

Stochastic Optimization Under Almost-Sure Affine Inequality Constraints

Amir Miri Lavasani

November 27, 2025

Contents

1	Introduction	3
2	Theory Background	5
2.1	Convex Optimization	5
2.2	Probability Theory	7
2.3	Stochastic Optimization	10
2.4	Norms and Inequalities	11
3	Sequential Penalty Methods	12
3.1	Consistency of Solutions	12
3.2	Sequential SGD	15
3.2.1	Bounding the surrogate error	16
3.2.2	Bounding the tracking error	18
3.2.3	Convergence rates	20
3.2.4	Iterate averaging	24
3.3	Fast Convergence with Iterate Moving Averages	25
4	Applications & Numerical Examples	32
4.1	Optimal Control	32
4.2	Reinforcement Learning	32
5	Summary and Outlook	33

1 Introduction

Stochastic optimization is concerned with problems of the form

$$\min_{x \in \mathcal{X}} \left\{ f(x) := \mathbb{E}(F(x, \xi)) \right\}.$$

Here, $\emptyset \neq \mathcal{X} \subset \mathbb{R}^d$, $F: \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$, $\xi: \Omega \rightarrow \mathbb{R}^m$ is a random vector on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\omega \mapsto F(x, \xi(\omega))$ is assumed to be measurable for all $x \in \mathbb{R}^d$, so that f is well-defined. Typically, the set \mathcal{X} is called the **feasible set** and f is called the **objective function**, or simply **objective**.

We consider a constrained stochastic optimization problem of the form

$$\begin{aligned} & \min_{x \in \mathbb{R}^d} \left\{ j(x) := \mathbb{E} \left(\frac{1}{2} \|y - b\|^2 + \frac{\lambda}{2} \|x\|^2 \right) \right\} \\ & \text{subject to (s. t.) } B(\xi)y = C(\xi)x \quad \text{almost surely (a. s.)} \\ & \quad y \leq c, \end{aligned}$$

where $y, b, c \in \mathbb{R}^d$, $\lambda \in (0, \infty)$, $\xi: \Omega \rightarrow \mathbb{R}^m$ is a random vector on some fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $B(z), C(z) \in \mathbb{R}^{n \times d}$ for all $z \in \mathbb{R}^m$. As a norm, we consider the standard Euclidian norm on \mathbb{R}^d . Under the assumption that $B(z)$ is invertible for all $z \in \mathbb{R}^m$, we have $y = B(\xi)^{-1}C(\xi)x$ almost surely and the problem can be rewritten as

$$\begin{aligned} & \min_{x \in \mathbb{R}^d} \left\{ j(x) = \mathbb{E} \left(\frac{1}{2} \|A(\xi)x - b\|^2 + \frac{\lambda}{2} \|x\|^2 \right) \right\} \\ & \text{s. t. } A(\xi)x \leq c \quad \text{a. s.}, \end{aligned} \tag{P}$$

where $A(z) := B(z)^{-1}C(z) \in \mathbb{R}^{d \times d}$ for all $z \in \mathbb{R}^m$ and we assume that $\mathbb{E}\|A(\xi)x\|^2$ is finite for all $x \in \mathbb{R}^d$.

A key difficulty in solving the problem is the almost sure constraint. For one, it is not directly clear whether there even exists a feasible point. For this, one would atleast need $A(\xi)$ to have bounded support. Additionally, even if its nonempty, the feasible set is in general still very difficult to compute explicitly, due to its probabilistic nature. However, even simple situations can be problematic: Suppose for instance that $\xi \in \{1, \dots, M\}$ for some $M \in \mathbb{N}$. Then, problem (P) is equivalent to

$$\begin{aligned} & \min_{x \in \mathbb{R}^d} \left\{ j(x) = \mathbb{E} \left(\frac{1}{2} \|A(\xi)x - b\|^2 + \frac{\lambda}{2} \|x\|^2 \right) \right\} \\ & \text{s. t. } A(i)x \leq c \quad \forall i \in \{1, \dots, M\}. \end{aligned} \tag{P^M}$$

The projection itself is the solution to a quadratic program, which has a generic cost of $\mathcal{O}(n^3M^3)$ operations if the matrices $A(i)$ have no special structure (sparsity, for example) that can be exploited [2]. This projection can become expensive to compute if $n \times M$ is large.

In this work, we will consider a different approach by introducing a family of unconstrained optimization problems

$$\min_{x \in \mathbb{R}^d} \left\{ j^k(x) := \mathbb{E} \left(\frac{1}{2} \|y - b\|^2 + \frac{\lambda}{2} \|x\|^2 \right) + \frac{\gamma_k}{2} \pi(x) \right\}, \tag{P^k}$$

where $\gamma_k \in (0, \infty)$ for all $k \in \mathbb{N}$ and $\pi: \mathbb{R}^d \rightarrow \mathbb{R}$ has the properties $\pi(x) \geq 0$ and $\pi(x) = 0$ if and only if x

is feasible for (P) . If π is also convex and (P) has at least one feasible point, then there always exists a solution $x_k^* \in \mathbb{R}^d$ to problem (P^k) for all $k \in \mathbb{N}$. There are three immediate questions:

1. If x^* denotes the solution to (P) , when can we guarantee that $x^\gamma \rightarrow x^*$?
2. How do we choose π ?
3. How can we use this to numerically solve (P) ?

Outline.

Related literature.

Contributions. Single-loop penalty methods, Batch sizes, relaxed gradient bound, general treatment of penalty function, analysis of averaging, analysis of iterate moving averages.

2 Theory Background

In this chapter, we state some classic definitions and results that we will make use of in the later sections. Proofs are omitted, but can be found in the cited sources.

2.1 Convex Optimization

The contents of this section can be found in [2, 7]. Throughout, we let $\|\cdot\|$ denote the standard Euclidian norm and $\langle \cdot, \cdot \rangle$ the standard inner product on \mathbb{R}^d .

Definition 2.1. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable, and $L > 0$. We say that f is **Lipschitz-smooth with constant L** , or simply **L -smooth**, if its gradient is Lipschitz continuous, i. e. there exists a constant $L \in (0, \infty)$ such that

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

for all $x, y \in \mathbb{R}^d$.

Proposition 2.2. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be L -smooth. Then

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

for all $x, y \in \mathbb{R}^d$.

Proposition 2.3. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be L_f -smooth and let $g: \mathbb{R}^d \rightarrow \mathbb{R}$ be L_g -smooth and define $h(x) := af(x) + bg(x)$ for some positive constants $a, b \in (0, \infty)$. Then h is Lipschitz-smooth with constant $aL_f + bL_g$.

Definition 2.4. A set $C \subset \mathbb{R}^d$ is called **convex** if, for all $x, y \in C$ and $t \in [0, 1]$, it holds that $tx + (1 - t)y \in C$. We say that a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is **convex** if

$$f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y)$$

for all $x, y \in \mathbb{R}^d$ and all $t \in (0, 1)$. We say that f is **concave** if $-f$ is convex.

Proposition 2.5. Every convex function on \mathbb{R}^d is continuous.

Definition 2.6. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be convex. A vector $g \in \mathbb{R}^d$ is a **subgradient** of f at $x \in \mathbb{R}^d$ if

$$f(y) \geq f(x) + \langle g, y - x \rangle$$

for all $y \in \mathbb{R}^d$. The set of all subgradients of f at x is denoted by $\partial f(x)$ and we call this set the **subdifferential of f at x** . If $\partial f(x) \neq \emptyset$, then we call f **subdifferentiable at x** . If $\partial f(x) \neq \emptyset$ for all $x \in \mathbb{R}^d$, we call f **subdifferentiable**.

Proposition 2.7. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable. Then f is subdifferentiable with $\partial f(x) = \{\nabla f(x)\}$ for all $x \in \mathbb{R}^d$. In particular,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

for all $x, y \in \mathbb{R}^d$.

Proposition 2.8. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be subdifferentiable. Then

$$\langle g_y - g_x, y - x \rangle \geq 0$$

for all $x, y \in \mathbb{R}^d$ and $g_x \in \partial f(x)$, $g_y \in \partial f(y)$.

Definition 2.9. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\mu \in (0, \infty)$. We say that f is (μ) -strongly convex if

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \mu t(1-t)\|x - y\|^2$$

for all $x, y \in \mathbb{R}^d$ and $t \in (0, 1)$.

Clearly, strongly convex functions are convex.

Proposition 2.10. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and $g: \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and $\alpha > 0$. Also, let $A \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$. Then, the functions

1. $x \mapsto f(x) + g(x)$,
2. $x \mapsto \alpha f(x)$,
3. $x \mapsto f(Ax + b)$,

are all convex. If f is μ -strongly convex, then the above functions are also all μ -strongly convex.

Proposition 2.11. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ be convex and nondecreasing. Then the composition $f \circ g$ is also convex.

Proposition 2.12. Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly convex and subdifferentiable. Then

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

for all $x, y \in \mathbb{R}^d$ and $g \in \partial f(x)$. This implies that, for all $g_x \in \partial f(x)$, $g_y \in \partial f(y)$, we have

$$\langle g_y - g_x, x - y \rangle \geq \frac{\mu}{2}\|x - y\|^2,$$

for all $x, y \in \mathbb{R}^d$. In particular, if f is differentiable and $x^* = \arg \min_{x \in \mathbb{R}^d} f(x)$, we have

$$f(x^*) \leq f(x) - \frac{\mu}{2}\|x - x^*\|^2$$

for all $x \in \mathbb{R}^d$.

Proposition 2.13. Let $g: \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and let $\mu \in (0, \infty)$. Then, the function $x \mapsto g(x) + \frac{\mu}{2}\|x\|^2$ is μ -strongly convex.

Proposition 2.14. If $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is strongly convex and $C \subset \mathbb{R}^d$ is convex, then f admits a unique minimizer on C , i.e. there exists a point $x^* \in C$ such that $f(x^*) < f(x)$ for all $x \in C$.

Proposition 2.15. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be twice differentiable. If f has positive definite hessian, then f is convex. If, additionally, there exists some $\mu \in (0, \infty)$ such that $f''(x) - \mu I_d$ is positive definite for all $x \in \mathbb{R}^d$, where I_d denotes the $d \times d$ identity matrix, then f is μ -strongly convex.

Example 2.16. Examples of (strongly) convex functions include

- affine function;
- quadratic functions $f(x) := x^\top Ax + b^\top x + c$ for $x, b \in \mathbb{R}^d$, $c \in \mathbb{R}$, and $A \in \mathbb{R}^{d \times d}$ positive definite.
If $A - \mu I_d$ is positive definite for some $\mu \in (0, \infty)$, then f is μ -strongly convex.
- $x \mapsto \exp(x)$, $x \mapsto -\log(x)$, $x \mapsto \max(0, x)$.

2.2 Probability Theory

The contents of this section can be found in standard probability texts, for example [6, 1].

Definition 2.17. Let Ω be a set and let 2^Ω denote its power set. A subset $\mathcal{F} \subset 2^\Omega$ is called a **σ -algebra over Ω** if it satisfies the following three conditions:

1. $\emptyset \in \mathcal{F}$.
2. If $A, B \in \mathcal{F}$, then $B \setminus A \in \mathcal{F}$.
3. For any countable sequence $A_1, A_2, \dots \in \mathcal{F}$, we have $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

If \mathcal{F} is a σ -algebra over Ω , then the tuple (Ω, \mathcal{F}) is called a **measurable space**. For any subset $\mathcal{G} \subset 2^\Omega$, we define the **σ -algebra generated by \mathcal{G}** as the intersection over all σ -algebras that contain \mathcal{G} as an element, and we denote this σ -algebra by $\sigma(\mathcal{G})$.

Example 2.18. An important example of a σ -algebra over \mathbb{R}^d is the **Borel σ -algebra $\mathcal{B}(\mathbb{R}^d)$** , which is defined to be the σ -algebra generated by the subset of all open sets on \mathbb{R}^d . Functions that are $\mathcal{B}(\mathbb{R}^d)$ -measurable are called **Borel measurable**.

Definition 2.19. Let (Ω, \mathcal{F}) and (E, \mathcal{G}) be measurable spaces. A map $f: \Omega \rightarrow E$ is called \mathcal{F}, \mathcal{G} -measurable if $f^{-1}(G) := \{\omega \in \Omega \mid f(\omega) \in G\} \in \mathcal{F}$ for all $G \in \mathcal{G}$. We may say f is \mathcal{F} -measurable or simply measurable if one or both of the σ -algebras are either clear from the context or not relevant.

Definition 2.20. Let (Ω, \mathcal{F}) be a measurable space. A map $\mu: \mathcal{F} \rightarrow [0, \infty]$ is called a **measure on (Ω, \mathcal{F})** if it satisfies the following two conditions:

1. $\mu(\emptyset) = 0$.
2. For any countable sequence $A_1, A_2, \dots \in \mathcal{F}$, we have $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.

If μ is a measure on (Ω, \mathcal{F}) , then the triplet $(\Omega, \mathcal{F}, \mu)$ is called a **measure space**.

Definition 2.21. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f: \Omega \rightarrow \mathbb{R}$ be measurable. We define the **(μ) -integral** of f , denoted $\int f d\mu$, in three steps:

1. If $f(\omega) = \sum_{i=1}^n c_i 1_{A_i}(\omega)$ for some $n \in \mathbb{N}$, $c_1, \dots, c_n > 0$, and disjoint measurable sets $A_1, \dots, A_n \in \mathcal{F}$, then we define

$$\int f d\mu := \sum_{i=1}^n c_i \mu(A_i).$$

In this case f is called a **simple function**. The set of all simple functions on Ω is denoted by $\mathcal{S}(\Omega)$.

2. If f is nonnegative, i.e. $f(\omega) \geq 0$ of all $\omega \in \Omega$, then we define

$$\int f d\mu := \sup_{g \in \mathcal{S}(\Omega), g \leq f} \int g d\mu.$$

3. If f is neither a simple function, nor nonnegative, but $\int |f| d\mu < \infty$, then we define

$$\int f d\mu := \int \max(0, f) d\mu - \int \max(0, -f) d\mu.$$

Otherwise, we say that the (μ) -integral of f does not exist. If any of these three conditions apply to f , we say that f is (μ) -**integrable**.

Proposition 2.22. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f: \Omega \rightarrow \mathbb{R}$ and $g: \Omega \rightarrow \mathbb{R}$ be integrable. Then*

$$(i) \int f + g d\mu = \int f d\mu + \int g d\mu.$$

$$(ii) \int cf d\mu = c \int f d\mu \text{ for all } c \in \mathbb{R}.$$

(iii) If $f \leq g$, then $\int f d\mu \leq \int g d\mu$. If additionally $f < g$ on some set $A \in \mathcal{F}$ with $\mu(A) > 0$, then $\int f d\mu < \int g d\mu$.

Definition 2.23. Let (Ω, \mathcal{F}) be a measurable space. If $\mathbb{P}: \mathcal{F} \rightarrow [0, 1]$ is a measure on (Ω, \mathcal{F}) , we call \mathbb{P} a **probability measure** and we call the triple $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability space**. In this context, elements of \mathcal{F} are called **events**.

Definition 2.24. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $A \in \mathcal{F}$ is said to hold **almost surely** (a.s. for short) if $\mathbb{P}(A) = 1$.

Definition 2.25. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and (E, \mathcal{G}) a measurable space. A map $X: \Omega \rightarrow E$ is called a **random variable** on $(\Omega, \mathcal{F}, \mathbb{P})$ if X is \mathcal{F}, \mathcal{G} -measurable. In the case $E = \mathbb{R}^d$, we may call X a **random vector**. Further, we define the notation $\mathbb{P}(X \in G) := \mathbb{P}(X^{-1}(G))$. We define the **distribution of X** to be the probability measure $\mathbb{P}^X := \mathbb{P} \circ X^{-1}$ on (E, \mathcal{G}) . Finally, we define $\sigma(X) := \sigma(\{X^{-1}(G), G \in \mathcal{G}\})$ and call this the **σ -algebra generated by X** .

Definition 2.26. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are called **independent** if $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$ for all $A, B \in \mathcal{F}$. X and Y are called **identically distributed** if $\mathbb{P}^X = \mathbb{P}^Y$. We use the abbreviation **i. i. d.** as shorthand for “independent and identically distributed”.

Definition 2.27. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a random variable on this space. If X is integrable, we define the **expected value of X** , denoted by $\mathbb{E}(X)$, as $\mathbb{E}(X) := \int X d\mathbb{P}$.

The following three properties will be used multiple times throughout this text without explicit mention. They follow directly from [proposition 2.22](#).

Proposition 2.28. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ and $Y: \Omega \rightarrow \mathbb{R}$ be integrable random variables. Then*

$$(i) \mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

$$(ii) \mathbb{E}(cX) = c\mathbb{E}(X) \text{ for all } c \in \mathbb{R}.$$

(iii) If $X \leq Y$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$. If additionally $X(\omega) < Y(\omega)$ for all ω in an event $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$, then $\mathbb{E}(X) < \mathbb{E}(Y)$.

Proposition 2.29. *Let $f: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$ be convex in its first argument, i.e. $x \mapsto f(x, \omega)$ is convex for all $\omega \in \Omega$. Then, the function $x \mapsto \mathbb{E}(f(x, \cdot))$ is convex.*

Proposition 2.30 (Jensen's inequality). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. Then we have*

$$\mathbb{E}(X^2) \geq \mathbb{E}(X)^2.$$

In particular: If X^2 is integrable, then X must also be integrable.

Definition 2.31. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. If X^2 is integrable, we define the **variance of X** , denoted by $\text{Var}(X)$, as $\text{Var}(X) := \mathbb{E}\|X - \mathbb{E}(X)\|^2$.

One fact from probability theory is that, for any integrable random variable $X: \Omega \rightarrow E$, it holds that $\int X d\mathbb{P} = \int I d\mathbb{P}^X$, where $I: E \rightarrow E$ is the identity operator. It is now easy to see that if two random variables X and Y are identically distributed, they have the same expected value and variance.

Proposition 2.32. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, \dots, X_n: \Omega \rightarrow \mathbb{R}$ be independent random variables, such that X_i^2 is integrable for all $i \in \{1, \dots, n\}$. Then $\text{Var}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$, for any $a_1, \dots, a_n \in \mathbb{R}$. If, additionally, they are all identically distributed, it holds that $\text{Var}(1/n \sum_{i=1}^n X_i) = \text{Var}(X_1)/n$.*

Proposition 2.33. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable such that X^2 is integrable. Then $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.*

Definition 2.34. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}^n$ be a random variable. Further, let $\mathcal{C} \subset \mathcal{F}$ be a σ -algebra. We call a random variable $Z: \Omega \rightarrow \mathbb{R}^n$ a **conditional expectation of X given \mathcal{C}** , if

1. Z is \mathcal{C} -measurable, and
2. for all $C \in \mathcal{C}$, it holds that

$$\int_C Z d\mathbb{P} = \int_C X d\mathbb{P}.$$

If Z is a conditional expectation of X given \mathcal{C} then we use the notation $\mathbb{E}(X | \mathcal{C}) := Z$

If $Y: \Omega \rightarrow \mathbb{R}^m$ is a random variable, such that $\mathcal{C} = \sigma(Y)$, then we use the notation $\mathbb{E}(X | Y) := \mathbb{E}(X | \sigma(Y))$. In that case, we call $\mathbb{E}(X | Y)$ the **conditional expectation of X given Y** . Further, for $\omega \in \Omega$ with $Y(\omega) = y \in \mathbb{R}^m$, the **conditional expectation of X given $Y = y$** , denoted by $\mathbb{E}(X | Y = y)$, is defined as $\mathbb{E}(X | Y = y) := \mathbb{E}(X | Y)(\omega)$.

Note that $\mathbb{E}(X | Y)$ is not unique. However, if Z_1 and Z_2 are both conditional expectations of X given Y , then we always have $Z_1 = Z_2$ almost surely. For simplicity, we will keep the “almost surely” implicit.

Remark 2.35. If X and Y are integrable random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the conditional expectation of X given $Y = y$ can be thought of as the expected value of X on a different probability space $(\Omega, \mathcal{F}, \mathbb{P}_{Y=y})$, where $\mathbb{P}_{Y=y}(A) := \mathbb{P}(A | Y = y)$. More precisely, $\mathbb{E}(X | Y = y) = \int X d\mathbb{P}_{Y=y}$. It follows that $\mathbb{E}(X | Y = \cdot)$ inherits all basic properties of $\mathbb{E}(\cdot)$.

Below, we state some special properties of conditional expectations.

Proposition 2.36 (Properties of $\mathbb{E}(X | Y)$). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}^n$, $Y: \Omega \rightarrow \mathbb{R}^m$ be integrable random variables. Further, let $F: \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$ such that $\omega \mapsto F(x, \omega)$ is \mathcal{F} -measurable for all $x \in \mathbb{R}^d$ and let $G: \mathbb{R}^n \rightarrow \mathbb{R}$ be Borel measurable. Then,*

- (i) $\mathbb{E}(\mathbb{E}(X | Y)) = \mathbb{E}(X)$.
- (ii) $\mathbb{E}(X | X) = X$.
- (iii) if $F(x, \cdot) \leq G(x)$ a.s. for all $x \in \mathbb{R}^d$, it follows that $\mathbb{E}(F(X, \cdot) | X) \leq G(X)$.

Definition 2.37. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}^n$, $Y: \Omega \rightarrow \mathbb{R}^m$ be integrable random variables. The **conditional variance of X given Y** , denoted by $\text{Var}(X | Y)$, is defined as

$$\text{Var}(X | Y) := \mathbb{E}\left(\|X - \mathbb{E}(X | Y)\|^2 \mid Y\right).$$

For $y \in \mathbb{R}^d$, the **conditional variance of X given $Y = y$** is defined as

$$\text{Var}(X | Y = y) := \mathbb{E}\left(\|X - \mathbb{E}(X | Y = y)\|^2 \mid Y = y\right).$$

It holds that $\text{Var}(X | Y = Y(\omega)) = \text{Var}(X | Y)(\omega)$ for all $\omega \in \Omega$.

Remark 2.38. Let X and Y be integrable random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Similarly to the conditional expectation of X given $Y = y$, the conditional variance $\text{Var}(X | Y = y)$ can be thought of as the variance of X on a different probability space $(\Omega, \mathcal{F}, \mathbb{P}_{Y=y})$ (see [remark 2.35](#)). Hence, all basic properties of $\text{Var}(\cdot)$ are inherited by the conditional variance. In particular, if X_1, \dots, X_n are i. i. d. random variables, it holds that $\text{Var}(X_1 + \dots + X_n | Y) = 1/n \text{Var}(X_1 | Y)$ – a fact that will be used later.

Proposition 2.39. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}^n$ be an integrable random variable. Further, let $F: \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$ such that $\omega \mapsto F(x, \omega)$ is \mathcal{F} -measurable for all $x \in \mathbb{R}^d$ and $\mathbb{E}(F(X, \cdot)^2) < \infty$, and let $G: \mathbb{R}^n \rightarrow \mathbb{R}$ be Borel measurable. Then, if $\text{Var}(F(x, \cdot)) \leq G(x)$ for all $x \in \mathbb{R}^d$, it also holds that

$$\text{Var}(F(X, \cdot) | X) \leq G(X).$$

2.3 Stochastic Optimization

Definition 2.40 ([5]). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function. A random vector $G: \Omega \rightarrow \mathbb{R}^d$ is called a **stochastic subgradient of f** at a point $x \in \mathbb{R}^d$ if $\mathbb{E}(G) \in \partial f(x)$, or equivalently

$$f(y) \geq f(x) + \langle \mathbb{E}(G), y - x \rangle$$

for all $y \in \mathbb{R}^d$. If, additionally, f is differentiable at x , we may simply refer to G as a **stochastic gradient**.

Example 2.41. Let $F: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$ be continuously differentiable in its first argument and let $f(x) := \mathbb{E}(F(x, \cdot))$ for all $x \in \mathbb{R}^d$. Then, for any $x \in \mathbb{R}^d$, the random vector $G_x: \Omega \rightarrow \mathbb{R}^d$, defined by $G_x(\omega) := \nabla_x F(x, \omega)$, is a stochastic gradient of f at x .

Definition 2.42. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Further, let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and $g: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}^m$. Consider a stochastic optimization problem $\min_{x \in \mathbb{R}^d} f(x)$ subject to the constraints $g(x, \omega) \leq 0$ almost surely. We define the **active set of $x \in \mathbb{R}^d$ for scenario $\omega \in \Omega$** , as the set

$$\mathcal{A}(x, \omega) := \{i \in \{1, \dots, m\}, g_i(x, \omega) = 0\},$$

where $g_i(x, \omega)$ is the i th component of $g(x, \omega)$, for $i \in \{1, \dots, m\}$.

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a **convex stochastic optimization problem** has the form

$$\min_{x \in \mathcal{X}} \left\{ f(x) := \mathbb{E}(F(x, \xi)) \right\},$$

where $\emptyset \neq \mathcal{X} \subset \mathbb{R}^d$, $\xi: \Omega \rightarrow \mathbb{R}^m$ is a random vector, and $F: \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a function that satisfies

- $x \mapsto F(x, \xi(\omega))$ is convex for almost every $\omega \in \Omega$;

- $\omega \rightarrow F(x, \xi(\omega))$ is \mathbb{P} -measurable for all $x \in \mathbb{R}^d$.

With these assumptions, the problem is well-defined and f is convex. We say that a problem of the above form is **unconstrained** if $\mathcal{X} = \mathbb{R}^d$. In that case, if additionally f is subdifferentiable, a standard method to solve such a problem is *stochastic gradient descent*.

Algorithm 1 (Stochastic Gradient Descent (SGD)). Let $x_1 \in \mathbb{R}^d$. For $k \in \mathbb{N}$, let $\tau_k \in (0, \infty)$ be a parameter, called **step size**. The **Stochastic Gradient Descent (SGD)** iterates $(x_k)_{k \in \mathbb{N}}$ are defined by

$$x_{k+1} := x_k - \tau_k G^k(x_k),$$

where $G^k(x_k)$ is a stochastic subgradient of f at x_k .

The idea and analysis of this method go back to Robbins and Monro [10]. The convergence of the iterates $(x_k)_{k \in \mathbb{N}}$, generated by [algorithm 1](#), to a minimum $x^* \in \arg \min_{x \in \mathbb{R}^d} f(x)$ (if it exists) depends heavily on the choice of step sizes $(\tau_k)_{k \in \mathbb{N}}$ and the behavior of $\mathbb{E}\|G^k(x_k)\|^2$. In case of a *strongly convex objective* f , if there exists a constant $M^2 \in (0, \infty)$ such that $\sup_{k \in \mathbb{N}} \mathbb{E}\|G^k(x_k)\|^2 \leq M^2$, the conditions

$$\sum_{k=1}^{\infty} \tau_k = \infty, \quad \sum_{k=1}^{\infty} \tau_k^2 < \infty,$$

ensure that [algorithm 1](#) converges to the minimum of f .

Proposition 2.43. *The iterates of [algorithm 1](#) satisfy*

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^d} \left\{ f(x_k) + \langle g^k(x_k), x - x_k \rangle + \frac{1}{2\tau_k} \|x - x_k\|^2 \right\}$$

for all $k \in \mathbb{N}$.

Note that the function $x \mapsto f(x_k) + \langle G^k(x_k), x - x_k \rangle + 1/2\tau_k \|x - x_k\|^2$ is strongly convex, since it is the compositon of the quadratic $x \mapsto f(x_k) + \langle G^k(x_k), x \rangle + (1/2\tau_k)x^\top x$ and the affine function $x \mapsto x - x_k$, (see [proposition 2.10](#) and [example 2.16](#)), and thus the above minimization problem is well defined ([proposition 2.14](#)).

In most applications, SGD is used with *iterate averaging*. In [9], it was shown that the averages $\bar{x}_k := \sum_{i=1}^k x_i$ efficiently converge in the case of convex objectives. Another popular averaging scheme uses *iterate moving averages*, where one considers the moving average $\hat{x}_{k+1} := (1 - \hat{\rho}_k)\hat{x}_k + \hat{\rho}_k x_{k+1}$, where $\hat{\rho}_k \in [0, 1]$ and $(x_k)_{k \in \mathbb{N}}$ still denote the standard SGD iterates. We will analyze the iterate average \bar{x}_k in [section 3.2.4](#) and the iterate moving average \hat{x}_k in [section 3.3](#).

2.4 Norms and Inequalities

3 Sequential Penalty Methods

Throughout this chapter, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All maps $\mathbb{R}^n \rightarrow \mathbb{R}^m$, $n, m \in \mathbb{N}$, are implicitly considered to be measurable with respect to the corresponding Borel σ -algebras on \mathbb{R}^n and \mathbb{R}^m (example 2.18). We consider a more general form of problem (P):

$$\begin{aligned} & \min_{x \in \mathbb{R}^d} f(x) \\ \text{s. t. } & A(\xi)x \leq c \quad \text{a.s.}, \end{aligned} \tag{Q}$$

where $f: \mathbb{R}^d \rightarrow \mathbb{R}$, $A(y) \in \mathbb{R}^{n \times d}$ for $y \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and $\xi: \Omega \rightarrow \mathbb{R}^m$ is a random variable. We also define the corresponding unconstrained problems

$$\min_{x \in \mathbb{R}^d} \left\{ f^k(x) := f(x) + \frac{\gamma_k}{2} \pi(x) \right\}, \tag{Q^k}$$

where $\pi: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\gamma_k \in (0, \infty)$ for all $k \in \mathbb{N}$. We state the following assumptions to refer back to.

Assumption 1. The objective f is μ -strongly convex for some $\mu > 0$.

Assumption 2. The penalty function π is convex and satisfies $\pi(x) \geq 0$ for all x and $\pi(x) = 0$ if and only if $A(\xi)x \leq c$ almost surely.

Assumption 3. The sequence of penalty parameters $(\gamma_k)_{k \in \mathbb{N}}$ is strictly increasing, unbounded, and satisfies $\gamma_k \in (0, \infty)$ for all $k \in \mathbb{N}$.

Assumption 4. There exists at least one feasible point for (Q).

3.1 Consistency of Solutions

Theorem 3.1. *In the situation of (Q) and (Q^k), assume that assumptions 1 to 4 hold. Let x_k^* be the solution to (Q^k) and x^* the solution to (Q) (in particular, these solutions exist and are unique). Then $x_k^* \rightarrow x^*$ and $f^k(x_k^*) \rightarrow f(x^*)$ as $k \rightarrow \infty$.*

First, some preparation.

Lemma 3.2. *If $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is radially unbounded, continuous, and $X \subset \mathbb{R}^d$ is nonempty and closed, then f attains a minimum over X , i.e. there exists $x^* \in X$ such that $f(x^*) = \inf_{x \in X} f(x)$.*

Proof. Let $x_0 \in X$. Since f is radially unbounded, there exists $r > 0$ such that $f(x) \geq f(x_0)$ for all $x \in \mathbb{R}^d$ with $\|x\| > r$, therefore any minimum of f – if it exists – must be contained in the closed ball of radius r around 0, which we denote by B_r . In particular, for $C := X \cap B_r$ we have

$$\inf_{x \in X} f(x) = \inf_{x \in C} f(x).$$

By continuity of f , its domain must be a closed set, which implies that C is compact. Assume now that f does not attain a minimum on C . Then there must exist a sequence $(x_k)_{k \in \mathbb{N}} \subset C$ such that $\lim_{k \rightarrow \infty} f(x_k) = \inf_{x \in C} f(x)$. Continuous functions map compact sets to compact sets, hence $\inf_{x \in C} f(x) \in f(C)$ and thus there must exist some $x^* \in C$ such that $f(x^*) = \inf_{x \in C} f(x)$. \square

Lemma 3.3. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be radially unbounded and assume that $\{f(u_k), k \in \mathbb{N}\}$ is bounded for some sequence $(u_k)_{k \in \mathbb{N}} \subset \mathbb{R}^d$. Then, $(u_k)_{k \in \mathbb{N}}$ is a bounded sequence.

Proof. Assume $(u_k)_{k \in \mathbb{N}}$ is not bounded. Then there must exist some subsequence $(u_{k_r})_{r \in \mathbb{N}}$ such that $\|u_{k_r}\| \rightarrow \infty$ for $r \rightarrow \infty$. However, f is radially unbounded, which implies $f(u_{k_r}) \rightarrow \infty$ for $r \rightarrow \infty$, which contradicts our assumption that $(f(u_{k_r}))_{r \in \mathbb{N}}$ is bounded. Hence $(u_{k_r})_{r \in \mathbb{N}}$ must be bounded. \square

Lemma 3.4. Let $U := \{u_k, k \in \mathbb{N}\}$ be a subset of \mathbb{R}^d . Suppose that any subsequence of U contains a subsequence that converges to $u \in \mathbb{R}^d$. Then $u_k \rightarrow u$ for $k \rightarrow \infty$.

Proof. Assume that $u_k \not\rightarrow u$. Then there must exist some $\epsilon > 0$ and a sequence of natural numbers $k_1 < k_2 < \dots$ such that

$$\|u_{k_r} - u\| \geq \epsilon \quad (3.1)$$

for all $r \in \mathbb{N}$. However, as a subsequence of U , the sequence $(u_{k_r})_{r \in \mathbb{N}}$ must simultaneously contain a subsequence that converges to u , which contradicts (3.1). Thus, our assumption $u_k \not\rightarrow u$ must be false. \square

We will now prove the main theorem.

Proof of Theorem 3.1. From strong convexity of f ([assumption 1](#)), convexity of π ([assumption 2](#)) and $\gamma_k > 0$ ([assumption 3](#)), it follows that f^k is also strongly convex for all $k \in \mathbb{N}$ ([proposition 2.10](#)). Thus, for every $k \in \mathbb{N}$, [proposition 2.14](#) implies that there exists a unique solution x_k^* to problem (Q^k) . Let x be any feasible point for (Q) , which exists by [assumption 4](#). Then, for any $k \in \mathbb{N}$,

$$f(x_k^*) \leq f^k(x_k^*) \leq f^k(x) = f(x). \quad (3.2)$$

In particular, $(f^k(x_k^*))_{k \in \mathbb{N}}$ is a bounded sequence. Since f is radially unbounded ([Make lemma.](#)) and π is nonnegative ([assumption 2](#)), f^k must also be radially unbounded. It follows, by [lemma 3.3](#), that the sequence $(x_k^*)_{k \in \mathbb{N}}$ is also bounded and thus it contains a subsequence $(x_{k_r}^*)_{r \in \mathbb{N}}$ that converges to a point $x_\infty^* \in \mathbb{R}^d$. For any $k \in \mathbb{N}$, we have

$$\begin{aligned} f^{k+1}(x_{k+1}^*) - f^k(x_k^*) &\geq f^{k+1}(x_{k+1}^*) - f^k(x_{k+1}^*) \\ &= \frac{\gamma_{k+1}}{2} \pi(x_{k+1}^*) - \frac{\gamma_k}{2} \pi(x_{k+1}^*) \\ &= \frac{\gamma_{k+1} - \gamma_k}{2} \pi(x_{k+1}^*) \\ &\geq 0, \end{aligned}$$

where we used [assumptions 2](#) and [3](#) in the last step. This implies that $(f^k(x_k^*))_{k \in \mathbb{N}}$ is a monotonically increasing sequence. We know from (3.2) that $(f^k(x_k^*))_{k \in \mathbb{N}}$ must also be bounded and thus $(f^k(x_k^*))_{k \in \mathbb{N}}$ must converge. In particular, we have

$$\limsup_{k \rightarrow \infty} [f^k(x_k^*) - f(x_k^*)] < \infty.$$

By plugging in definitions for f^k and f , we get

$$\limsup_{k \rightarrow \infty} \frac{\gamma_k}{2} \pi(x_k^*) < \infty.$$

The function π is convex ([assumption 2](#)) and thus continuous ([proposition 2.5](#)). Hence, since $\lim_{k \rightarrow \infty} \gamma_k = \infty$ ([assumption 3](#)), the above limit can be finite only if

$$\pi(x_\infty^*) = \lim_{r \rightarrow \infty} \pi(x_{k_r}^*) = 0,$$

which implies that x_∞^* is feasible, by [assumption 2](#). To prove optimality of x_∞^* , let x^* be the solution to [\(Q\)](#). Then we have, again from [\(3.2\)](#),

$$f(x_\infty^*) = \lim_{r \rightarrow \infty} f(x_{k_r}^*) \leq \lim_{r \rightarrow \infty} f^{k_r}(x_{k_r}^*) = \lim_{k \rightarrow \infty} f^k(x_k^*) \leq f(x^*),$$

which implies $f(x_\infty^*) = f(x^*)$ by feasibility of x_∞^* and optimality of x^* . This in turn implies that all inequalities must, in fact, be equalities and thus

$$\lim_{k \rightarrow \infty} f^k(x_k^*) = f(x^*),$$

as desired. Finally, by uniqueness of x^* we must have $x_\infty^* = x^*$, proving that x^* is a limit point of $(x_k^*)_{k \in \mathbb{N}}$. Note that [assumption 3](#) still holds if we replace $(\gamma_k)_{k \in \mathbb{N}}$ by any subsequence $(\gamma_{k_l})_{l \in \mathbb{N}}$ and the same arguments imply that x^* is also a limit point of $(\gamma_{k_l})_{l \in \mathbb{N}}$. Hence, by [lemma 3.4](#), we in fact have $\lim_{k \rightarrow \infty} x_k^* = x^*$. \square

Assumption 5. The random matrix $A(\xi)$ has finite second moment, which means $\mathbb{E}\|A(\xi)\|_F^2 < \infty$, where $\|M\|_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^n M_{ij}^2}$ for $M \in \mathbb{R}^{n \times n}$.

Lemma 3.5. *In the situation of [\(P\)](#), assume that [assumption 5](#) holds. Then, j satisfies [assumption 1](#).*

Proof. It holds that

$$\begin{aligned} j(x) &\stackrel{\text{def}}{=} \mathbb{E} \left(\frac{1}{2} \|A(\xi)x - b\|^2 \right) + \frac{\lambda}{2} \|x\|^2 \\ &\leq \mathbb{E}(\|A(\xi)\|^2) \|x\|^2 + \|b\|^2 + \frac{\lambda}{2} \|x\|^2 \longrightarrow \infty, \quad \|x\| \rightarrow \infty, \end{aligned}$$

thus j is radially unbounded. Further, for all $x \in \mathbb{R}^d$,

$$j''(x) = \mathbb{E}(A(\xi)^\top A(\xi)) + \lambda I_d,$$

where I_n is the $d \times d$ -identity matrix. By definition,

$$(A(\xi)^\top A(\xi))_{ij} = \sum_{k=1}^n A(\xi)_{ki} A(\xi)_{kj}.$$

[Assumption 5](#) implies that $\mathbb{E}(A(\xi)_{ij}^2) < \infty$ for all $i, j \in \{1, \dots, d\}$. Hence, by Cauchy-Schwarz,

$$\mathbb{E}|(A(\xi)^\top A(\xi))_{ij}| \leq \sum_{k=1}^n \mathbb{E}|A(\xi)_{ki} A(\xi)_{kj}| \leq \sum_{k=1}^n \mathbb{E}(A(\xi)_{ki}^2) \mathbb{E}(A(\xi)_{kj}^2) < \infty.$$

Thus, for all $x \in \mathbb{R}^d$, $j''(x)$ exists and, since $\lambda > 0$, $j''(x) - \lambda I$ is positive definite. By [proposition 2.15](#), it follows that j is (λ) -strongly convex. \square

Applying this to our problem of interest, we have the following useful result for determining reasonable penalty functions π .

Corollary 3.6. *In the situation of [\(P\)](#) and [\(P^k\)](#), assume that [assumptions 2 to 5](#) hold. Then there exists a unique solution x^* to [\(P\)](#) and, for all $k \in \mathbb{N}$, there exists a unique solution x_k^* to [\(P^k\)](#). These solutions satisfy $\lim_{k \rightarrow \infty} x_k^* = x^*$, $\lim_{k \rightarrow \infty} j^k(x_k^*) = j(x^*)$.*

Proof. By [lemma 3.5](#), j satisfies [assumption 1](#). The claim now follows from [theorem 3.1](#). \square

3.2 Sequential SGD

We will now analyze a form of stochastic gradient descent to efficiently solve (Q).

Algorithm 2. For $k \in \mathbb{N}$, let $x_1 \in \mathbb{R}^d$, $\tau_k, \gamma_k \in (0, \infty)$ and $b_k \in \mathbb{N}$. The **Sequential SGD (SSGD)** iterates have the form

$$x_{k+1} := x_k - \tau_k \tilde{G}^k(x_k),$$

where

$$\tilde{G}^k(x) := \frac{1}{b_k} \sum_{j=1}^{b_k} G^k(x, \xi_k^j),$$

$(\xi_i^j)_{i=1, \dots, k, j=1, \dots, b_k}$ are i. i. d. samples from the distribution of ξ and $G^k(x, \xi)$ is a stochastic subgradient of f^k at x . We refer to τ_k as a **step size**, γ_k as a **penalty parameter** and b_k as a **batch size**.

Let x^* be the solution to (Q). Our goal is to determine appropriate parameters τ_k, γ_k and $b_k \in \mathbb{N}$, such $\mathbb{E}\|x_k - x^*\|$ converges to zero as fast as possible. There are several difficulties here. One is that we do not use gradients from our main objective in (Q), but from the surrogate objective (Q^k). In addition, this surrogate depends on γ_k , which may need to satisfy $\gamma_k \rightarrow \infty$ – that is, the surrogate objective changes between iterations. Further, the squared gradient norm $\mathbb{E}\|G^k(x, \xi)\|^2$ grows quadratically in γ_k and $\|x\|$, which goes against standard assumptions in the literature. These difficulties prevent us from being able to directly apply standard analysis techniques like the ones found in [9], for example. Because of this, we will first decompose $\mathbb{E}\|x_k - x^*\|$ as follows:

$$\mathbb{E}\|x_k - x^*\| \leq \mathbb{E}\|x_k - x_k^*\| + \|x_k^* - x^*\|, \quad (3.3)$$

where we used the triangle inequality. In the following sections, we will derive bounds for the two terms on the right-hand side and use those bounds to determine appropriate sequences $(\tau_k)_{k \in \mathbb{N}}$, $(\gamma_k)_{k \in \mathbb{N}}$ and $(b_k)_{k \in \mathbb{N}}$ to guarantee convergence of the algorithm. In the following, all statements involving random variables are understood to hold almost surely, unless otherwise stated. Looking at (3.3), there are two terms we need to bound: $\mathbb{E}\|x_k - x_k^*\|$ and $\|x_k^* - x^*\|$. We refer to the former as the **tracking error** and the latter as the **surrogate error**. We will refer to the following additional assumptions.

Assumption 6. We can sample arbitrarily many independent random variables $\xi_i^j, i, j \in \mathbb{N}$, from the distribution of ξ .

Assumption 7. The objective f is L -smooth.

Assumption 8. The objective f and the penalty π are differentiable and have Lipschitz gradients. More specifically, there exist $L, L_\pi \in (0, \infty)$ such that, for all $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ and $\|\nabla \pi(x) - \nabla \pi(y)\| \leq L_\pi\|x - y\|$.

Assumption 9. If f^k is subdifferentiable at a point $x \in \mathbb{R}^d$ for some $k \in \mathbb{N}$, let $g^k(x) \in \partial f(x)$ and let $G^k(x, \xi)$ be a stochastic gradient of f^k at x . Then there exists a constant $C > 0$, such that

$$\text{Var}(G^k(x, \xi)) \leq C (\|x\|^2 + \|x\|^2 \gamma_k^2 + \gamma_k^2 + 1).$$

Assumption 10. There exists a $K \in \mathbb{N}$ such that, for all $k \geq K$, the active sets $\mathcal{A}(x_k^*, \omega)$ are almost surely identical to $\mathcal{A}(x^*, \omega)$, i. e. $\mathcal{A}(x_k^*, \omega) = \mathcal{A}(x^*, \omega)$ for all $k \geq K$ and almost every $\omega \in \Omega$.

Lemma 3.7. Let [assumption 5](#) hold. Then, the objective j and the penalty π in (P^k) are both Lipschitz smooth. More precisely, j is $\mathbb{E}\|A^\top(\xi)A(\xi)\|_F$ -smooth and π is $\mathbb{E}\|A^\top(\xi)A(\xi)\|_F$ -smooth.

Proof. We will first show that any quadratic is Lipschitz-smooth. Let $f(x) := 1/2 x^\top A x + b^\top x + c$ for $x \in \mathbb{R}^d$, matrix $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$, and $c \in \mathbb{R}$. Differentiating f , we get

$$\nabla f(x) = A^\top (Ax - b),$$

hence, for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| = \|A^\top A(x - y)\| \leq \|A^\top A\|_F \|x - y\|. \quad (3.4)$$

Thus f is $\|A^\top A\|_F$ -smooth. Note that we can write j as

$$j(x) = \frac{1}{2} x^\top \mathbb{E}(A^\top(\xi)A(\xi) + \lambda I_d)x - \langle \mathbb{E}(A(\xi))x, b \rangle + \frac{1}{2} b^\top b,$$

so, by (3.4), the fact that $\|\mathbb{E}(A^\top(\xi)A(\xi)) + \lambda I_d\|_F \leq \mathbb{E}\|A^\top(\xi)A(\xi) + \lambda I_d\|_F$, and [assumption 5](#), it follows that j is $\mathbb{E}\|A^\top(\xi)A(\xi) + \lambda I_d\|_F$ -smooth.

Next, we consider π . Let $g(t) := \max(0, t)$ for $t \in \mathbb{R}$. Clearly, for $t, s \in (0, \infty)$, we have $|g(t) - g(s)| = |t - s|$. If $t \geq 0$ and $s \leq 0$, we have $|g(t) - g(s)| = t \leq t - s = |t - s|$. By symmetry, the same holds if $t \leq 0$ and $s \geq 0$. We conclude that $|g(t) - g(s)| \leq |t - s|$ for all $t, s \in \mathbb{R}$, making g 1-Lipschitz continuous. Now,

$$\nabla \pi(x) = 2\mathbb{E}(A^\top(\xi)(A(\xi)x - c)_+),$$

where $(x)_+$ applies g element-wise to all components of $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$. We thus have

$$\begin{aligned} \|\nabla \pi(x) - \nabla \pi(y)\| &= 2\|\mathbb{E}(A^\top(\xi)(A(\xi)x - c)_+ - A^\top(\xi)(A(\xi)y - c)_+)\| \\ &\leq \mathbb{E}\left\| (A^\top(\xi)A(\xi)x - A^\top(\xi)c)_+ - (A^\top(\xi)A(\xi)y - A^\top(\xi)c)_+ \right\| \\ &\leq \mathbb{E}\|A^\top(\xi)A(\xi)(x - y)\| \\ &\leq \mathbb{E}\|A^\top(\xi)A(\xi)\|_F \|x - y\|, \end{aligned}$$

as desired. \square

3.2.1 Bounding the surrogate error

In this section, we will restrict ourselves to special cases for π , in order to bound the surrogate error $\|x_k^* - x^*\|$.

Theorem 3.8. *In the situations of (Q) and (Q^k), let $\pi(x) := \mathbb{E}\|(0, A(\xi)x - c)_+\|^2$ and assume that [assumptions 1, 3 to 5, 7 and 10](#) hold. Then there exists a unique solution x^* to (Q) and, for all $k \in \mathbb{N}$, there exists a unique solution x_k^* to (Q^k). Further, we have*

$$\|x_k^* - x^*\| = \mathcal{O}(\gamma_k^{-1}).$$

The following statement will be used multiple times, so we state it here, before proving [theorem 3.8](#).

Lemma 3.9. *Let $A(\xi)$ satisfy [assumption 5](#). Then, the function $\pi(x) := \mathbb{E}\|(0, A(\xi)x - c)_+\|^2$, $x \in \mathbb{R}^d$, satisfies [assumption 2](#).*

Proof. The function $r(t) := \max(0, t)$ is convex and $t \mapsto t^2$ is convex and nondecreasing on \mathbb{R} ([example 2.16](#)). Hence, their composition, $h(t) := \max(0, t)^2$, is also convex ([proposition 2.11](#)). Now, for

$x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, we interpret $r(x)$ as the vector $(r(x_1), \dots, r(x_d))^\top$. Then, the function

$$\phi(x) := \|r(x)\|^2 = \sum_{i=1}^d h(x_i)$$

is also convex, since it is the sum of convex functions ([proposition 2.10](#)). Thus, the function $\psi: \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$, defined by $\psi(x, y) := \phi(A(y)x - c)$, is convex in its first argument, by [proposition 2.10](#). Finally, by [proposition 2.29](#), $\pi(x) = \mathbb{E}(\psi(x, \xi))$ is convex. Clearly, $\pi(x) \geq 0$ for all $x \in \mathbb{R}^d$ and $\pi(x) = 0$ if and only if $x \in \mathbb{R}^d$ is feasible for [\(P\)](#), so π satisfies [assumption 2](#). \square

Proof of Theorem 3.8. [Assumption 5](#) ensures that π is well-defined. Existence and uniqueness of x^* and x_k^* for all $k \in \mathbb{N}$ follows from [assumptions 1](#) and [4](#), [lemma 3.9](#), and [propositions 2.10](#) and [2.14](#). By [assumption 10](#), there exists $K \in \mathbb{N}$ such that $\mathcal{A}(x_k^*, \omega) = \mathcal{A}(x^*, \omega)$ for all $k \geq K$ and almost every $\omega \in \Omega$. This implies that the gradient of π ,

$$\nabla \pi(x) = 2\mathbb{E}(A(\xi)^\top \max(0, A(\xi)x - c)),$$

is affine on the set $\{x_k^*, k \geq K\}$. Specifically, if $\mathcal{A}(x^*) = \{i_1, \dots, i_r\} \subset \{1, \dots, d\}$ for some $r \in \{1, \dots, d\}$, then, for all $k \geq K$,

$$\nabla \pi(x_k^*) = 2\mathbb{E}(A(\xi)^\top P \cdot (A(\xi)x_k^* - c)) = 2\mathbb{E}(A(\xi)^\top PA(\xi))x_k^* - 2\mathbb{E}(A(\xi)^\top)Pc, \quad (3.5)$$

where $P = (p_{ij})_{i,j \in \{1, \dots, d\}} \in \mathbb{R}^{d \times d}$, such that $p_{ii} = 1$ if $i \notin \mathcal{A}(x_K^*)$ and $p_{ij} = 0$, otherwise. Let $k \geq K$. Since $\nabla \pi(x^*) = 0$, we have

$$\|\nabla \pi(x_k^*)\| = \|\nabla \pi(x_k^*) - \nabla \pi(x^*)\| = 2\|\mathbb{E}(A(\xi)^\top PA(\xi))\| \|x_k^* - x^*\|.$$

By optimality of x_k^* for f^k , it holds that

$$0 = \nabla f^k(x_k^*) = \nabla f(x_k^*) + \frac{\gamma_k}{2}\pi(x_k^*),$$

which implies

$$\|\pi(x_k^*)\| \leq \frac{2}{\gamma_k} \|\nabla f(x_k^*)\| \leq \frac{2\|\nabla f(x_k^*) - \nabla f(x^*)\|}{\gamma_k} + \frac{2\|\nabla f(x^*)\|}{\gamma_k}. \quad (3.6)$$

[Assumption 7](#) and [proposition 2.2](#) imply that

$$\|\nabla f(x_k^*) - \nabla f(x^*)\| \leq L\|x_k^* - x^*\|,$$

and, combining that with (3.6), we obtain

$$\|\pi(x_k^*)\| \leq \frac{2L}{\gamma_k} \|x_k^* - x^*\| + \frac{2\|\nabla f(x^*)\|}{\gamma_k}.$$

Together with the equality (3.5), we have

$$2\|\mathbb{E}(A(\xi)^\top PA(\xi))\| \|x_k^* - x^*\| \leq \frac{2L}{\gamma_k} \|x_k^* - x^*\| + \frac{2\|\nabla f(x^*)\|}{\gamma_k}. \quad (3.7)$$

Set $q := 2\|\mathbb{E}(A(\xi)^\top PA(\xi))\|$. By [assumption 3](#), there exists $K' \geq K$ such that $2L/\gamma_k < q$ for all $k \geq K'$.

Hence, rearranging (3.6), we obtain

$$\|x_k^* - x^*\| \leq \frac{2 \|\nabla f(x^*)\|}{\gamma_k(q - 2 L_f / \gamma_k)} = \mathcal{O}(\gamma_k^{-1}),$$

as desired. \square

Corollary 3.10. *In the situations of (P) and (P^k), let $\pi(x) := \mathbb{E}\|(0, A(\xi)x - c)_+\|^2$ and assume that assumptions 3 to 5 and 10 hold. Then there exists a unique solution x^* to (P) and, for all $k \in \mathbb{N}$, there exists a unique solution x_k^* to (P^k). Further, we have*

$$\|x_k^* - x^*\| = \mathcal{O}(\gamma_k^{-1}).$$

Proof. By lemma 3.5, j satisfies assumption 1. Since j is quadratic, j is also Lipschitz smooth, so assumption 7 also holds. The claim now follows from theorem 3.8. \square

3.2.2 Bounding the tracking error

The following analysis is an adaptation of techniques used in [4]. One notable difference is that we do not assume uniformly bounded variance or second moment of the stochastic gradients. We introduce the following notation

Notation 3.11. For any $k \in \mathbb{N}$, we define $A_k := \|x_k - x_k^*\|^2$, $a_k := \mathbb{E}(A_k)$, and $\Delta_k := \|x_k^* - x_{k+1}^*\|$. Further, we define $\xi_k^{[b_k]} := (\xi_k^1, \dots, \xi_k^{b_k}) \in \mathbb{R}^{m \times b_k}$ and $\mathbb{E}_k(X) := \mathbb{E}(X | \xi_{k-1}^{[b_{k-1}]}, \dots, \xi_1^{[b_1]})$. If assumption 8 holds, then f^k is $(L + \gamma_k L_\pi)$ -smooth. In that case, we define $L_k := L + \gamma_k L_\pi$.

Theorem 3.12. *In the situations of (Q) and (Q^k), assume that assumptions 1 to 4, 6, 8 and 9 hold. Then, for all $k \in \mathbb{N}$, the iterates $(x_k)_{k \in \mathbb{N}}$ of algorithm 2 satisfy*

$$a_{k+1} \leq (1 - \tilde{\rho}_k)a_k + 2M_k^2(1 + \gamma_k^2)\tau_k^2 + (1 + \eta_k^{-1})\Delta_k^2,$$

where $\tilde{\rho}_k := \mu\tau_k/2 - 2(M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2$ and $M_k^2 = \mathcal{O}(b_k^{-1})$.

Remark 3.13. Note that the strong convexity assumption is crucial for the above result to be useful, or otherwise there would not exist a step size τ_k that would lead to a contraction factor in front of a_k .

We will use the following two lemmata in the proof of theorem 3.12.

Lemma 3.14. *In the situations of (Q) and (Q^k), assume that assumptions 1 to 4, 6, 8 and 9 hold. Then, for all $k \in \mathbb{N}$, the iterates $(x_k)_{k \in \mathbb{N}}$ of algorithm 2 satisfy*

$$\mathbb{E}_k \|\tilde{G}^k(x_k)\|^2 \leq (M_k^2(1 + \gamma_k^2) + L_k^2) A_k + M_k^2(1 + \gamma_k^2),$$

where $M_k^2 = \mathcal{O}(b_k^{-1})$.

Proof. By [proposition 2.32](#) and [assumption 9](#), we have

$$\begin{aligned}
\mathbb{V}\text{ar}_k(\tilde{G}^k(x_k)) &= \frac{1}{b_k} \mathbb{V}\text{ar}_k(G^k(x_k, \xi)) \\
&\leq \frac{C}{b_k} (\|x_k\|^2 + \|x_k\|^2 \gamma_k^2 + \gamma_k^2 + 1) \\
&\leq \frac{C}{b_k} (2(\|x_k - x_k^*\|^2 + \|x_k^*\|^2) + 2\gamma_k^2(\|x_k - x_k^*\|^2 + \|x_k^*\|^2) + \gamma_k^2 + 1) \\
&= \frac{C}{b_k} (2(1 + \gamma_k^2)\|x_k - x_k^*\|^2 + (2\|x_k^*\|^2 + 1)\gamma_k^2 + 2\|x_k^*\|^2 + 1) \\
&= \frac{C}{b_k} (2(1 + \gamma_k^2)A_k + (2\|x_k^*\|^2 + 1)\gamma_k^2 + 2\|x_k^*\|^2 + 1) \\
&\leq \frac{1}{b_k} (2C(1 + \gamma_k^2)A_k + M^2(1 + \gamma_k^2)),
\end{aligned}$$

where $M^2 := \sup_{k \in \mathbb{N}} C(2\|x_k^*\|^2 + 1)$. Note that, by [theorem 3.1](#), the sequence $(x_k^*)_{k \in \mathbb{N}}$ converges, and so $M^2 < \infty$. Next, note that, by optimality of x_k^* for f^k and [assumption 8](#), we have

$$\|\nabla f^k(x_k)\|^2 = \|\nabla f^k(x_k) - \nabla f^k(x_k^*)\|^2 \leq L_k^2 A_k.$$

Putting the two bounds together and using [proposition 2.33](#), we get

$$\begin{aligned}
\mathbb{E}_k \|\tilde{G}^k(x_k)\|^2 &= \mathbb{V}\text{ar}(\tilde{G}^k(x_k)) + \|\nabla f^k(x_k)\|^2 \\
&\leq \frac{1}{b_k} (2C(1 + \gamma_k^2)A_k + M^2(1 + \gamma_k^2)) + L_k^2 A_k \\
&= \left(\frac{2C}{b_k}(1 + \gamma_k^2) + L_k^2 \right) A_k + \frac{M^2}{b_k}(1 + \gamma_k^2) \\
&= (M_k^2(1 + \gamma_k^2) + L_k^2) A_k + M_k^2(1 + \gamma_k^2),
\end{aligned}$$

where $M_k^2 := \max(2C, M^2)/b_k$. \square

Lemma 3.15. *In the situations of [\(Q\)](#) and [\(Q^k\)](#), assume that [assumptions 1 to 4, 6, 8 and 9](#) hold. Then, for all $k \in \mathbb{N}$, the iterates of [algorithm 2](#) satisfy*

$$\mathbb{E} \|x_{k+1} - x_k^*\|^2 \leq (1 - q_k)a_k + M_k^2(1 + \gamma_k^2)\tau_k^2,$$

for all $k \in \mathbb{N}$, where $q_k := \mu\tau_k - (M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2$ and $M_k^2 = \mathcal{O}(b_k^{-1})$.

Proof. Plugging in the definition of x_{k+1} and expanding, we get

$$\begin{aligned}
\|x_{k+1} - x_k^*\|^2 &= \|x_k - x_k^* - \tau_k \tilde{G}^k(x_k)\|^2 \\
&= A_k + \tau_k^2 \|\tilde{G}^k(x_k)\|^2 - 2\tau_k \langle x_k - x_k^*, \tilde{G}^k(x_k) \rangle.
\end{aligned}$$

Applying \mathbb{E}_k on both sides, we get

$$\mathbb{E}_k \|x_{k+1} - x_k^*\|^2 = A_k + \tau_k^2 \mathbb{E}_k \|\tilde{G}^k(x_k)\|^2 - 2\tau_k \langle x_k - x_k^*, g^k(x_k) \rangle.$$

Strong convexity of f^k yields

$$\mathbb{E}_k \|x_{k+1} - x_k^*\|^2 \leq (1 - \mu\tau_k)A_k + \tau_k^2 \mathbb{E}_k \|\tilde{G}^k(x_k)\|^2. \quad (3.8)$$

[Lemma 3.14](#) yields

$$\mathbb{E}_k \|\tilde{G}^k(x_k)\|^2 \leq (M_k^2(1 + \gamma_k^2) + L_k^2) A_k + M_k^2(1 + \gamma_k^2).$$

Plugging this into (3.8), we get

$$\mathbb{E}_k \|x_{k+1} - x_k^*\|^2 \leq \left(1 - \mu\tau_k + (M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2\right) A_k + M_k^2(1 + \gamma_k^2)\tau_k^2.$$

Now, taking expectations of both sides, [proposition 2.36](#) yields the claim. \square

We will now prove the first main theorem of this subsection.

Proof of Theorem 3.12. Let $k \in \mathbb{N}$. First, we have

$$\begin{aligned} A_{k+1} &= \mathbb{E}_k \|x_{k+1} - x_k^* + x_k^* - x_{k+1}^*\|^2 \\ &= \mathbb{E}_k \|x_{k+1} - x_k^*\|^2 + \Delta_k^2 + 2\mathbb{E}_k \langle x_{k+1} - x_k^*, x_k^* - x_{k+1}^* \rangle. \end{aligned}$$

We can apply the Cauchy-Schwarz and Young inequalities to obtain

$$A_{k+1} \leq (1 + \eta_k) \mathbb{E}_k \|x_{k+1} - x_k^*\|^2 + (1 + \eta_k^{-1}) \Delta_k^2,$$

for all $\eta_k > 0$. Hence, by applying $\mathbb{E}(\cdot)$ on both sides,

$$a_{k+1} \leq (1 + \eta_k) \mathbb{E} \|x_{k+1} - x_k^*\|^2 + (1 + \eta_k^{-1}) \Delta_k^2. \quad (3.9)$$

Using [lemma 3.15](#), we can bound the first term:

$$(1 + \eta_k) \mathbb{E} \|x_{k+1} - x_k^*\|^2 \leq (1 + \eta_k)(1 - q_k) a_k + (1 + \eta_k) M_k^2(1 + \gamma_k^2) \tau_k^2. \quad (3.10)$$

We choose $\eta_k = \min(1, \mu\tau_k/2)$ and have

$$\begin{aligned} (1 + \eta_k)(1 - q_k) &= (1 + \eta_k) \left(1 - \mu\tau_k + (M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2\right) \\ &= 1 + \eta_k - (1 + \eta_k)\mu\tau_k + (1 + \eta_k)(M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2 \\ &\leq 1 + \frac{\mu}{2}\tau_k - \mu\tau_k + 2(M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2 \\ &= 1 - \frac{\mu}{2}\tau_k + 2(M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2, \end{aligned}$$

Plugging this into (3.10) and using $1 + \eta_k \leq 2$ again, we arrive at the bound

$$(1 + \eta_k) \mathbb{E} \|x_{k+1} - x_k^*\|^2 \leq \left(1 - \frac{\mu}{2}\tau_k + 2(M_k^2(1 + \gamma_k^2) + L_k^2)\tau_k^2\right) a_k + 2M_k^2(1 + \gamma_k^2) \tau_k^2.$$

Together with (3.9), we obtain the first claim. The proof of the second claim follows analogously after applying the second part of [lemma 3.15](#) to bound (3.9). \square

3.2.3 Convergence rates

In the previous sections, we proved bounds on the iterates of [algorithm 2](#). We will now use these bounds to choose asymptotically optimal policies for the parameters $(\tau_k)_{k \in \mathbb{N}}$, $(\gamma_k)_{k \in \mathbb{N}}$, and $(b_k)_{k \in \mathbb{N}}$ in [algorithm 2](#), for solving (P).

Assumption 11. The random matrix $A(\xi)$ satisfies $\mathbb{E}\|A(\xi)\|_F^4 < \infty$.

Theorem 3.16. In the situations of (P) and (P^k), let $\pi(x) := \mathbb{E}\|(0, A(\xi)x - c)_+\|^2$ and assume that assumptions 3, 4, 6, 10 and 11 hold. Then, for any $\epsilon \in (0, 1/3)$, algorithm 2 with parameters $\tau_k = k^{-2/3}$, $\gamma_k = k^{1/3-\epsilon}$ and $b_k = 1 + k^{2(1/3-\epsilon)}$, converges, and yields iterates $(x_k)_{k \in \mathbb{N}}$ that satisfy

$$\mathbb{E}\|x_k - x^*\| = \mathcal{O}(\gamma_k^{-1}) = \mathcal{O}(k^{-1/3+\epsilon}),$$

where x^* denotes the solution to (P). Furthermore, it holds that

$$\mathbb{E}(\pi(x_k)) = \mathcal{O}(\gamma_k^{-2}) = \mathcal{O}(k^{-2/3+2\epsilon}).$$

Lemma 3.17. In the situations of (Q) and (Q^k), assume that assumptions 1 to 4 hold. Additionally, assume that f and π are differentiable. Then we have

$$\Delta_k \leq \frac{\gamma_{k+1} - \gamma_k}{\gamma_k} \frac{G}{\mu},$$

for all $k \in \mathbb{N}$, where $G := 2 \sup_{k \in \mathbb{N}} \|\nabla f(x_k^*)\| < \infty$.

Proof. Let $k \in \mathbb{N}$. The claim clearly holds if $x_k^* = x_{k+1}^*$. Assume for the rest of the proof that $x_k^* \neq x_{k+1}^*$. We have

$$f(x) = f^k(x) - \frac{\gamma_k}{2} \pi(x),$$

which implies

$$\nabla f(x_k^*) = -\frac{\gamma_k}{2} \nabla \pi(x_k^*), \quad (3.11)$$

by optimality of x_k^* for f^k . We can apply strong convexity of f (assumption 1) and proposition 2.12 to obtain

$$\begin{aligned} \frac{\mu}{2} \Delta_k^2 &\leq \langle x_k^* - x_{k+1}^*, \nabla f(x_k^*) - \nabla f(x_{k+1}^*) \rangle \\ &= \langle x_k^* - x_{k+1}^*, -\frac{\gamma_k}{2} \nabla \pi(x_k^*) + \frac{\gamma_{k+1}}{2} \nabla \pi(x_{k+1}^*) \rangle \\ &= \langle x_k^* - x_{k+1}^*, \frac{\gamma_{k+1} - \gamma_k}{2} \nabla \pi(x_{k+1}^*) + \frac{\gamma_k}{2} (\nabla \pi(x_{k+1}^*) - \nabla \pi(x_k^*)) \rangle \\ &= \frac{\gamma_{k+1} - \gamma_k}{2} \langle x_k^* - x_{k+1}^*, \nabla \pi(x_{k+1}^*) \rangle - \frac{\gamma_k}{2} \langle x_{k+1}^* - x_k^*, \nabla \pi(x_{k+1}^*) - \nabla \pi(x_k^*) \rangle. \end{aligned}$$

Convexity of π (assumption 2) implies that $\langle x - y, \pi(x) - \pi(y) \rangle \geq 0$ for all $x, y \in \mathbb{R}^d$, by proposition 2.8. Hence, by positivity of γ_k (assumption 3), we have

$$\frac{\mu}{2} \Delta_k^2 \leq \frac{\gamma_{k+1} - \gamma_k}{2} \langle x_k^* - x_{k+1}^*, \nabla \pi(x_{k+1}^*) \rangle$$

and an application of the Cauchy-Schwarz inequality along with the fact $\gamma_{k+1} > \gamma_k$ (assumption 3), yield

$$\frac{\mu}{2} \Delta_k^2 \leq \frac{\gamma_{k+1} - \gamma_k}{2} \cdot \Delta_k \cdot \|\nabla \pi(x_{k+1}^*)\|.$$

Dividing both sides by $\mu/2 \Delta_k$, we get

$$\Delta_k \leq \frac{\gamma_{k+1} - \gamma_k}{\mu} \|\nabla \pi(x_{k+1}^*)\|.$$

Substituting ∇f for $\nabla \pi$ via (3.11), we arrive at

$$\Delta_k \leq \frac{\gamma_{k+1} - \gamma_k}{\mu} \frac{2 \|\nabla f(x_{k+1}^*)\|}{\gamma_k} = \frac{\gamma_{k+1} - \gamma_k}{\gamma_k} \frac{2 \|\nabla f(x_{k+1}^*)\|}{\mu}.$$

Finally, by [theorem 3.1](#), we know that $x_k^* \rightarrow x^*$ for $k \rightarrow \infty$. Hence, continuity of ∇f implies that $(\nabla f(x_k^*))_{k \in \mathbb{N}}$ also converges, and in particular $\sup_{k \in \mathbb{N}} \|\nabla f(x_k^*)\| < \infty$. Hence,

$$\Delta_k \leq \frac{G}{\mu} \frac{\gamma_{k+1} - \gamma_k}{\gamma_k},$$

where $G := 2 \sup_{k \in \mathbb{N}} \|\nabla f(x_k^*)\|$, as desired. \square

Lemma 3.18 (Chung's lemma). *Let $(\alpha_k)_{k \in \mathbb{N}}$ be a nonnegative scalar sequence and $k_0 \in \mathbb{N}$ be such that*

$$\alpha_{k+1} \leq \left(1 - \frac{a}{k^s}\right) \alpha_k + \mathcal{O}\left(\frac{b}{k^{s+t}}\right)$$

for all $k \geq k_0$ and some $0 < s \leq 1$, $a, b, t > 0$. Then, it holds that

$$\alpha_k = \mathcal{O}\left(\frac{1}{k^t}\right).$$

Proof. See [3]. \square

We can now prove the main theorem of this section.

Proof of [Theorem 3.16](#). Let $k \in \mathbb{N}$. By the triangle inequality,

$$\mathbb{E}\|x_k - x^*\| \leq \mathbb{E}\|x_k - x_k^*\| + \|x_k^* - x^*\|. \quad (3.12)$$

First, we analyze $\mathbb{E}\|x_k - x_k^*\|$. We want to use [theorem 3.12](#), but for this we first need to show that [assumptions 1, 2, 8](#) and [9](#) hold. By Jensen's inequality ([proposition 2.30](#)), [assumption 11](#) implies [assumption 5](#). Hence, by [lemma 3.5](#), [assumption 1](#) is satisfied by j . Further, by [lemma 3.9](#), π satisfies [assumption 2](#). [Assumption 8](#) was verified in [lemma 3.7](#). What's left to show is that [assumption 9](#) holds. Let $Q(\xi) := A(\xi)^\top A(\xi)$, $\tilde{b}(\xi) := A(\xi)^\top b$, and $\tilde{c} := A(\xi)^\top c$. A stochastic gradient of j^k at $x \in \mathbb{R}^d$, denoted by $G^k(x, \xi)$, is given by

$$\begin{aligned} G^k(x, \xi) &= A(\xi)^\top (A(\xi)x - b + \gamma(0, A(\xi)x - c)_+) + \lambda x \\ &= Q(\xi)x - \tilde{b}(\xi) + \gamma(0, Q(\xi)x - \tilde{c}(\xi))_+ + \lambda x. \end{aligned}$$

We have

$$\begin{aligned} \|G^k(x, \xi)\| &\leq \|Q(\xi)x\| + \|\tilde{b}(\xi)\| + \gamma(\|Q(\xi)x\| + \|\tilde{c}(\xi)\|) + \lambda\|x\| \\ &\leq \|Q(\xi)\|_F\|x\| + \|\tilde{b}(\xi)\| + \gamma(\|Q(\xi)\|_F\|x\| + \|\tilde{c}(\xi)\|) + \lambda\|x\| \\ &= (\|Q(\xi)\|_F + \lambda)\|x\| + \gamma\|Q(\xi)\|_F\|x\| + \gamma\|\tilde{c}(\xi)\| + \|\tilde{b}(\xi)\|. \end{aligned}$$

Using the inequality $(a + b + c + d)^2 \leq 4(a^2 + b^2 + c^2 + d^2)$, $\forall a, b, c, d \in \mathbb{R}$, we can conclude

$$\mathbb{E}\|G^k(x, \xi)\|^2 \leq 4(\mathbb{E}(\|Q(\xi)\|_F + \lambda)^2 \|x\|^2 + \gamma^2 \mathbb{E}(\|Q(\xi)\|_F^2) \|x\|^2 + \gamma^2 \mathbb{E}\|\tilde{c}(\xi)\|^2 + \mathbb{E}\|\tilde{b}(\xi)\|^2).$$

Note that all expectations are finite, by [assumption 11](#). Indeed, it holds that $\|Q(\xi)\|_F^2 = \|A(\xi)^\top A(\xi)\|_F^2 \leq \|A(\xi)\|_F^4 < \infty$, and thus

$$\mathbb{E}(\|Q(\xi)\|_F + \lambda)^2 \leq 2\mathbb{E}\|Q(\xi)\|_F^2 + 2\lambda \leq 2\mathbb{E}\|A(\xi)\|_F^4 + 2\lambda < \infty.$$

The terms $\mathbb{E}\|\tilde{b}(\xi)\|^2$ and $\mathbb{E}\|\tilde{c}(\xi)\|^2$ are similarly bounded by a constant times $\mathbb{E}\|A(\xi)\|_F^2$, which is also finite by [assumption 11](#) and Jensen's inequality ([proposition 2.30](#)). Hence, [assumption 9](#) is satisfied.

With our choices for γ_k and b_k , [theorem 3.12](#) now yields

$$a_{k+1} \leq (1 - \tilde{\rho}_k)a_k + 2D\tau_k^2 + (1 + \eta_k^{-1})\Delta_k^2,$$

where $D \in (0, \infty)$ is a constant, $\eta_k = \min(1, \mu\tau_k/2)$ and $\tilde{\rho}_k = \mu\tau_k/2 - 2(D + L_k^2)\tau_k^2$. For large enough $k \in \mathbb{N}$, we have $(D + L_k^2)\tau_k^2 \approx \gamma_k^2\tau_k^2 = k^{-2/3-2\epsilon}$, and, since this decays faster than τ_k , we then have

$$\tilde{\rho}_k \geq \frac{\mu\tau_k}{4}. \quad (3.13)$$

By [lemma 3.17](#), we know that $\Delta_k = \mathcal{O}((\gamma_{k+1} - \gamma_k)/\gamma_k)$. To further analyze this, consider the function $h(x) := x^\alpha$ on \mathbb{R} for some $\alpha > 0$. By the mean value theorem, there exists some $\theta \in [x, y]$, s.t.

$$\frac{h(y) - h(x)}{y - x} = h'(\theta) = \alpha\theta^{\alpha-1}.$$

In particular, if $\alpha \leq 1$, it holds that

$$\frac{h(y) - h(x)}{y - x} \leq \alpha x^{\alpha-1}.$$

Setting $\alpha = 1/3 - \epsilon$ gives $h(k) = \gamma_k$, so

$$\gamma_{k+1} - \gamma_k \leq \alpha k^{\alpha-1}.$$

Hence,

$$\Delta_k^2 = \mathcal{O}(k^{-2}).$$

Combining this with (3.13), there exists $k_0 \in \mathbb{N}$ such that

$$a_{k+1} \leq \left(1 - \frac{\mu}{4k^{2/3}}\right) a_k + \mathcal{O}\left(\frac{1}{k^{4/3}}\right),$$

for all $k \geq k_0$. Hence, by [lemma 3.18](#), we have $a_k = \mathcal{O}(k^{-2/3})$ and an application of Jensen's inequality ([proposition 2.30](#)) yields $\mathbb{E}\|x_k - x_k^*\| = \mathcal{O}(k^{-1/3})$.

Finally, by [theorem 3.8](#), we know $\|x_k^* - x^*\| = \mathcal{O}(k^{-1/3+\epsilon})$. Combining this with (3.12), we get

$$\mathbb{E}\|x_k - x^*\| = \mathcal{O}(k^{-1/3+\epsilon}). \quad (3.14)$$

The remaining claim follows from the facts that π has Lipschitz gradients and $\pi(x^*) = \nabla\pi(x^*) = 0$, which together imply ([proposition 2.2](#))

$$\begin{aligned} \mathbb{E}(\pi(x_k)) &= \mathbb{E}(\pi(x_k) - \pi(x^*)) \\ &\leq \mathbb{E}\left(\langle x_k - x^*, \nabla\pi(x^*) \rangle + \frac{L_\pi}{2}\|x_k - x^*\|^2\right) \\ &= \frac{L_\pi}{2}\mathbb{E}\|x_k - x^*\|^2 \end{aligned}$$

for some constant $L_\pi \in (0, \infty)$. The claim now follows from (3.14). \square

Remark 3.19. The ϵ in the definition of γ_k is needed in order to ensure that the factor $1 - \tilde{\rho}_k$ is eventually smaller than 1 for all k large enough, without needing to know the constants involved in $\tilde{\rho}_k$.

3.2.4 Iterate averaging

We will now analyze the convergence properties of the iterate average $\bar{x}_k := 1/k \sum_{i=1}^k x_i$, where x_i denotes the i th iterate of [algorithm 2](#).

Lemma 3.20. *Let $(\alpha_k)_{k \in \mathbb{N}}$ be a sequence of real numbers such that $\alpha_k = \mathcal{O}(k^{-a})$ for some $a \in (0, 1)$. Then, we have*

$$\frac{1}{k} \sum_{i=1}^k \alpha_i = \mathcal{O}(k^{-a}).$$

Proof. If $\alpha_k = \mathcal{O}(k^{-a})$, then there exists a constant $c \in (0, \infty)$ and $k_0 \in \mathbb{N}$, such that $\alpha_k \leq c k^{-a}$ for all $k \geq k_0$, hence

$$\begin{aligned} \frac{1}{k} \sum_{i=1}^k \alpha_i - \frac{1}{k} \sum_{i=1}^{k_0-1} \alpha_i &= \frac{1}{k} \sum_{i=k_0}^k \alpha_i \\ &\leq \frac{c}{k} \sum_{i=k_0}^k i^{-a} \\ &\leq \frac{c}{k} \int_{k_0}^k x^{-a} dx \\ &= \frac{c(k^{1-a} - k_0^{1-a})}{k(1-a)} = \mathcal{O}(k^{-a}). \end{aligned}$$

Since $1/k \sum_{i=1}^{k_0-1} \alpha_i = \mathcal{O}(k^{-1})$ and $a \in (0, 1)$, we obtain $1/k \sum_{i=1}^k \alpha_i = \mathcal{O}(k^{-a})$. \square

Theorem 3.21. *In the situations of [\(P\)](#) and [\(P^k\)](#), let $\pi(x) = \mathbb{E}\|(A(\xi)x - b)_+\|^2$, and assume that [assumptions 3, 4, 6, 10](#) and [11](#) hold. Then, for all $\epsilon \in (0, 1/3)$, [algorithm 2](#) with parameters $\tau_k = k^{-2/3}$, $\gamma_k = k^{1/3-\epsilon}$ and $b_k = 1 + k^{2(1/3-\epsilon)}$, converges, and yields iterates $(x_k)_{k \in \mathbb{N}}$ that satisfy*

$$\mathbb{E}|j^k(\bar{x}_k) - j(x^*)| = \mathcal{O}(\gamma_k^{-1}) = \mathcal{O}(k^{-1/3+\epsilon}),$$

where $\bar{x}_k := 1/k \sum_{i=1}^k x_i$, for $k \in \mathbb{N}$, and x^* denotes the solution to [\(P\)](#).

Proof. Smoothness of j and π is verified in [lemma 3.7](#). Let $k \in \mathbb{N}$. We have

$$\begin{aligned} \mathbb{E}|j^k(\bar{x}_k) - j(x^*)| &\leq \mathbb{E}|j^k(\bar{x}_k) - j^k(x_k^*)| + |j^k(x_k^*) - j(x^*)| \\ &= \mathbb{E}(j^k(\bar{x}_k) - j^k(x_k^*)) + (j(x^*) - j^k(x_k^*)). \end{aligned} \tag{3.15}$$

We will first analyze $\mathbb{E}(j^k(\bar{x}_k) - j^k(x_k^*))$. By convexity of j and π , we have

$$j^k(\bar{x}_k) \stackrel{\text{def}}{=} j(\bar{x}_k) + \frac{\gamma_k}{2} \pi(\bar{x}_k) \leq \frac{1}{k} \sum_{i=1}^k j(x_i) + \frac{\gamma_k}{2k} \sum_{i=1}^k \pi(x_i). \tag{3.16}$$

Next, note that $j^i(x_i^*) \leq j^k(x_k^*)$ for all $i \in \{1, \dots, k\}$, and thus

$$\begin{aligned} 0 \leq j^k(\bar{x}_k) - j^k(x_k^*) &\leq \frac{1}{k} \sum_{i=1}^k (j(x_i) - j^k(x_k^*)) + \frac{\gamma_k}{2k} \sum_{i=1}^k \pi(x_i) \\ &\leq \frac{1}{k} \sum_{i=1}^k (j(x_i) - j^i(x_i^*)) + \frac{\gamma_k}{2k} \sum_{i=1}^k \pi(x_i) \\ &\stackrel{\text{def}}{=} \frac{1}{k} \sum_{i=1}^k (j^i(x_i) - j^i(x_i^*)) + \frac{1}{2k} \sum_{i=1}^k (\gamma_k - \gamma_i) \pi(x_i), \end{aligned}$$

where, in the last step, we used that $j(x) = j^i(x) - \frac{\gamma_i}{2}\pi(x)$ for all $x \in \mathbb{R}^d$, $i \in \mathbb{N}$. By [theorem 3.16](#), we have $\pi(x_i) = \mathcal{O}(\gamma_i^{-2})$, thus

$$\frac{1}{k} \sum_{i=1}^k (\gamma_k - \gamma_i) \pi(x_i) = \mathcal{O} \left(\frac{1}{k} \sum_{i=1}^k \frac{\gamma_k - \gamma_i}{\gamma_i^2} \right).$$

Using [lemma 3.20](#), we have

$$\frac{1}{k} \sum_{i=1}^k \frac{\gamma_k - \gamma_i}{\gamma_i^2} = \gamma_k \left(\frac{1}{k} \sum_{i=1}^k \frac{1}{\gamma_i^2} \right) - \frac{1}{k} \sum_{i=1}^k \frac{1}{\gamma_i} = \gamma_k \cdot \mathcal{O}(\gamma_k^{-2}) - \mathcal{O}(\gamma_k^{-1}) = \mathcal{O}(\gamma_k^{-1})$$

and thus

$$\frac{1}{k} \sum_{i=1}^k (\gamma_k - \gamma_i) \pi(x_i) = \mathcal{O}(\gamma_k^{-1}).$$

Next, note that $\nabla j^i(x_i^*) = 0$ for all $i \in \mathbb{N}$. Furthermore, since j^i is a linear combination of two Lipschitz-smooth functions with constants which we will call L and L_π , respectively, [proposition 2.3](#) implies that j^i must also be Lipschitz smooth with constant $L_i := L + \gamma_i L_\pi = \mathcal{O}(\gamma_i)$, for all $i \in \mathbb{N}$. Hence, by use of [proposition 2.2](#) and [theorem 3.16](#), we obtain

$$j^i(x_i) - j^i(x_i^*) \leq L_i \|x_i - x_i^*\|^2 = \mathcal{O}(\gamma_i^{-1}),$$

for all $i \in \mathbb{N}$. An application of [lemma 3.20](#) now yields

$$\frac{1}{k} \sum_{i=1}^k j^i(x_i) - j^i(x_i^*) = \mathcal{O}(\gamma_k^{-1}),$$

and combining with (3.16), we get

$$j^k(\bar{x}_k) - j^k(x_k^*) = \mathcal{O}(\gamma_k^{-1}). \quad (3.17)$$

Similarly, since $\pi(x^*) = 0$, we have

$$j(x^*) - j^k(x_k^*) = j^k(x^*) - j^k(x_k^*) \leq L_k \|x^* - x_k^*\|^2$$

and an application of [theorem 3.8](#) yields

$$j(x^*) - j^k(x_k^*) = \mathcal{O}(\gamma_k^{-1}). \quad (3.18)$$

Combining (3.15), (3.17) and (3.18), we arrive at the desired result. \square

3.3 Fast Convergence with Iterate Moving Averages

We will now analyze an accelerated version of the SSGD algorithm, which makes use of *iterate moving averages*.

Algorithm 3. For $k \in \mathbb{N}$, let $x_1 \in \mathbb{R}^d$, $\tau_k \in (0, 4/\mu)$, $\gamma_k \in (0, \infty)$ and $b_k \in \mathbb{N}$. In the setting of (Q^k), the

Iterate Moving Average SSGD (IMA-SSGD) iterates have the form

$$\begin{aligned} x_{k+1} &:= x_k - \tau_k \tilde{G}^k(x_k) \\ \hat{x}_{k+1} &:= \left(1 - \frac{\mu\tau_k}{4 - \mu\tau_k}\right) \hat{x}_k + \frac{\mu\tau_k}{4 - \mu\tau_k} x_{k+1}, \end{aligned}$$

where

$$\tilde{G}^k(x) := \frac{1}{b_k} \sum_{j=1}^{b_k} G^k(x, \xi_k^j),$$

$(\xi_i^j)_{i=1,\dots,k, j=1,\dots,b_k}$ are i. i. d. samples from the distribution of ξ and $G^k(x, \xi)$ is a stochastic subgradient of f^k at x . We refer to τ_k as a **step size**, γ_k as a **penalty parameter** and b_k as a **batch size**.

Our analysis of [algorithm 3](#) is an adaptation of methods used by Cutler and Drusvyatskiy in [4], who in turn adapt averaging techniques used by Ghadimi and Lan in [8]. The analysis hinges on the following fundamental lemma.

Lemma 3.22 (Averaging lemma). *Let $h : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$ be a convex function and let $(x_t)_{t \in \mathbb{N}_0}$ be a sequence of vectors in \mathbb{R}^d . Suppose that there are constants $c_1, c_2 \in \mathbb{R}$, a sequence of nonnegative scalars $(\rho_t)_{t \in \mathbb{N}}$, and scalar sequences $(V_t)_{t \in \mathbb{N}_0}$, $(\omega_t)_{t \in \mathbb{N}}$, satisfying*

$$\rho_t h(x_t) \leq (1 - c_1 \rho_t) V_{t-1} - (1 + c_2 \rho_t) V_t + \omega_t$$

for all $t \in \mathbb{N}$. Define $\hat{\Gamma}_0 := 0$,

$$\hat{\rho}_t := \frac{(c_1 + c_2)\rho_t}{1 + c_2\rho_t} \quad \text{and} \quad \hat{\Gamma}_t := \prod_{i=1}^t (1 - \hat{\rho}_i),$$

for all $t \in \mathbb{N}$. Further, let $\hat{x}_0 := x_0$ and recursively define the averages

$$\hat{x}_t := (1 - \hat{\rho}_t) \hat{x}_{t-1} + \hat{\rho}_t x_t$$

for all $t \in \mathbb{N}$. Suppose that the relations $c_1 + c_2 > 0$, $1 - c_1 \rho_t > 0$, and $1 + c_2 \rho_t > 0$ hold for all $t \in \mathbb{N}$. Then, the following estimate holds for all $t \in \mathbb{N}_0$:

$$\frac{h(\hat{x}_t)}{c_1 + c_2} + V_t \leq \hat{\Gamma}_t \left(\frac{h(x_0)}{c_1 + c_2} + V_0 + \sum_{i=1}^t \frac{\omega_i}{\hat{\Gamma}_i (1 + c_2 \rho_i)} \right).$$

Proof. See lemma 42 in [4]. □

Before we make use of the averaging lemma, we will derive some preparatory results.

Lemma 3.23. *Let [assumptions 1](#), [6](#) and [8](#) hold. Then, [algorithm 3](#) with step sizes $\tau_k \in (0, 1/L_k)$ yields iterates $(x_k)_{k \in \mathbb{N}}$, such that*

$$2\tau_k \mathbb{E}(f^k(x_{k+1}) - f^k(x)) \leq (1 - \mu\tau_k + 2m_k\tau_k^2) \mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + m_k M_x^2 \tau_k^2,$$

for all $x \in \mathbb{R}^d$, where $m_k := C(1 + \gamma_k^2)/b_k(1 - L_k\tau_k)$ and $M_x^2 := 2\|x\|^2 + 1$.

Proof. For $k \in \mathbb{N}$, let $\tau_k \in (0, 1/L_k)$ and define $z_k := \nabla f^k(x_k) - \tilde{G}^k(x_k)$. Then,

$$\begin{aligned} f^k(x_{k+1}) &\leq f^k(x_k) + \langle \nabla f^k(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2} \|x_{k+1} - x_k\|^2 \\ &= f^k(x_k) + \langle \tilde{G}^k(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2} \|x_{k+1} - x_k\|^2 + \langle z_k, x_{k+1} - x_k \rangle. \end{aligned}$$

By Cauchy-Schwarz and Young's inequality, for all $\epsilon_k > 0$, we have

$$\begin{aligned} f^k(x_{k+1}) &\leq f^k(x_k) + \langle \tilde{G}^k(x_k), x_{k+1} - x_k \rangle + \frac{L_k + \epsilon_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + \frac{\epsilon_k}{2} \|z_k\|^2 \\ &= f^k(x_k) + \langle \tilde{G}^k(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\tau_k} \|x_{k+1} - x_k\|^2 \\ &\quad + \frac{L_k + \epsilon_k^{-1} + \tau_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + \frac{\epsilon_k}{2} \|z_k\|^2, \end{aligned} \tag{3.19}$$

where in the last step we added and subtracted $1/2\tau_k \|x_{k+1} - x_k\|$. Using [proposition 2.43](#), we see that x_{k+1} is the minimizer of the $1/2\tau_k$ -strongly convex function $x \mapsto \langle \tilde{G}^k(x_k), x - x_k \rangle + 1/2\tau_k \|x - x_k\|^2$. Hence, by the last statement in [proposition 2.12](#), we have

$$\langle \tilde{G}^k(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\tau_k} \|x_{k+1} - x_k\|^2 \leq \langle \tilde{G}^k(x_k), x - x_k \rangle + \frac{1}{2\tau_k} \|x - x_k\|^2 - \frac{1}{2\tau_k} \|x - x_{k+1}\|^2$$

for all $x \in \mathbb{R}^d$. Plugging this into (3.19), we obtain

$$\begin{aligned} f^k(x_{k+1}) &\leq f^k(x_k) + \langle \tilde{G}^k(x_k), x - x_k \rangle + \frac{1}{2\tau_k} \|x - x_k\|^2 - \frac{1}{2\tau_k} \|x - x_{k+1}\|^2 \\ &\quad + \frac{L_k + \epsilon_k^{-1} + \tau_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + \frac{\epsilon_k}{2} \|z_k\|^2. \end{aligned}$$

We would like to use strong convexity of f^k to proceed. To do this, we first need to add and subtract $\langle \nabla f^k(x_k), x - x_k \rangle$. Applying [proposition 2.12](#) then yields

$$\begin{aligned} f^k(x_{k+1}) &\leq f^k(x) - \frac{\mu}{2} \|x - x_k\|^2 - \langle z_k, x_k - x \rangle + \frac{1}{2\tau_k} \|x - x_k\|^2 - \frac{1}{2\tau_k} \|x - x_{k+1}\|^2 \\ &\quad + \frac{L_k + \epsilon_k^{-1} + \tau_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + \frac{\epsilon_k}{2} \|z_k\|^2 \end{aligned}$$

for all $x \in \mathbb{R}^d$. Simplifying, and noting that $\langle z_k, x_k - x \rangle = -\langle z_k, x - x_k \rangle$, we thus have

$$\begin{aligned} f^k(x_{k+1}) &\leq f^k(x) + \left(\frac{1}{2\tau_k} - \frac{\mu}{2} \right) \|x - x_k\|^2 + \langle z_k, x - x_k \rangle - \frac{1}{2\tau_k} \|x - x_{k+1}\|^2 \\ &\quad + \frac{L_k + \epsilon_k^{-1} + \tau_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + \frac{\epsilon_k}{2} \|z_k\|^2 \end{aligned}$$

for all $x \in \mathbb{R}^d$ and $\epsilon_k > 0$. Choosing $\epsilon_k := \tau_k/(1 - L_k\tau_k)$, we obtain

$$f^k(x_{k+1}) \leq f^k(x) + \left(\frac{1}{2\tau_k} - \frac{\mu}{2} \right) \|x - x_k\|^2 + \langle z_k, x - x_k \rangle - \frac{1}{2\tau_k} \|x - x_{k+1}\|^2 + \frac{\tau_k}{2(1 - L_k\tau_k)} \|z_k\|^2,$$

for all $x \in \mathbb{R}^d$. Taking expectations, we can drop the inner product term, and subsequently multiplying by $2\tau_k$ yields

$$2\tau_k \mathbb{E}(f^k(x_{k+1}) - f^k(x)) \leq (1 - \mu\tau_k) \mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + \frac{\tau_k^2}{1 - L_k\tau_k} \mathbb{E}\|z_k\|^2, \tag{3.20}$$

for all $x \in \mathbb{R}^d$. Note that, by definition,

$$\mathbb{E}_k \|z_k\|^2 = \mathbb{E}_k \|\tilde{G}^k(x_k) - \nabla f^k(x_k)\|^2 = \mathbb{E}_k \|\tilde{G}^k(x_k) - \mathbb{E}_k(G^k(x_k))\|^2 = \text{Var}_k(\tilde{G}^k(x_k)),$$

thus

$$\mathbb{E}\|z_k\|^2 = \mathbb{E}(\text{Var}_k(\tilde{G}^k(x_k))), \quad (3.21)$$

by proposition 2.36. Assumption 9 and proposition 2.39 imply, for all $x \in \mathbb{R}^d$,

$$\begin{aligned} \text{Var}_k(\tilde{G}^k(x_k)) &= \frac{1}{b_k} \text{Var}_k(G^k(x_k)) \\ &\leq \frac{C}{b_k} \left(\|x_k\|^2 + \|x_k\|^2 \gamma_k^2 + \gamma_k^2 + 1 \right) \\ &\leq \frac{C}{b_k} \left(2\|x_k - x\|^2 + 2\|x\|^2 + 2\|x_k - x\|^2 \gamma_k^2 + 2\|x\|^2 \gamma_k^2 + \gamma_k^2 + 1 \right) \\ &= \frac{C}{b_k} \left(2(1 + \gamma_k^2) \|x_k - x\|^2 + 2(1 + \gamma_k^2) \|x\|^2 + \gamma_k^2 + 1 \right) \\ &= \frac{C(1 + \gamma_k^2)}{b_k} \left(2\|x_k - x\|^2 + 2\|x\|^2 + 1 \right) \\ &= \frac{C(1 + \gamma_k^2)}{b_k} \left(2\|x_k - x\|^2 + M_x^2 \right), \end{aligned}$$

where $M_x^2 := 2\|x\|^2 + 1$, and we used that $(a + b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$. Taking expectations on both sides, and using (3.21), we have

$$\mathbb{E}\|z_k\|^2 \leq \frac{C(1 + \gamma_k^2)}{b_k} \left(2\mathbb{E}\|x_k - x\|^2 + M_x^2 \right).$$

Define $m_k := C(1 + \gamma_k^2)/b_k(1 - L_k \tau_k)$. Combining the above with (3.20), we obtain

$$2\tau_k \mathbb{E}(f^k(x_{k+1}) - f^k(x)) \leq (1 - \mu\tau_k + 2m_k\tau_k^2) \mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + m_k M_x^2 \tau_k^2,$$

for all $x \in \mathbb{R}^d$. □

In [4], the authors make an assumption, which in our setting would essentially boil down to imposing global boundedness of $\|\nabla \pi(x)\|$. For our purposes, however, this would be too strong, which motivates the following relaxed assumption.

Assumption 12. The penalty function π is differentiable and there exist constants $D_1, D_2 \in (0, \infty)$ such that

$$\|\nabla \pi(x) - \nabla \pi(y)\| \leq D_1 + D_2 \|x - y\|$$

for all $x, y \in \mathbb{R}^d$.

Note that Assumption 12 holds if π is differentiable and Lipschitz continuous or has Lipschitz continuous gradients.

Lemma 3.24. *In the situation of (Q^k), assume that assumption 12 holds. Then, for all $t, i \in \mathbb{N}$, $u, v \in \mathbb{R}^d$, we have*

$$(f^t(u) - f^t(v)) - (f^i(u) - f^i(v)) \leq \frac{(D_1^2 + \|\nabla \pi(v)\|^2)(\gamma_t - \gamma_i)^2}{2\epsilon} + \left(\frac{D_2(\gamma_t - \gamma_i) + \epsilon/2}{2} \right) \|u - v\|^2,$$

for all $\epsilon > 0$.

Proof. Let $u, v \in \mathbb{R}^d$. By definition, we have

$$f^k(u) - f^k(v) = f(u) - f(v) + \frac{\gamma_k}{2}(\pi(u) - \pi(v))$$

for all $k \in \mathbb{N}$. Hence, for all $t, i \in \mathbb{N}$,

$$(f^t(u) - f^t(v)) - (f^i(u) - f^i(v)) = \frac{\gamma_t - \gamma_i}{2}(\pi(u) - \pi(v)). \quad (3.22)$$

For $\tau \in [0, 1]$, let $u_\tau := v + \tau(u - v)$. By the fundamental theorem of calculus and Cauchy-Schwarz, we have

$$\begin{aligned} \pi(u) - \pi(v) &= \int_0^1 \langle \nabla \pi(u_\tau), u - v \rangle d\tau \\ &\leq \sup_{\tau \in [0, 1]} \|\nabla \pi(u_\tau)\| \|u - v\|. \end{aligned}$$

We can use [assumption 12](#) and the triangle inequality to obtain

$$\begin{aligned} \|\nabla \pi(u_\tau)\| &\leq \|\nabla \pi(u_\tau) - \nabla \pi(v)\| + \|\nabla \pi(v)\| \\ &\leq D_1 + D_2 \tau \|u - v\| + \|\nabla \pi(v)\| \\ &\leq D_1 + D_2 \|u - v\| + \|\nabla \pi(v)\| \end{aligned}$$

for all $\tau \in [0, 1]$. Thus,

$$(\gamma_t - \gamma_i)(\pi(u) - \pi(v)) \leq (D_1 + \|\nabla \pi(v)\|)(\gamma_t - \gamma_i)\|u - v\| + D_2(\gamma_t - \gamma_i)\|u - v\|^2.$$

By Young's inequality, for all $\epsilon > 0$,

$$(\gamma_t - \gamma_i)(\pi(u) - \pi(v)) \leq \frac{(D_1 + \|\nabla \pi(v)\|)^2(\gamma_t - \gamma_i)^2}{2\epsilon} + \frac{\epsilon}{2}\|u - v\|^2 + D_2(\gamma_t - \gamma_i)\|u - v\|^2,$$

hence, using the fact that $(a + b)^2 \leq 2(a^2 + b^2)$ for all $a, b \in \mathbb{R}$, we have

$$(\gamma_t - \gamma_i)(\pi(u) - \pi(v)) \leq \frac{(D_1^2 + \|\nabla \pi(v)\|^2)(\gamma_t - \gamma_i)^2}{\epsilon} + \left(D_2(\gamma_t - \gamma_i) + \frac{\epsilon}{2}\right)\|u - v\|^2.$$

Using this bound in [\(3.22\)](#), we arrive at

$$(f^t(u) - f^t(v)) - (f^i(u) - f^i(v)) \leq \frac{(D_1^2 + \|\nabla \pi(v)\|^2)(\gamma_t - \gamma_i)^2}{2\epsilon} + \left(\frac{D_2(\gamma_t - \gamma_i) + \epsilon/2}{2}\right)\|u - v\|^2,$$

as desired. \square

Lemma 3.25. *In the situation of [\(Q^k\)](#), assume that [assumptions 1 to 4, 6 and 8](#) hold. Further, let $(x_k)_{k \in \mathbb{N}}$ be iterates generated by [algorithm 3](#), with $\gamma_k := \gamma \cdot k^\alpha$ for $\alpha \in (0, 1)$, $\gamma \in (0, \infty)$, $b_k \geq 8C\tau_k(1 + \gamma_k^2)/\mu$, and $\tau_k \in (0, 1/2L_k)$ for all $k \in \mathbb{N}$. Then, there exists a sequence $(b_k)_{k \in \mathbb{N}}$ with $b_k = \mathcal{O}(\gamma_k)$, and a natural number $K \in \mathbb{N}$, such that for all $k \in \mathbb{N}$ and $t \in \mathbb{N}_0$ with $k, t \geq K$, it holds that*

$$2\tau_t \mathbb{E}(f^k(x_{t+1}) - f^k(x_k^*)) \leq \left(1 - \frac{\mu}{2}\tau_t\right) \mathbb{E}\|x_k^* - x_t\|^2 - \left(1 - \frac{\mu}{4}\tau_t\right) \mathbb{E}\|x_k^* - x_{t+1}\|^2 + \frac{\mu}{2}M^2\tau_t,$$

where $M^2 \in (0, \infty)$.

Proof. Let $k \in \mathbb{N}$, $t \in \mathbb{N}_0$, and let $\tau_t \in (0, 1/2L_t)$. Note that [assumption 8](#) implies [assumption 12](#). Hence,

we can apply [lemmas 3.23](#) and [3.24](#) and have, for all $\epsilon > 0$,

$$\begin{aligned}
2\tau_t \mathbb{E}(f^k(x_{t+1}) - f^k(x_k^*)) &\leq 2\tau_t(f^t(x_{t+1}) - f^t(x_k^*)) \\
&\quad + \tau_t \left(\frac{(D_1^2 + \|\nabla \pi(x_k^*)\|^2)(\gamma_k - \gamma_t)^2}{\epsilon} + (D_2(\gamma_k - \gamma_t) + \epsilon/2) \|x_k^* - x_{t+1}\|^2 \right) \\
&\leq (1 - \mu\tau_t + 2m_t\tau_t^2) \mathbb{E}\|x_k^* - x_t\|^2 - \mathbb{E}\|x_k^* - x_{t+1}\|^2 + m_t M_{x_k^*}^2 \tau_t^2 \\
&\quad + \tau_t \left(\frac{(D_1^2 + \|\nabla \pi(x_k^*)\|^2)(\gamma_k - \gamma_t)^2}{\epsilon} + (D_2(\gamma_k - \gamma_t) + \epsilon/2) \|x_k^* - x_{t+1}\|^2 \right) \\
&= (1 - \mu\tau_t + 2m_t\tau_t^2) \mathbb{E}\|x_k^* - x_t\|^2 \\
&\quad - (1 - \tau_t(D_2(\gamma_k - \gamma_t) + \epsilon/2)) \mathbb{E}\|x_k^* - x_{t+1}\|^2 + m_t M_{x_k^*}^2 \tau_t^2 \\
&\quad + \tau_t \frac{(D_1^2 + \|\nabla \pi(x_k^*)\|^2)(\gamma_k - \gamma_t)^2}{\epsilon},
\end{aligned}$$

where $m_t = C(1 + \gamma_t^2)/b_t(1 - L_t\tau_t)$ and $M_{x_k^*}^2 = 2\|x_k^*\|^2 + 1$. Note that, by [theorem 3.1](#), we have $\sup_{k \in \mathbb{N}} M_{x_k^*}^2 \leq M^2 \in (0, \infty)$. Since $\tau_t \in (0, 1/2L_t)$, we have $1 - L_t\tau_t \geq 1/2$, and thus

$$m_t \leq \frac{2C(1 + \gamma_t^2)}{b_t}.$$

With the choice of batch size $b_t \geq 8C\tau_t(1 + \gamma_t^2)/\mu$, we then have

$$2m_t\tau_t^2 \leq 2\left(\frac{\mu}{4\tau_t}\right)\tau_t^2 = \frac{\mu}{2}\tau_t,$$

and

$$m_t M_{x_k^*}^2 \tau_t^2 \leq \frac{\mu}{4} M^2 \tau_t.$$

Since $\gamma_i = \gamma \cdot i^\alpha$ with $\alpha \in (0, 1)$, for all $i \in \mathbb{N}$, there exists $k_0 \in \mathbb{N}$, such that for all $k, t \geq k_0$, it holds that $\gamma_k - \gamma_t \leq \mu/(8D_2)$ (this follows from the mean-value theorem, see for example the proof of [theorem 3.16](#) for the argument). Hence, with the choice $\epsilon := (D_1^2 + 1)\mu/4 \leq \mu/4$, we have

$$D_2(\gamma_k - \gamma_t) + \frac{\epsilon}{2} \leq \frac{\mu}{4},$$

for all $k, t \geq k_0$. The extra factor $D_1^2 + 1$ in the definition of ϵ exists to simplify terms later. Since x_k^* is optimal for f^k , we have

$$0 = \nabla f^k(x_k^*) = \nabla f(x_k^*) + \frac{\gamma_k}{2} \nabla \pi(x_k^*),$$

which implies $\|\nabla \pi(x_k^*)\| = \|\nabla f(x_k^*)\|/\gamma_k \leq G/\gamma_k$, where $G := \sup_{k \in \mathbb{N}} \|\nabla f(x_k^*)\|$ and $G < \infty$ due to convergence of $(x_k^*)_{k \in \mathbb{N}}$ ([theorem 3.1](#)) and continuity of ∇f . Since, $\lim_{k \rightarrow \infty} \gamma_k = \infty$ and $(\gamma_k)_{k \in \mathbb{N}}$ is increasing, there exists some $k_1 \in \mathbb{N}$, such that $G/\gamma_k \leq 1$ for all $k \geq k_1$. Keeping in mind the definition of ϵ , we obtain

$$\frac{(D_1^2 + \|\nabla \pi(x_k^*)\|^2)(\gamma_k - \gamma_t)^2}{\epsilon} \leq \frac{(D_1^2 + 1)(\gamma_k - \gamma_t)^2}{\epsilon} = \frac{4(\gamma_k - \gamma_t)^2}{\mu}$$

for all $k \in \mathbb{N}$ with $k \geq k_1$. Note that there exists a natural number, which we will also refer to as k_0 for

simplicity, such that $(\gamma_k - \gamma_t)^2 \leq (\mu^2/16)M^2$ for all $k, t \geq k_0$. Hence,

$$\frac{(D_1^2 + \|\nabla \pi(x_k^*)\|^2)(\gamma_k - \gamma_t)^2}{\epsilon} \leq \frac{\mu}{4} M^2,$$

for all $k, t \geq k_0$. Putting everything together, we arrive at the bound

$$2\tau_t \mathbb{E}(f^k(x_{t+1}) - f^k(x_k^*)) \leq \left(1 - \frac{\mu}{2}\tau_t\right) \mathbb{E}\|x_k^* - x_t\|^2 - \left(1 - \frac{\mu}{4}\tau_t\right) \mathbb{E}\|x_k^* - x_{t+1}\|^2 + \frac{\mu}{2}M^2\tau_t,$$

for all $k, t \geq \max(k_0, k_1)$. \square

Corollary 3.26. *In the situation of [lemma 3.25](#), assume additionally that $\tau_k < \mu/4$ for all $k \in \mathbb{N}$. Then, it holds that the iterates $(\hat{x}_k)_{k \in \mathbb{N}}$ generated by [algorithm 3](#) satisfy*

$$f^k(\hat{x}_k) - f^k(x_k^*) \leq \hat{\Gamma}_k \left(f^k(x_0) - f^k(x_k^*) + \frac{\mu}{8}V_0 + \frac{\mu^2 M^2}{16} \sum_{i=1}^k \frac{\tau_{i-1}}{\hat{\Gamma}_i(1 - \frac{\mu}{4}\tau_{i-1})} \right),$$

where

$$\hat{\Gamma}_k := \prod_{i=1}^k \left(1 - \frac{\mu\tau_{i-1}}{4 - \mu\tau_{i-1}}\right),$$

for all $k \in \mathbb{N}$.

Proof. Let $h(x) := f^k(x) - f^k(x_k^*)$ for $x \in \mathbb{R}^d$, $\rho_{t+1} := 2\tau_t$, $c_1 = \mu/4$, $c_2 = -\mu/8$, $V_i := \mathbb{E}\|x_k^* - x_i\|^2$, $i \in \{t, t+1\}$, and $\omega_{t+1} := (\mu/2)M^2\tau_t$. Then, by [lemma 3.25](#), we have

$$\rho_{t+1}h(x_{t+1}) \leq (1 - c_1\rho_{t+1})V_t - (1 + c_2\rho_{t+1})V_{t+1} + \omega_{t+1},$$

for all $t \in \mathbb{N}_0$. Setting $k := t + 1$, we thus have

$$\rho_k h(x_k) \leq (1 - c_1\rho_k)V_{k-1} - (1 + c_2\rho_k)V_k + \omega_k.$$

for all $k \in \mathbb{N}$. For $\tau_{k-1} \in (0, 4/\mu)$, we have $0 < c_1\rho_k < 1$, $-1 < c_2\rho_k < 0$, hence $1 - c_1\rho_k > 0$ and $1 + c_2\rho_k > 0$. Of course, $c_1 + c_2 = \mu/8 > 0$. Thus, all conditions of [lemma 3.22](#) are satisfied. Setting

$$\hat{\rho}_k := \frac{(c_1 + c_2)\rho_k}{1 + c_2\rho_k} = \frac{\mu\tau_{k-1}}{4 - \mu\tau_{k-1}}, \quad \hat{x}_k := \left(1 - \frac{\mu\tau_{k-1}}{4 - \mu\tau_k}\right) \hat{x}_{k-1} + \frac{\mu\tau_{k-1}}{4 - \mu\tau_{k-1}} x_k,$$

and

$$\Gamma^k := \prod_{i=1}^k (1 - \hat{\rho}_i),$$

we can now conclude

$$h(\hat{x}_k) \leq \hat{\Gamma}_k \left(h(x_0) + (c_1 + c_2)V_0 + (c_1 + c_2) \sum_{i=1}^k \frac{\omega_i}{\hat{\Gamma}_i(1 + c_2\rho_i)} \right),$$

as desired. \square

4 Applications & Numerical Examples

4.1 Optimal Control

4.2 Reinforcement Learning

5 Summary and Outlook

Restate problem

Summarize main contributions

Outlook: Extension to settings beyond strong-convexity assumption. High-probability bounds. Adaptive gradient methods. More general penalties. Non-asymptotic bounds. Extension to online setting.

Bibliography

- [1] Joseph K Blitzstein and Jessica Hwang. *Introduction to probability*. Chapman and Hall/CRC, 2019.
- [2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [3] Kai Lai Chung. “On a stochastic approximation method”. In: *The Annals of Mathematical Statistics* (1954), pp. 463–483.
- [4] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. “Stochastic Optimization under Distributional Drift”. In: *Journal of Machine Learning Research* 24.147 (2023), pp. 1–56. URL: <http://jmlr.org/papers/v24/21-1410.html>.
- [5] John C Duchi. “Introductory lectures on stochastic optimization”. In: *The mathematics of data* 25 (2018), pp. 99–186.
- [6] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.
- [7] Guillaume Garrigos and Robert M Gower. “Handbook of convergence theorems for (stochastic) gradient methods”. In: *arXiv preprint arXiv:2301.11235* (2023).
- [8] Saeed Ghadimi and Guanghui Lan. “Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework”. In: *SIAM Journal on Optimization* 22.4 (2012), pp. 1469–1492.
- [9] A. Nemirovski et al. “Robust Stochastic Approximation Approach to Stochastic Programming”. In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609. DOI: [10.1137/070704277](https://doi.org/10.1137/070704277). eprint: <https://doi.org/10.1137/070704277>. URL: <https://doi.org/10.1137/070704277>.
- [10] Herbert Robbins and Sutton Monro. “A stochastic approximation method”. In: *The annals of mathematical statistics* (1951), pp. 400–407.