}}^2 = 2\omega^2 f_{\mathbf{S}_k}(x_k)$, and

$$\|x_{k+1} - x_*\|_{\mathbf{B}}^2 = \|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega) f_{\mathbf{S}_k}(x_k). \tag{47}$$

Moreover, $\mathrm{E}\left[\|x_{k+1} - x_k\|_{\mathbf{B}}^2\right] = 2\omega^2 \mathrm{E}\left[f(x_k)\right]$, and

$$\mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2\right] = \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] - 2\omega(2-\omega)\mathrm{E}\left[f(x_k)\right]. \tag{48}$$

Remarks: Equation (47) says that for any $x_* \in \mathcal{L}$, in the k-th iteration of Algorithm 1 the distance of the current iterate from x_* decreases by the amount $2\omega(2-\omega) f_{\mathbf{S}_k}(x_k)$.

Lower Bound on a Quadratic

Lemma 34

Let Assumption 3 be satisfied. Then the inequality

$$x^\top \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}x \geq \lambda_{\min}^{+}(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})x^\top x \qquad (49)$$

holds for all $x \in \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right)$.

Proof.

It is known that for any matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, the inequality

$$x^\top \mathbf{M}^\top \mathbf{M} x \geq \lambda_{\min}^{+}(\mathbf{M}^\top \mathbf{M})x^\top x$$

holds for all $x \in \mathrm{Range}\left(\mathbf{M}^\top\right)$. Applying this with $\mathbf{M} = (\mathrm{E}\left[\mathbf{Z}\right])^{1/2}\mathbf{B}^{-1/2}$, we see that (49) holds for all $x \in \mathrm{Range}\left(\mathbf{B}^{-1/2}(\mathrm{E}\left[\mathbf{Z}\right])^{1/2}\right)$. However,

$$\begin{aligned}\mathrm{Range}\left(\mathbf{B}^{-1/2}(\mathrm{E}\left[\mathbf{Z}\right])^{1/2}\right) &= \mathrm{Range}\left(\mathbf{B}^{-1/2}(\mathrm{E}\left[\mathbf{Z}\right])^{1/2}(\mathbf{B}^{-1/2}(\mathrm{E}\left[\mathbf{Z}\right])^{1/2})^\top\right)\\ &= \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\right) = \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right),\end{aligned}$$

where the last identity follows by combining Assumption 3 and Theorem 18. □

Proof of Lemma 33 - I

Recall that Algorithm 1 performs the update

$$x_{k+1} = x_k - \omega \mathbf{B}^{-1}\mathbf{Z}_k(x_k - x_*).$$

From this we get

$$\begin{aligned}\|x_{k+1} - x_k\|_{\mathbf{B}}^2 &= \omega^2\|\mathbf{B}^{-1}\mathbf{Z}_k(x_k - x_*)\|_{\mathbf{B}}^2\\ &\overset{(20)}{=} \omega^2(x_k - x_*)^\top \mathbf{Z}_k(x_k - x_*)\\ &\overset{(21)}{=} 2\omega^2 f_{\mathbf{S}_k}(x_k). \qquad (50)\end{aligned}$$

In a similar vein,

$$\begin{aligned}\|x_{k+1} - x_*\|_{\mathbf{B}}^2 &= \|(\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*)\|_{\mathbf{B}}^2\\ &= (x_k - x_*)^\top(\mathbf{I} - \omega\mathbf{Z}_k\mathbf{B}^{-1})\mathbf{B}(\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*)\\ &\overset{(20)}{=} (x_k - x_*)^\top(\mathbf{B} - \omega(2-\omega)\mathbf{Z}_k)(x_k - x_*)\\ &\overset{(21)}{=} \|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega)f_{\mathbf{S}_k}(x_k), \qquad (51)\end{aligned}$$

Proof of Lemma 33 - II

establishing (47).

Taking expectation in (50) and using the tower property, we get

$$\begin{aligned} \mathrm{E}\left[\|x_{k+1}-x_k\|_{\mathbf{B}}^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\|x_{k+1}-x_k\|_{\mathbf{B}}^2 \mid x_k\right]\right] \\ &\overset{(50)}{=} 2\omega^2\mathrm{E}\left[\mathrm{E}\left[f_{\mathbf{S}_k}(x_k) \mid x_k\right]\right] \\ &= 2\omega^2\mathrm{E}\left[f(x_k)\right], \end{aligned}$$

where in the last step we have used the definition of f.

Taking expectation in (47), we get

$$\begin{aligned} \mathrm{E}\left[\|x_{k+1}-x_*\|_{\mathbf{B}}^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\|x_{k+1}-x_*\|_{\mathbf{B}}^2 \mid x_k\right]\right] \\ &\overset{(51)}{=} \mathrm{E}\left[\|x_k-x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega)f(x_k)\right] \\ &= \mathrm{E}\left[\|x_k-x_*\|_{\mathbf{B}}^2\right] - 2\omega(2-\omega)\mathrm{E}\left[f(x_k)\right]. \end{aligned}$$

Quadratic Bounds

Lemma 35 (Quadratic bounds)

For all $x \in \mathbb{R}^n$ and $x_* \in \mathcal{L}$ we have

$$\lambda_{\min}^+ \cdot f(x) \leq \frac{1}{2}\|\nabla f(x)\|_{\mathbf{B}}^2 \leq \lambda_{\max} \cdot f(x). \tag{52}$$

and

$$f(x) \leq \frac{\lambda_{\max}}{2}\|x-x_*\|_{\mathbf{B}}^2. \tag{53}$$

Moreover, if Assumption 3 holds, then for all $x \in \mathbb{R}^n$ and $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$ we have

$$\frac{\lambda_{\min}^+}{2}\|x-x_*\|_{\mathbf{B}}^2 \leq f(x). \tag{54}$$

Proof of Lemma 35 - I

In view of (16) and (32), we obtain a spectral characterization of f:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \left(u_i^\top \mathbf{B}^{1/2}(x - x_*)\right)^2, \tag{55}$$

where x_* is any point in $\mathcal{L}$. On the other hand, in view of (27) and (32), we have

$$\begin{aligned}
\|\nabla f(x)\|_{\mathbf{B}}^2 &= \|\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*)\|_{\mathbf{B}}^2 & (56)\\
&= (x - x_*)^\top \mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*) \\
&= (x - x_*)^\top \mathbf{B}^{1/2}(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathbf{B}^{1/2}(x - x_*) \\
&= (x - x_*)^\top \mathbf{B}^{1/2}\mathbf{U}(\mathbf{U}^\top\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\mathbf{U})^2\mathbf{U}^\top\mathbf{B}^{1/2}(x - x_*) \\
&\overset{(32)}{=} (x - x_*)^\top \mathbf{B}^{1/2}\mathbf{U}\Lambda^2\mathbf{U}^\top\mathbf{B}^{1/2}(x - x_*) \\
&= \sum_{i=1}^{n} \lambda_i^2 \left(u_i^\top \mathbf{B}^{1/2}(x - x_*)\right)^2. & (57)
\end{aligned}$$

Inequality (52) follows by comparing (55) and (56), using the bounds

$$\lambda_{\min}^{+}\lambda_i \leq \lambda_i^2 \leq \lambda_{\max}\lambda_i,$$

which hold for i for which $\lambda_i > 0$.

Proof of Lemma 35 - II

We now move to the bounds involving norms. First, note that for any $x_* \in \mathcal{L}$ we have

$$\begin{aligned}
f(x) &\overset{(16)}{=} \frac{1}{2}(x - x_*)^\top \mathrm{E}\left[\mathbf{Z}\right](x - x_*) & (58)\\
&= \frac{1}{2}(\mathbf{B}^{1/2}(x - x_*))^\top(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathbf{B}^{1/2}(x - x_*).
\end{aligned}$$

The upper bound follows by applying the inequality

$$\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2} \preceq \lambda_{\max}\mathbf{I}.$$

If $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$, then in view of (18), we have

$$\mathbf{B}^{1/2}(x - x_*) \in \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right).$$

Applying Lemma 34 to (58), we get the lower bound.

Strong Convergence

Theorem 36 (Strong convergence)

Let Assumption 3 (exactness) hold and set $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$*. Let* $\{x_k\}$ *be the random iterates produced by Algorithm 1, where the relaxation parameter satisfies* $0 < \omega < 2$*, and let* $r_k \overset{def}{=} \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$*. Then for all* $k \geq 0$ *we have*

$$(1 - \omega(2-\omega)\lambda_{\max})^k r_0 \leq r_k \leq (1 - \omega(2-\omega)\lambda_{\min}^+)^k r_0. \qquad (59)$$

The best rate is achieved when $\omega = 1$.

Proof.

Let $\phi_k = \mathrm{E}\left[f(x_k)\right]$. We have

$$r_{k+1} \overset{(48)}{=} r_k - 2\omega(2-\omega)\phi_k \overset{(54)}{\leq} r_k - \omega(2-\omega)\lambda_{\min}^+ r_k,$$

and

$$r_{k+1} \overset{(48)}{=} r_k - 2\omega(2-\omega)\phi_k \overset{(53)}{\geq} r_k - \omega(2-\omega)\lambda_{\max} r_k.$$

Inequalities (59) follow from this by unrolling the recurrences.

□

Convergence of $f(x_k)$

Convergence of $f(x_k)$

Theorem 37 (Convergence of f)

Choose $x_0 \in \mathbb{R}^n$, and let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by Algorithm 1, where the relaxation parameter satisfies $0 < \omega < 2$.

(i) *Let $x_* \in \mathcal{L}$. The average iterate $\hat{x}_k \overset{def}{=} \frac{1}{k}\sum_{t=0}^{k-1} x_t$ for all $k \geq 1$ satisfies*

$$\mathrm{E}\left[f(\hat{x}_k)\right] \leq \frac{\|x_0 - x_*\|_{\mathbf{B}}^2}{2\omega(2-\omega)k}. \tag{60}$$

(ii) *Now let Assumption 3 hold. For $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$ and $k \geq 0$ we have*

$$\mathrm{E}\left[f(x_k)\right] \leq \left(1 - \omega(2-\omega)\lambda_{\min}^{+}\right)^k \frac{\lambda_{\max}\|x_0 - x_*\|_{\mathbf{B}}^2}{2}. \tag{61}$$

The best rate is achieved when $\omega = 1$.

Proof of Theorem 37

(i) Let $\phi_k = \mathrm{E}\left[f(x_k)\right]$ and $r_k = \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$. By summing up the identities from (48), we get

$$2\omega(2-\omega)\sum_{t=0}^{k-1}\phi_t = r_0 - r_k.$$

Therefore, using Jensen's inequality, we get

$$\mathrm{E}\left[f(\hat{x}_k)\right] \leq \mathrm{E}\left[\frac{1}{k}\sum_{t=0}^{k-1} f(x_t)\right] = \frac{1}{k}\sum_{t=0}^{k-1}\phi_t = \frac{r_0 - r_k}{2\omega(2-\omega)k} \leq \frac{r_0}{2\omega(2-\omega)k}.$$

(ii) Combining inequality (53) with Theorem 36, we get

$$\mathrm{E}\left[f(x_k)\right] \leq \frac{\lambda_{\max}}{2}\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \overset{(59)}{\leq} \left(1 - \omega(2-\omega)\lambda_{\min}^{+}\right)^k \frac{\lambda_{\max}\|x_0 - x_*\|_{\mathbf{B}}^2}{2}.$$

Introduction to Randomized Methods in Convex Optimization
Peter Richtárik

5. Parallel and Accelerated Methods

Parallel Method ("Minibatch Method")

Parallel Method ("Minibatch Method")

Algorithm 2 Parallel Method

```
Parameters: distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$; stepsize/relaxation parameter $\omega \in \mathbb{R}$; parallelism parameter $\tau$ (aka "minibatch size")
Choose $x_0 \in \mathbb{R}^n$                                   ▷ Initialization
for $k = 0, 1, 2, \ldots$ do
    for $i = 1, 2, \ldots, \tau$ do
        Draw $\mathbf{S}_{ki} \sim \mathcal{D}$
        Set $z_{k+1,i} = x_k - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_{ki}(\mathbf{S}_{ki}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_{ki})^\dagger \mathbf{S}_{ki}^\top(\mathbf{A}x_k - b)$
    Set $x_{k+1} = \frac{1}{\tau}\sum_{i=1}^{\tau} z_{k+1,i}$              ▷ Average the results
```

- Note that for $\tau = 1$, the parallel method (Algorithm 2) **reduces to the basic method (Algorithm 1).**
- We take one step of the basic method τ times, independently, started from x_k. The results are then averaged to obtain x_{k+1}.
- The τ computations **can (but do not have to!) be performed in parallel,** whence the name of the method.

Convergence of the Parallel Method

Theorem 38

Let Assumption 3 hold and set $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by Algorithm 2, where the relaxation parameter satisfies $0 < \omega < 2/\xi(\tau)$, where $\xi(\tau) \overset{def}{=} \frac{1}{\tau} + \left(1 - \frac{1}{\tau}\right)\lambda_{\max}$. Then

$$\mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2\right] \leq \rho(\omega, \tau) \cdot \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right],$$

and

$$\mathrm{E}\left[f(x_k)\right] \leq \rho(\omega, \tau)^k \frac{\lambda_{\max}}{2}\|x_0 - x_*\|_{\mathbf{B}}^2,$$

where

$$\rho(\omega, \tau) \overset{def}{=} 1 - \omega\left[2 - \omega\xi(\tau)\right]\lambda_{\min}^+.$$

Understanding the Behaviour of the Parallel Method - I

The **convergence factor**

$$\rho(\omega, \tau) = 1 - \omega\left[2 - \omega\underbrace{\left(\tfrac{1}{\tau} + \left(1 - \tfrac{1}{\tau}\right)\lambda_{\max}\right)}_{\xi(\tau)}\right]\lambda_{\min}^+$$

depends on the choice of the stepsize ω and on the minibatch size τ.

- The **stepsize rate function**

  $$\omega \mapsto \rho(\omega, \tau),$$

  is minimized for $\omega(\tau) \overset{\text{def}}{=} 1/\xi(\tau)$ and the associated **optimal rate** is

  $$\rho(\omega(\tau), \tau) = 1 - \frac{\lambda_{\min}^+}{\frac{1}{\tau} + \left(1 - \frac{1}{\tau}\right)\lambda_{\max}}. \tag{62}$$

- The **minibatch rate function**

  $$\tau \mapsto \rho(\omega(\tau), \tau)$$

  is **decreasing on $[1, \infty)$,** with

  $$\rho(\omega(1), 1) = 1 - \lambda_{\min}^+, \qquad \lim_{\tau \to \infty} \rho(\omega(\tau), \tau) = 1 - \frac{\lambda_{\min}^+}{\lambda_{\max}}.$$

Understanding the Behaviour of the Parallel Method - II

Convergence Rate for $\tau = 1$ (with optimal stepsize $\omega = \omega(\tau)$):

$$k \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{\|x_0 - x_*\|_{\mathbf{B}}^2}{\epsilon}\right) \quad \Rightarrow \quad \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \leq \epsilon$$

Convergence Rate for $\tau = +\infty$ (with optimal stepsize $\omega = \omega(\tau)$):

$$k \geq \frac{\lambda_{\max}}{\lambda_{\min}^{+}} \log\left(\frac{\|x_0 - x_*\|_{\mathbf{B}}^2}{\epsilon}\right) \quad \Rightarrow \quad \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \leq \epsilon$$

Recall what we proved about the basic method:

- The **weak convergence rate of the basic method is "fast"**:

$$\tilde{\mathcal{O}}\left(\lambda_{\max}/\lambda_{\min}^{+}\right)$$

- The **strong convergence rate of the basic method is "slow"**:

$$\tilde{\mathcal{O}}\left(1/\lambda_{\min}^{+}\right)$$

So, how does minibatching improve the basic method?

- The **strong convergence rate of the parallel method** interpolates between slow and fast!

Accelerated Method

Accelerated Method

In order to obtain further acceleration, we suggest to perform an update step in which x_{k+1} depends on both x_k and x_{k-1}. In particular, we take two *dependent* steps of Algorithm 1, one from x_k and one from x_{k-1}, and then take an affine combination of the results. That is, the process is started with $x_0, x_1 \in \mathbb{R}^n$, and for $k \geq 1$ involves an iteration of the form

$$\boxed{x_{k+1} = \gamma \phi_\omega(x_k, \mathbf{S}_k) + (1-\gamma)\phi_\omega(x_{k-1}, \mathbf{S}_{k-1})} \tag{63}$$

where the matrices $\{\mathbf{S}_k\}$ are independent samples from $\mathcal{D}$, and $\gamma \in \mathbb{R}$ is an **acceleration parameter.**

Remarks:

- By choosing $\gamma = 1$ (no acceleration), we recover the Basic Method.
- Theory suggests that γ should be always between 1 and 2. In particular, for well conditioned problems (small ζ), one should choose $\gamma \approx 1$, and for ill conditioned problems (large ζ), one should choose $\gamma \approx 2$.
- By a proper combination of overrelaxation (choice of ω) with acceleration (choice of γ), Algorithm 3 enjoys the **accelerated convergence rate of $\tilde{\mathcal{O}}(\sqrt{\zeta})$, where ζ is the condition number.**

Accelerated Method

Algorithm 3 Accelerated Method

```
Parameters: distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$; stepsize/relaxation parameter $\omega > 0$; acceleration parameter $\gamma > 0$
Choose $x_0, x_1 \in \mathbb{R}^n$ such that $x_0 - x_1 \in \mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\right)$ (for instance, choose $x_0 = x_1$)
Draw $\mathbf{S}_0 \sim \mathcal{D}$
Set $z_0 = \phi_\omega(x_0, \mathbf{S}_0)$
for $k = 1, 2, \dots$ do
    Draw a fresh sample $\mathbf{S}_k \sim \mathcal{D}$
    Set $z_k = \phi_\omega(x_k, \mathbf{S}_k)$
    Set $x_{k+1} = \gamma z_k + (1-\gamma) z_{k-1}$        ▷ Main update step
Output $x_k$
```

Convergence

Theorem 39 (Complexity of Algorithm 3)

Let Assumption 3 (exactness) be satisfied and let $\{x_k\}_{k=0}^{\infty}$ *be the sequence of random iterates produced by Algorithm 3, started with* $x_0, x_1 \in \mathbb{R}^n$ *satisfying the relation* $x_0 - x_1 \in \mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\right)$*, with* **relaxation parameter** $0 < \omega \leq 1/\lambda_{\max}$ *and* **acceleration parameter** $\gamma = 2/(1+\sqrt{\mu})$*, where* $\mu \in (0, \omega\lambda_{\min}^+)$*. Let* $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$*. Then there exists a constant* $C > 0$*, such that for all* $k \geq 2$ *we have*

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq (1-\sqrt{\mu})^{2k} C. \tag{64}$$

(i) *If we choose* $\omega = 1/\lambda_{\max}$ *(overrelaxation), then we can pick* $\mu = 0.99/\zeta$ *(recall that* $\zeta = \lambda_{\max}/\lambda_{\min}^+$ *is the condition number), which leads to the rate*

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{\frac{0.99\lambda_{\min}^+}{\lambda_{\max}}}\right)^{2k} C. \tag{65}$$

(ii) *If we choose* $\omega = 1$ *(no overrelaxation), then we can pick* $\mu = 0.99\lambda_{\min}^+$*, which leads to the rate*

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{0.99\lambda_{\min}^+}\right)^{2k} C. \tag{66}$$

Comments

Alternative Way of Writing Convergence Rate (65):

$$k \geq \frac{1}{2\sqrt{0.99}}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^+}}\log\left(\frac{C}{\epsilon}\right) \quad \Rightarrow \quad \|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \epsilon$$

Alternative Way of Writing Convergence Rate (66):

$$k \geq \frac{1}{2\sqrt{0.99}}\sqrt{\frac{\mathbf{1}}{\lambda_{\min}^+}}\log\left(\frac{C}{\epsilon}\right) \quad \Rightarrow \quad \|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \epsilon$$

- All three methods: basic (Algorithm 1), parallel (Algorithm 2) and accelerated (Algorithm 3) enjoy **linear convergence.** That is, their complexity has logarithmic dependence on $1/\epsilon$. This means that the error decays exponentially fast.
- However, **the leading constants in the complexity bounds are different.**
- Both the **basic and parallel methods** depend either on $1/\lambda^+_{\min}$ (slow) or $\lambda_{\max}/\lambda^+_{\min}$ (fast), depending on how we set the parameters ω, τ and γ, and whether we are interested in weak or strong convergence.
- However, the **accelerated method depends on the square root of these quantities.** This is why the method is called accelerated.

Introduction to Randomized Methods in Convex Optimization
Peter Richtárik

6. Duality

Motivation

- Recall that assuming exactness, and under certain assumptions in the stepsize ω, the iterates of the **basic method** converge[4] in the weak sense and/or in the strong sense to

$$x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0).$$

- That is, the basic method actually solves the optimization problem:

$$\begin{array}{rl} \text{minimize} & P(x) \overset{\text{def}}{=} \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2 \\ \text{subject to} & \mathbf{A}x = b \\ & x \in \mathbb{R}^n. \end{array} \tag{67}$$

- We will call (67) the **primal problem,** and P the **primal objective function.**
- In optimization, one can associate with each optimization problem a closely related optimization problem, called the **dual problem.**
- We shall now investigate several very interesting relationships between the primal and the dual problems.

[4] This is also true for the parallel and accelerated methods. However, we shall not deal with them in this lecture.

Dual Problem

Dual Problem: Concave Quadratic Maximization

The **dual problem** to (67) is the optimization problem

$$\begin{aligned} \text{maximize} \quad & D(y) \overset{\text{def}}{=} (b - \mathbf{A}x_0)^\top y - \tfrac{1}{2}\|\mathbf{A}^\top y\|^2_{\mathbf{B}^{-1}} \\ \text{subject to} \quad & y \in \mathbb{R}^m. \end{aligned} \tag{68}$$

- $D : \mathbb{R}^m \to \mathbb{R}$ is the **dual objective function** (quadratic)
- The dimension of the dual variable (y) is m (# rows of $\mathbf{A}$). The dimension of the primal variable (x) is n (# columns of $\mathbf{A}$).
- A more detailed look at the terms:
  - The first term, $(b - \mathbf{A}x_0)^\top y$, is linear in y.
  - The second term can be written as $-\frac{1}{2}y^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y$.
  - Thus, the **gradient and Hessian of D** are given by:

$$\nabla D(y) = b - \mathbf{A}x_0 - \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y, \qquad \nabla^2 D(y) = -\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \tag{69}$$

- Note that $\nabla^2 D(y)$ is a **negative semidefinite matrix.** Equivalently, $-\nabla^2 D(y)$ is a **positive semidefinite matrix.** Hence
  - D is a concave quadratic function
  - $-D$ is a convex quadratic function

Weak Duality

Lemma 40 (Weak Duality)

For any **primal feasible point** *x (i.e., $x \in \mathbb{R}^n$ for which $\mathbf{A}x = b$) and for any* **dual feasible point** *(i.e., $y \in \mathbb{R}^m$), we have*

$$P(x) \geq D(y).$$

Proof.

For any $x \in \mathbb{R}^n$ for which $\mathbf{A}x = b$ and for any $y \in \mathbb{R}^m$ we have

$$\begin{aligned} P(x) - D(y) & \overset{(67)+(68)}{=} \tfrac{1}{2}\|x - x_0\|^2_{\mathbf{B}} + \tfrac{1}{2}\|\mathbf{A}^\top y\|^2_{\mathbf{B}^{-1}} + (x - x_0)^\top \mathbf{A}^\top y \\ & = \tfrac{1}{2}\|\mathbf{B}^{1/2}(x - x_0)\|^2 + \tfrac{1}{2}\|\mathbf{B}^{-1/2}\mathbf{A}^\top y\|^2 + (x_0 - x)^\top \mathbf{A}^\top y \\ & = \tfrac{1}{2}\|\mathbf{B}^{-1/2}\mathbf{A}^\top y + \mathbf{B}^{1/2}(x_0 - x)\|^2 \\ & = \tfrac{1}{2}\|x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y - x\|^2_{\mathbf{B}} \geq 0. \end{aligned}$$

□

Optimality Conditions

Definition 41 (Duality Mapping)

The **duality mapping** is the function $x(y) : \mathbb{R}^m \to \mathbb{R}^n$ defined by

$$x(y) \overset{\text{def}}{=} x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y. \tag{70}$$

Theorem 42

(i) **Dual boundedness.** *D is bounded above $\Leftrightarrow$ the primal problem is feasible*

(ii) **Dual optimality.**

$$y \text{ is dual optimal} \quad \Leftrightarrow \quad \mathbf{A}x(y) = b \tag{71}$$

(iii) **Primal optimality.**

$$x = x_* \quad \Leftrightarrow \quad \mathbf{A}x = b \quad \text{and} \quad x = x(y) \text{ for some } y \tag{72}$$

(iv) **x_* can be obtained from any dual optimal point:**

$$y_* \text{ is dual optimal} \quad \Rightarrow \quad x(y_*) = x_* \tag{73}$$

Convex Quadratic Optimization

Exercise 7

Consider a general **convex quadratic optimization** *problem*

$$\min_{y\in\mathbb{R}^m} \tfrac{1}{2}y^\top \mathbf{Q}y + d^\top y,$$

and assume that the problem is bounded. Show that the problem can be equivalently written in the form (70) *for suitable $\mathbf{A}, \mathbf{B}$, x_0 and b.*

Proof of Theorem 42

(i) Since D is a concave quadratic function, it has a maximizer if and only if there exists y such that $\nabla D(y) = 0$ (in which case any such y is a maximizer). In view of (69), this happens if and only if the following linear system has a solution:

$$\mathbf{AB}^{-1}\mathbf{A}^\top y = b - \mathbf{A}x_0. \tag{74}$$

This system has a solution if and only if

$$b - \mathbf{A}x_0 \in \operatorname{Range}\left(\mathbf{AB}^{-1}\mathbf{A}^\top\right) \overset{Fact\ 17iii}{=} \operatorname{Range}(\mathbf{A}).$$

(ii) Using the reasoning in (i), we know that y is dual optimal $\Leftrightarrow$ y solves (74). It remains to notice that (74) can equivalently be written as $\mathbf{A}x(y) = b$.

(iii) Do this as an exercise. *Hint:* Use weak duality; in particular, the derived expression for $P(x) - D(y)$.

(iv) This follows by combining (ii) and (iii).

Dual Suboptimality vs Primal Suboptimality

The dual-to-primal mapping enjoys the following insightful property:

Theorem 43

Let y_* be any dual optimal point and $y \in \mathbb{R}^m$. Then

$$D(y_*) - D(y) = \tfrac{1}{2}\|x_* - x(y)\|^2_{\mathbf{B}}. \tag{75}$$

Proof.

$$\begin{aligned}
D(y_*) - D(y) &\overset{(68)}{=} (b - \mathbf{A}x_0)^\top (y_* - y) - \tfrac{1}{2} y_*^\top \mathbf{AB}^{-1}\mathbf{A}^\top y_* + \tfrac{1}{2} y^\top \mathbf{AB}^{-1}\mathbf{A}^\top y \\
&\overset{(70)+(71)}{=} y_*^\top \mathbf{AB}^{-1}\mathbf{A}^\top (y_* - y) - \tfrac{1}{2} y_*^\top \mathbf{AB}^{-1}\mathbf{A}^\top y_* + \tfrac{1}{2} y^\top \mathbf{AB}^{-1}\mathbf{A}^\top y \\
&= \tfrac{1}{2}(y - y_*)^\top \mathbf{AB}^{-1}\mathbf{A}^\top (y - y_*) \\
&\overset{(70)}{=} \tfrac{1}{2}\|x(y) - x(y_*)\|^2_{\mathbf{B}}.
\end{aligned}$$

It remains to use (73) which states that $x(y_*) = x_*$. □

Dual Algorithms Solve the Primal Problem

Let $\{y_k\}_{k=0}^{\infty}$ be any sequence for which

$$D(y_k) \rightarrow D(y_*).$$

Such a sequence can be obtained by **any algorithm that solves the dual problem.** In view of Theorem 43, we automatically have

$$x(y_k) \rightarrow x(y_*) = x_*.$$

Now, define an **associated primal algorithm** via the iterates:

$$x_k \overset{\text{def}}{=} x(y_k). \tag{76}$$

Conclusion: **Any convergent dual algorithm automatically leads to a convergent primal algorithm.**

Stochastic Dual Subspace Ascent

Algorithm: Stochastic Dual Subspace Ascent (SDSA)

Consider the following algorithm for solving the dual problem (68):

$$\boxed{y_{k+1} = y_k + \mathbf{S}_k \lambda_k} \tag{77}$$

Above, $\mathbf{S}_k$ is a fresh sample from $\mathcal{D}$, and λ_k is a suitably chosen **"stepsize"** parameter. We refer to this method by the name **stochastic dual subspace ascent (SDSA).**

- **Why stochastic?** Because the iterates are random vectors, which follows from the fact that $\mathbf{S}_k$ is a random matrix.
- **Why subspace?** The step, $\mathbf{S}_k\lambda_k$, can potentially be any point in a specific random subspace of $\mathbb{R}^m$. In particular, this is the space $\mathrm{Range}\,(\mathbf{S}_k)$, i.e., the subspace spanned by the columns of the random matrix $\mathbf{S}_k$. We hope that by focusing on a random subspace (of a sufficiently small dimension) in each iteration, we can perform the iteration much faster, particularly if m is big.
- **Why ascent?** We wish the method to always improve the dual function value (or, at least, not to make it worse): $D(y_{k+1}) \geq D(y_k)$. We achieve this by an appropriate choice of λ_k. In particular, in SDSA we pick the best vector λ_k; i.e., the vector for which $D(y_k + \mathbf{S}_k\lambda_k)$ is maximized!

How to Compute the Best λ_k? I

In SDSA we pick the stepsize parameter λ_k via

$$\lambda_k \overset{\text{def}}{=} \arg\max_{\lambda} D(y_k + \mathbf{S}_k\lambda).$$

Since the function $\psi(\lambda) = D(y_k + \mathbf{S}_k\lambda)$ is a concave quadratic, λ is its maximizer if and only if

$$\nabla\psi(\lambda) = 0. \tag{78}$$

Since

$$\begin{aligned}
\nabla\psi(\lambda) &= \mathbf{S}_k^\top \nabla D(y_k + \mathbf{S}_k\lambda) \overset{(69)}{=} \mathbf{S}_k^\top(b - \mathbf{A}x_0 - \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top(y_k + \mathbf{S}_k\lambda)) \\
&= \mathbf{S}_k^\top \left[b - \mathbf{A}\underbrace{\left(x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y_k\right)}_{\overset{(70)}{=} x(y_k)} \right] - \mathbf{S}_k^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}_k\lambda \\
&= \mathbf{S}_k^\top(b - \mathbf{A}x(y_k)) - \mathbf{S}_k^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}_k\lambda,
\end{aligned}$$

How to Compute the Best λ_k? II

equation (78) is equivalent to the **linear system:**

$$\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k \lambda = \mathbf{S}_k^\top (b - \mathbf{A}x(y_k)). \tag{79}$$

If we wish to be greedy, we may choose λ_k as any solution of the linear system (79). In SDSA, we pick a **particular solution** of (79): **the least-norm solution.** In view of Exercise 5, the least-norm solution of a linear system is given by applying the pseudoinverse of the system matrix to the right hand side. Thus, we get:

$$\begin{aligned} \lambda_k &\overset{\text{def}}{=} \arg\min_{\lambda} \{\|\lambda\| \;:\; (79) \textit{ holds}\} \qquad (80)\\ &\overset{\text{Exercise 5}}{=} (\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top (b - \mathbf{A}x(y_k)). \qquad (81) \end{aligned}$$

Plugging this back into the SDSA iteration (77), we get

$$\boxed{y_{k+1} \overset{(77)+(81)}{=} y_k - \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top (\mathbf{A}x(y_k) - b)} \tag{82}$$

Duality of SDSA and the Basic Method with Unit Stepsize

A natural question: How do the iterates of the primal algorithm (defined in (76)) associated with the dual iterates of SDSA (defined in (82)) look like?

$$\begin{aligned} x(y_{k+1}) &\overset{(70)}{=} x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y_{k+1}\\ &\overset{(82)}{=} x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y_k - \mathbf{B}^{-1}\mathbf{A}^\top \underbrace{\mathbf{S}_k(\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top}_{\mathbf{H}_k} (\mathbf{A}x(y_k) - b)\\ &\overset{(76)}{=} x(y_k) - \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}_k(\mathbf{A}x(y_k) - b). \end{aligned}$$

Observe:

- If we set $y_0 = 0$,then $x(y_0) = x_0$
- **This is the basic method with unit stepsize!** (see (7))

Thus, we obtain the following result:

Theorem 44 (The Basic Method with Unit Stepsize is a "Mirror Image" of SDSA)

Let $y_0 = 0$ and let $\{y_k\}_{k=0}^\infty$ be the iterates (82) of SDSA. Then the primal iterates $x_k = x(y_k)$ associated with SDSA exactly correspond to the basic method with unit stepsize ($\omega = 1$).

Convergence of SDSA

By applying Theorem 43 to SDSA (with starting point $y_0 = 0$) and iterates $\{y_k\}$, we get

$$D(y_*) - D(y_k) = \tfrac{1}{2}\|x_* - x_k\|_{\mathbf{B}}^2,$$

where in view of Theorem 44, $\{x_k\}$ are the iterates of the basic method with unit stepsize.

By taking expectations on both sides of the above identity, we get

$$\mathrm{E}\left[D(y_*) - D(y_k)\right] = \tfrac{1}{2}\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]. \tag{83}$$

By applying Theorem 36 (strong convergence of the basic method) to (83), with $\omega = 1$, we get:

Theorem 45 (Convergence of SDSA)

Choose any $x_0 \in \mathbb{R}^n$. Let Assumption 3 (exactness) hold and set $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Let $y_0 = 0$ and $\{y_k\}_{k=0}^{\infty}$ be the random iterates produced by SDSA (see (82)). Further, let $t_k \overset{def}{=} \mathrm{E}\left[D(y_*) - D(y_k)\right]$. Then for all $k \geq 0$ we have

$$(1 - \lambda_{\max})^k t_0 \leq t_k \leq (1 - \lambda_{\min}^+)^k t_0. \tag{84}$$

Special Case: $\mathbf{S}_k$ is a Random Vector

If $\mathbf{S}_k$ has a single column only, then SDSA is moving in the **random direction $\mathbf{S}_k \in \mathbb{R}^m$,** using **stepsize $\lambda_k \in \mathbb{R}$.**

Special cases:

- If $\mathbf{S}_k$ is a **random coordinate vector,** i.e., if $\mathcal{D}$ is given by $\mathbf{S}_k = e_i$ (the ith unit basis vector in $\mathbb{R}^m$) with probability $p_i > 0$, then SDSA is called **stochastic dual coordinate ascent (SDCA).**
- If $\mathbf{S}_k$ is a **random Gaussian vector,** then SDSA is called **stochastic dual Gaussian ascent (SDGA).**

Bibliographic Comments

Sections 1-5 are based on [1, 3]. Section 6 is based on [2].

References

[1] Robert M. Gower and P.R. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications* 36(4), 1660-1690, 2015.

[2] Robert M. Gower and P.R. Stochastic dual ascent for solving linear systems. *ArXiv:1512.06890*, 2015.

[3] P.R. and Martin Takáč. Stochastic reformulations of linear systems: algorithms and convergence theory. *ArXiv:1706.01108*, 2017.