Penalty Methods for Almost Surely Constrained Convex Optimization

With Applications to Optimal Control and Machine Learning

Amir Miri Lavasani

January 9, 2026

Contents

1 Introduction

Convex optimization is concerned with problems of the form

$$\min_{x \in \mathcal{X}} f(x),$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a convex function and $\mathcal{X} \subset \mathbb{R}^d$ is a convex set. Typically, the set $\mathcal{X}$ is called the *feasible set*, elements $x \in \mathcal{X}$ are *feasible points*, and f is called the *objective function*, or simply *objective*. Oftentimes, a point $x \in \mathbb{R}^d$ is referred to as a *decision variable*. In practice, the feasible set $\mathcal{X}$ is often defined implicitly through the use of auxillary functions. Additionally, the constraints may involve random variables, which are supposed to capture uncertainty in the problem, which should be controlled. The general form of a *chance constrained* convex optimization problem formulated in such a way has the following form.

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} f(x) \\ &\text{subject to (s.t.)} \quad \mathbb{P}(h(x,\xi) = 0) \geq \delta_1 \\ &\qquad\qquad\qquad\quad \mathbb{P}(g(x,\xi) \leq 0) \geq \delta_2\,, \end{aligned}$$

where $\delta_1, \delta_2 \in (0,1]$, $g(\cdot,\xi)\colon \mathbb{R}^d \times \Omega \to \mathbb{R}^n$, and $h(\cdot,\xi)\colon \mathbb{R}^d \times \Omega \to \mathbb{R}^p$ is affine in the first argument for (almost) all $\omega \in \Omega$. We will consider the special case of convex optimization problems with *almost sure affine inequality constraints*, where we assume $x \mapsto g(x,\xi)$ is affine, $h(x,\xi) \equiv 0$, and $\delta_2 = 1$. Thus, our problem of interest has the form

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} f(x) \\ &\text{s.t.} \quad A(\xi)x + b(\xi) \leq 0 \quad \text{almost surely (a.s.)}, \end{aligned} \tag{P$^{\text{new}}$}$$

where $A(\xi)$ is a random matrix and $b(\xi)$ is a random vector. Further, we will focus on the case where f is *strongly convex*.

Example 1.1 (Inventory control)**.** This example is adapted from section 4.8.2 in [Eva13]. Let $T \in (0,\infty)$ represent a time period, and, for $t \in (0,T)$, we let ξ_t be a random variable. We introduce the variables

$$\begin{aligned} x(t) &= \text{amount of inventory at time } t \\ \alpha(t) &= \text{rate of ordering from manufacturers} \\ d(t,\xi_t) &= \text{customer demand (random)} \\ \gamma &= \text{cost of ordering one unit} \\ \beta &= \text{cost of storing one unit.} \end{aligned}$$

The cost of inventory at time $t \in (0,T)$ is given by

$$c(\alpha(\cdot),t) = \gamma\alpha(t) + \beta x(t)\,,$$

and the total cost over the entire time period is

$$C(\alpha(\cdot)) = \int_0^T c(t)\,\mathrm{d}t\,.$$

Initially, we hold $x(0) := x_0 \in (0,\infty)$ units of items. The relationship between $x(t)$, $\alpha(t)$, and $d(t,\xi)$ is

$$\dot{x}(t) = \alpha(t) - d(t,\xi_t) \quad \text{a. s.}$$

Our goal is to choose an ordering policy $t \mapsto \alpha(t)$ such that the total cost $C(\alpha(\cdot))$ is minimized, all while ensuring that demand is filled, i. e. we want $x(t) \geq 0$ a. s. for all $t \in (0,T)$. To turn this problem into one that looks like ($\mathrm{P}^{\mathrm{new}}$), we descretize the continuous time period $(0,\,T)$ to a discrete one $\{t_0, t_1, \ldots, t_{n-1}\}$ for some $n \in \mathbb{N}$, and $0 < t_0 < t_1 < \cdots < t_{n-1} \in (0,T)$. For $k \in \{0,\ldots,n-1\}$, we set $x_k := x(t_k)$, $\alpha_k := \alpha(t_k)$, and $d_k(\xi) := d(t_k, \xi_{t_k})$ with $\xi := (\xi_1,\ldots,\xi_n)^\top$. This leads us to the following discrete version of the above stochastic optimization problem (see section 5.1 for details)

$$\begin{aligned} &\min_{\alpha\in\mathbb{R}^n} \mathbb{E}\left(\sum_{k=0}^{n-1} \gamma\alpha_k + \beta\sum_{j=0}^{k-1}(\alpha_j - d_j(\xi))\right) \\ &\text{s. t.} \quad x_0 + \sum_{j=0}^{k-1}(\alpha_j - d_j(\xi)) \geq 0 \quad \text{a. s. for all } k \in \{1,\ldots,n\}. \end{aligned} \tag{1.1}$$

Let $A = (a_{ij})_{i,j\in\{1,\ldots,n\}} \in \mathbb{R}^{n\times n}$ be defined by $a_{ij} = 1$ if $i \geq j$ and $a_{ij} = 0$, otherwise. Then we can rewrite the objective as

$$\mathbb{E}\left(\sum_{k=0}^{n-1} \gamma\alpha_k + \beta\langle A_k, \alpha_k - d_k\rangle\right)$$

and the contraints as

$$x_0 + \langle A_k, \alpha_k - d_k\rangle \geq 0 \quad \text{a. s. for all } k \in \{1,\ldots,n\},$$

where A_k denotes the k-th row of A. Let I_n denote the $n \times n$ matrix on $\mathbb{R}^n$, $\mathbf{1} := (1,\ldots,1)^\top \in \mathbb{R}^n$, and let $d(\xi) := (d_0(\xi),\ldots,d_{n-1}(\xi))^\top$, $\alpha(\xi) := (\alpha_0(\xi),\ldots,\alpha_{n-1}(\xi))^\top$. Further, we define $v := \mathbf{1}^\top(\beta A + \gamma I_n)$, $\eta(\xi) := \mathbf{1}^\top\beta A d(\xi)$, and $b(\xi) := Ad(\xi) - x_0\mathbf{1}$. Adding a quadratic regularizer $\alpha \mapsto \lambda\|\alpha\|_2^2$ to the objective, with $\|\cdot\|_2$ the standard euclidian norm on $\mathbb{R}^n$ and $\lambda > 0$, we can once again rewrite the above as

$$\begin{aligned} &\min_{\alpha\in\mathbb{R}^n} \mathbb{E}\Big(\langle v, \alpha\rangle - \eta(\xi) + \lambda\|\alpha\|_2^2\Big) \\ &\text{s. t.} \quad -A\alpha + b(\xi) \leq 0 \quad \text{a. s.}, \end{aligned}$$

which is a convex optimization problem of the form ($\mathrm{P}^{\mathrm{new}}$). As an extension, one could consider the case where the costs γ and β change over time and/or are subject to randomness. We will come back to this problem in section 5.1.

Example 1.2 (Maximal margin classification)**.** This example and its presentation are based on section 6.3 in [Fer+19]. A common problem in machine learning is binary classification, where we are given a dataset $D := \{\,(a_i, b_i) \mid i = 1,\ldots,\mathrm{n}\,\}$ of size $n \in \mathbb{N}$, where each $a_i \in \mathbb{R}^d$ is called *feature* and each $b_i \in \{0,1\}$ is called *label*. As a specific example, one can think of the problem of classifying images of pets as either a cat ("0") or dog ("1"). The features and labels can be seen as samples from an underlying probability distribution $\mathbb{P}$ on the space $\mathbb{R}^d \times \{0,1\}$. Given this dataset, we aim to learn a rule to classify new samples $a_{n+1} \in \mathbb{R}^d$ from the distribution of the features as either 0 or 1. Ideally this rule would choose the class that maximizes $\mathbb{P}(a_{n+1}, \cdot)$. One approach to find such a rule is called *support vector*

machine (SVM), which is an algorithm for constructing a hyperplane that seperates the two subsets $\{\, a \in \mathbb{R}^d \mid (a,0) \in D \,\}$ and $\{\, a \in \mathbb{R}^d \mid (a,1) \in D \,\}$ in a way that maximizes the so-called "margin", which is defined as the distance of the hyperplane to the closest point of either of the two subsets. New samples are then classified based on which side of the hyperplane they lie in. Whether or not such a hyperplane even exists depends on the nature of the dataset. Nevertheless, we can always write down the problem of finding such a hyperplane as a convex optimization problem with affine constraints:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \ \frac{1}{2}\|x\|^2 \\ &\text{s.t.} \quad b_i \langle a_i, x \rangle \geq 1 \quad \text{for all } i \in \{1, \dots, n\}. \end{aligned} \tag{1.2}$$

If a solution to the above problem exists, we say that the dataset is *linearly seperable*. In applications, the above problem is usually relaxed to the unconstrained problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x\|^2 + C \sum_{i=1}^{n} \max(0, 1 - b_i \langle a_i, x \rangle),$$

where $C \in (0, \infty)$. This problem always has a solution, regardless of whether or not the dataset is actually linearly seperable. However, the quality of its solution for the classification task highly depends on the choice of C [Has+04]. In particular, if C is too small, the resulting classifier will perform poorly. The methods we develop in chapter 3 will effectively allow us to directectly tackle the original problem (1.2) by letting $C \to \infty$ and thus we can circumvent the difficulty of choosing a suitable C. This application is presented in section 5.2.

We will come back to these two examples in chapter 5. Our goal is to develop methods to solve problems like ($\mathrm{P}^{\mathrm{new}}$). The difficulty here is that we are essentially dealing with infinitely many constraints, if the support of $x \mapsto A(\xi)x + b(\xi)$ is infinite. However, even in the case of finite support there are difficulties, as the number of constraints might still be so large that popular methods like projected stochastic gradient descent (see [Nem+09]) can become infeasible, as is the case in large-scale machine learning problems. Here, each realizatixon of the random variable ξ would correspond to one data point of a dataset that can be too large to be processed all at once. Instead, one is forced to process the data points one by one or in batches. One method to address this problem is *random constraint projections* TODO ... In this work, we will focus on a different method, which is based on a well-known method to deal with constraints indirectly by penalizing infeasible points: Instead of solving ($\mathrm{P}^{\mathrm{new}}$) directly, one instead solves a sequence of unconstrained problems of the form

$$\min_{x \in \mathbb{R}^d} \left\{ f^k(x) := f(x) + \gamma_k \pi^k(x) \right\}, \tag{1.3}$$

where $(\gamma_k)_{k \in \mathbb{N}}$ is an unbounded sequence of positive numbers, and, for all $k \in \mathbb{N}$, the function $\pi^k : \mathbb{R}^d \to [0, \infty)$ is a function that is meant to penalize infeasible points, i. e. we want $\pi^k(x)$ to be small when x is feasible, and large otherwise. We will refer to the parameters $(\gamma_k)_{k \in \mathbb{N}}$ as *penalty parameters* and the functions π^k as *penalty functions*. Assuming that π^k is convex for all $k \in \mathbb{N}$, there always exists a unique solution $x_k^\star$ of (1.3). Then, a natural question to ask is:

Under which conditions does the sequence of minimizers $(x_k^\star)_{k \in \mathbb{N}}$, corresponding to the sequence of functions $(f^k)_{k \in \mathbb{N}}$ defined in (1.3), converge to the solution $x^\star$ of ($\mathrm{P}^{\mathrm{new}}$)?

If $\lim_{k \to \infty} x_k^\star = x^\star$, we call the sequence $(x_k^\star)_{k \in \mathbb{N}}$ *consistent*. Unsurprisingly, consistency depends on the properties of $(\pi^k)_{k \in \mathbb{N}}$ and $(\gamma_k)_{k \in \mathbb{N}}$, as well as $A(\xi)$ and $b(\xi)$ to establish existence of $x^\star$. We will investigate consistency in section 3.1.

We will highlight two special examples of penalty functions that satisfy the conditions needed for

consistency.

Example 1.3 (Square-hinge penalty)**.** We let $\pi^k \equiv \pi_{\text{hin}}$ for all $k \in \mathbb{N}$, with

$$\pi_{\text{hin}}(x) := \frac{1}{2}\mathbb{E}\left(\sum_{i=1}^{m} \max(0, \langle A_i(\xi), x\rangle + b_i(\xi))^2\right),$$

and we refer to π_{hin} as the **square-hinge penalty**. Here, $A_i(\xi)$ dentotes the i-th row of $A(\xi)$ and $b_i(\xi)$ denotes the i-th row of $b(\xi)$. This penalty is convex and smooth (see TODO). Smoothness is highly desirable for convergence arguments, however, a disadvantage introduced by smoothness is the fact that infeasible points that are close to being feasible are barely penalized. As a consequence, while $\lim_{k\to\infty} x_k^\star = x^\star$ can indeed be guaranteed, in general none of the iterates $(x_k^\star)_{k\in\mathbb{N}}$ will be feasible.

Example 1.4 (Huber-like penalty)**.** This penalty appears in [NT20]. Let $(\delta_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers. For $k \in \mathbb{N}$, we define the sequence of **Huber-like penalties** as

$$\pi_{\text{hub}}^k(x) := \mathbb{E}\left(\sum_{i=1}^{m} \pi_{\delta_k}(x; A_i(\xi), b_i(\xi))\right),$$

where, as in the previous example, $A_i(\xi)$ is the i-th row of $A(\xi)$, $b_i(\xi)$ is the i-th row of $b(\xi)$, and

$$\pi_\delta(x; a, b) := \begin{cases} \frac{\langle a,x\rangle + b}{\|a\|} & \text{if } \langle a, x\rangle + b > \delta, \\ \frac{(\langle a,x\rangle + b + \delta)^2}{4\delta\|a\|} & \text{if } -\delta \le \langle a, x\rangle + b \le \delta, \\ 0, & \text{if } \langle a, x\rangle + b < -\delta, \end{cases}$$

for $\delta \in (0, \infty)$, $x, a \in \mathbb{R}^d$, and $b \in \mathbb{R}$. This penalty has some nice properties: It is convex, smooth, and has uniformly bounded gradients, $\sup_{k\in\mathbb{N}}(\sup_{x\in\mathbb{R}^d} \nabla\pi^k(x)) < \infty$. On top of this, the sequence $(\pi^k)_{k\in\mathbb{N}}$ converges to the penalty $x \mapsto \sum_{i=1}^m \max(0, A_i(\xi)x + b_i(\xi))$, which has the highly desirable property of being an *exact* penalty: TODO ...

Related literature. Penalty methods have been analyzed extensively in the deterministic case. See for example [WN+99; Ber97], where this method is implemented using a two-loop approach that proceeds as follows: In each iteration $k \in \mathbb{N}$, one solves the unconstrained problem $\min_{x\in\mathbb{R}^d} f^k(x)$ with some standard method like gradient descent or newton's method. If one is satisfied with the resulting solution, the algorithm stops. Otherwise, one repeats the process for some new penalty parameter $\gamma_{k+1} > \gamma_k$.
Outline.
Contributions. Single-loop penalty methods, batch sizes, relaxed gradient bound, general treatment of penalty functions, analysis of averaging, analysis of iterate moving averages.

2 Preliminaries

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 [BV04; GG23]. 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 $f\colon \mathbb{R}^d \to \mathbb{R}$ 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 $f\colon \mathbb{R}^d \to \mathbb{R}$ 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 $f\colon \mathbb{R}^d \to \mathbb{R}$ be L_f-smooth and let $g\colon \mathbb{R}^d \to \mathbb{R}$ be L_g-smooth and define $h(x) := a\,f(x) + b\,g(x)$ 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 $f\colon \mathbb{R}^d \to \mathbb{R}$ 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.

The following two propositions together imply that inequality constraints of the form present in ($\mathrm{P}^{\mathrm{new}}$) induce a convex feasibility set, provided the constraint defining map g has the property that $x \mapsto g(x, y)$ is convex for all $y \in \mathbb{R}^m$.

Proposition 2.5. *Let $I \subset \mathbb{R}$ be a (possibly infinite) set and let $\{C_i\}_{i\in I}$ be a family of convex subsets of $\mathbb{R}^d$. Then the (possibly infinite) intersection $\cap_{i\in I} C_i$ is a convex subset of $\mathbb{R}^d$.*

Proposition 2.6. *For $p \in \mathbb{N}$ and $i \in \{1, \ldots, p\}$, let $g_i\colon \mathbb{R}^d \to \mathbb{R}$ be convex functions and define $g(x) := (g_1(x), \ldots, g_p(x))^\top \in \mathbb{R}^p$ for $x \in \mathbb{R}^d$. Then, the set*

$$\left\{\, x \in \mathbb{R}^d \mid g(x) \leq 0 \right\}$$

is a convex subset of $\mathbb{R}^d$.

Together, the last two propositions yield the following.

Proposition 2.7. *Let $Y \subset \mathbb{R}^m$ be an arbitrary subset. Further, let $p \in \mathbb{N}$ and for all $i \in \{1, \dots, p\}$, let $g_i \colon \mathbb{R}^d \times Y \to \mathbb{R}$ be a function, such that $x \mapsto g_i(x, y)$ is convex for all $y \in Y$. Further, define $g(x, y) := (g_1(x, y), \dots, g_p(x, y))^\top \in \mathbb{R}^p$ for $(x, y) \in \mathbb{R}^d \times Y$. Then, the set*

$$\left\{ x \in \mathbb{R}^d \mid g(x, y) \leq 0 \text{ for all } y \in Y \right\}$$

is a convex subset of $\mathbb{R}^d$.

Proposition 2.8. *Every convex function on $\mathbb{R}^d$ is continuous.*

Definition 2.9. TODO: Remove subgradients, as we will only consider smooth functions. Let $f \colon \mathbb{R}^d \to \mathbb{R}$ 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 2.10. *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ 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 2.11. *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ 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 2.12. Let $f \colon \mathbb{R}^d \to \mathbb{R}$ 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 2.13. *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ and $g \colon \mathbb{R}^d \to \mathbb{R}$ 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 2.14. *Let $f \colon \mathbb{R} \to \mathbb{R}$ and $g \colon \mathbb{R} \to \mathbb{R}$ be convex and nondecreasing. Then the composition $f \circ g$ is also convex.*

Proposition 2.15. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ 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 $x^\star = \arg\min_{x\in\mathbb{R}^d} f(x)$, we have

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

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

Proposition 2.16. *Let $g\colon \mathbb{R}^d \to \mathbb{R}$ 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 2.17. *If $f\colon \mathbb{R}^d \to \mathbb{R}$ 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 $x^\star \in C$ such that $f(x^\star) \leq f(x)$ for all $x \in C$.*

Proposition 2.18. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ 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 2.19. Examples of (strongly) convex functions include

- affine function;
- quadratic functions $f(x) := x^\top A x + b^\top x + c$ 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.
- $x \mapsto \exp(x)$, $x \mapsto -\log(x)$, $x \mapsto \max(0, x)$.

Definition 2.20. Let $C \subset \mathbb{R}^d$ be a closed convex set. The **projection onto C** is the map $\Pi_C\colon \mathbb{R}^d \to C$, defined by

$$\Pi_C(x) := \inf_{y\in C} \|x - y\|_2,$$

where $\|\cdot\|_2$ is the standard euclidian norm on $\mathbb{R}^d$.

Definition 2.21. Let $C \subset \mathbb{R}^d$ be a closed convex set. The **outword normal cone** to C at $x \in C$, denoted by $N_C(x)$, is defined as

$$N_C(x) := \{\, v \in \mathbb{R}^d \mid \langle v, x - y\rangle \geq 0 \ \ \forall y \in C \,\}.$$

Proposition 2.22 (TODO: Cite)**.** *Let $C \subset \mathbb{R}^d$ be a closed convex set and let $x \in C$. Then*

1. *$x - \Pi_C(x) \in N_C(x)$;*
2. *If $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable and constant on C, it holds that*

$$\nabla f(x) \in N_C(\Pi_C(x))$$

for all $x \in \mathbb{R}^d$. If f is merely subdifferentiable, but convex, then

$$\partial f(x) \subset N_C(\Pi_C(x))$$

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

2.2 Probability Theory

The contents of this section can be found in standard probability texts, for example [Dur19; BH19].

Definition 2.23. 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 $B \backslash A \in \mathcal{F}$.
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 $(\Omega, \mathcal{F})$ is called a **measurable space**. For any subset $\mathcal{G} \subset 2^\Omega$, we define the **$\sigma$-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 2.24. 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 2.25. Let $(\Omega, \mathcal{F})$ and $(E, \mathcal{G})$ be measurable spaces. A map $f \colon \Omega \to E$ 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 2.26. Let $(\Omega, \mathcal{F})$ be a measurable space. A map $\mu : \mathcal{F} \to [0, \infty]$ is called a **measure on** $(\Omega, \mathcal{F})$ 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 $(\Omega, \mathcal{F})$, then the triplet $(\Omega, \mathcal{F}, \mu)$ is called a **measure space**.

Definition 2.27. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f \colon \Omega \to \mathbb{R}$ be measurable. We define the **(μ-)integral** of f, denoted $\int f \, \mathrm{d}\mu$, 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
$$\int f \, \mathrm{d}\mu := \sum_{i=1}^n c_i \mu(A_i).$$
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
$$\int f \, \mathrm{d}\mu := \sup_{g \in \mathcal{S}(\Omega),\, g \leq f} \int g \, \mathrm{d}\mu.$$
3. If f is neither a simple function, nor nonnegative, but $\int |f| \, \mathrm{d}\mu < \infty$, then we define
$$\int f \, \mathrm{d}\mu := \int \max(0, f) \, \mathrm{d}\mu - \int \max(0, -f) \, \mathrm{d}\mu.$$

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 2.28. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f\colon \Omega \to \mathbb{R}$ and $g\colon \Omega \to \mathbb{R}$ be integrable. Then*

(i) $\int f + g \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu + \int g \,\mathrm{d}\mu$.

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

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

Definition 2.29. Let $(\Omega, \mathcal{F})$ be a measurable space. If $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a measure on $(\Omega, \mathcal{F})$, 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 2.30. 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 2.31. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{G})$ a measurable space. A map $X\colon \Omega \to E$ 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 2.32. 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 2.33. 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 $\mathbb{E}(X) := \int X \,\mathrm{d}\mathbb{P}$.

The following three properties will be used multiple times throughout this text without explicit mention. They follow directly from proposition 2.28.

Proposition 2.34. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ be integrable random variables. Then*

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

(ii) $\mathbb{E}(cX) = c\,\mathbb{E}(X)$ *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 2.35. *Let $f\colon \mathbb{R}^d \times \Omega \to \mathbb{R}$ 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 2.36 (Jensen's inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathbb{R}$ 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 2.37. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathbb{R}$ be a random variable. If X^2 is integrable, we define the **variance of** X, denoted by $\mathbb{V}\mathrm{ar}(X)$, as $\mathbb{V}\mathrm{ar}(X) := \mathbb{E}\|X - \mathbb{E}(X)\|^2$.

One fact from probability theory is that, for any integrable random variable $X : \Omega \to E$, it holds that $\int X \,\mathrm{d}\mathbb{P} = \int I \,\mathrm{d}\mathbb{P}^X$, where $I\colon E \to E$ 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 2.38. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, \dots, X_n\colon \Omega \to \mathbb{R}$ be independent random variables, such that X_i^2 is integrable for all $i \in \{1, \dots, n\}$. Then $\mathbb{V}\mathrm{ar}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \,\mathbb{V}\mathrm{ar}(X_i)$, for any $a_1, \dots, a_n \in \mathbb{R}$. If, additionally, they are all identically distributed, it holds that $\mathbb{V}\mathrm{ar}(1/n \sum_{i=1}^n X_i) = \mathbb{V}\mathrm{ar}(X_1)/n$.*

Proposition 2.39. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathbb{R}$ be a random variable such that X^2 is integrable. Then $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.*

Definition 2.40. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathbb{R}^n$ be a random variable. Further, let $\mathcal{C} \subset \mathcal{F}$ be a σ-algebra. We call a random variable $Z\colon \Omega \to \mathbb{R}^n$ 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
$$\int_C Z \,\mathrm{d}\mathbb{P} = \int_C X \,\mathrm{d}\mathbb{P}.$$

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

If $Y\colon \Omega \to \mathbb{R}^m$ 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 $E(X \,|\, Y)$ 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 2.41. 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 probabiliy space $(\Omega, \mathcal{F}, \mathbb{P}_{Y=y})$, where $\mathbb{P}_{Y=y}(A) := \mathbb{P}(A \,|\, Y = y)$. More precisely, $\mathbb{E}(X \,|\, Y = y) = \int X \,\mathrm{d}\mathbb{P}_{Y=y}$. It follows that $\mathbb{E}(X \,|\, Y = \cdot)$ inherits all basic properties of $\mathbb{E}(\cdot)$ from proposition 2.34.

Below, we state some special properties of conditional expectations.

Proposition 2.42 (Properties of $\mathbb{E}(X \,|\, Y)$)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathbb{R}^n$, $Y\colon \Omega \to \mathbb{R}^m$ be integrable random variables. Further, let $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$ such that $\omega \mapsto F(x, \omega)$ is $\mathcal{F}$-measurable for all $x \in \mathbb{R}^d$ and let $G : \mathbb{R}^n \to \mathbb{R}$ be Borel measurable. Then,*

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

(ii) $\mathbb{E}(X \,|\, X) = X$.

(iii) *if $F(x, \cdot) \le G(x)$ a. s. for all $x \in \mathbb{R}^d$, it follows that $\mathbb{E}(F(X, \cdot) \,|\, X) \le G(X)$.*

Definition 2.43. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X\colon \Omega \to \mathbb{R}^n$, $Y\colon \Omega \to \mathbb{R}^m$ be integrable random variables. The **conditional variance of X given Y**, denoted by $\mathbb{V}\mathrm{ar}(X \,|\, Y)$, is defined as

$$\mathbb{V}\mathrm{ar}(X \,|\, Y) := \mathbb{E}\Big(||X - \mathbb{E}(X \,|\, Y)||^2 \,\Big|\, Y\Big).$$

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

$$\mathbb{V}\mathrm{ar}(X \,|\, Y = y) := \mathbb{E}\Big(||X - \mathbb{E}(X \,|\, Y = y)||^2 \,\Big|\, Y = y \Big).$$

It holds that $\mathbb{V}\mathrm{ar}(X \,|\, Y = Y(\omega)) = \mathbb{V}\mathrm{ar}(X \,|\, Y)(\omega)$ for all $\omega \in \Omega$.

Remark 2.44. 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 $\mathbb{V}\mathrm{ar}(X \,|\, Y = y)$ can be thought of as the variance of X on a different probability space $(\Omega, \mathcal{F}, \mathbb{P}_{Y=y})$ (see remark 2.41). Hence, all basic properties of $\mathbb{V}\mathrm{ar}(\cdot)$ are inherited by the conditional variance. In particular, if $X_1, \dots, X_n$ are i. i. d. random variables, it holds that $\mathbb{V}\mathrm{ar}(X_1 + \cdots + X_n \,|\, Y) = 1/n\, \mathbb{V}\mathrm{ar}(X_1 \,|\, Y)$ – a fact that will be used later.

Proposition 2.45. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X \colon \Omega \to \mathbb{R}^n$ be an integrable random variable. Further, let $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$ 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 $G : \mathbb{R}^n \to \mathbb{R}$ be Borel measurable. Then, if $\mathbb{V}\mathrm{ar}(F(x, \cdot)) \leq G(x)$ for all $x \in \mathbb{R}^d$, it also holds that*

$$\mathbb{V}\mathrm{ar}(F(X, \cdot) \,|\, X) \leq G(X).$$

2.3 Stochastic Optimization

Definition 2.46 ([Duc18])**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $f \colon \mathbb{R}^d \to \mathbb{R}$ be a convex function. A random vector $G \colon \Omega \to \mathbb{R}^d$ 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 2.47. Let $F \colon \mathbb{R}^d \times \Omega \to \mathbb{R}$ 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 $G_x \colon \Omega \to \mathbb{R}^d$, defined by $G_x(\omega) := \nabla_x F(x, \omega)$, is a stochastic gradient of f at x.

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

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

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

- $x \to F(x, \xi(\omega))$ is convex for almost every $\omega \in \Omega$;
- $\omega \to F(x, \xi(\omega))$ 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 Gradiend 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 [RM51]. The convergence of the iterates $(x_k)_{k\in\mathbb{N}}$, generated by algorithm 1, to a minimum $x^\star \in \arg\min_{x\in\mathbb{R}^d} f(x)$ (if it exists) depends heavily on the choice of step sizes $(\tau_k)_{k\in\mathbb{N}}$ and the behavior of $\mathbb{E}\|G^k(x_k)\|^2$. If f is strongly convex and 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 2.48. *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 2.13 and example 2.19), and thus the above minimization problem is well defined (proposition 2.17).

In most applications, SGD is used with *iterate averaging*. In [Nem+09], it was shown that the averages $\bar{x}_k := \sum_{i=1}^k x_k$ efficiently converge in the case of convex objectives. Another popular averagins 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

We will almost exclusively deal with the standard euclidian norm $x \mapsto \|x\|_2$ on $\mathbb{R}^d$, defined by

$$\|x\|_2 := \sqrt{x_1^2 + \cdots + x_d^2}.$$

Besides the euclidian norm, we will also need the *Frobenius norm* $A \mapsto \|A\|_F$, defined by

$$\|A\|_F := \sqrt{\sum_{i=1}^d \sum_{j=1}^d A_{ij}^2}$$

on the space of $d \times d$ matrices $A = (A_{ij})_{i,j\in\{1,\dots,d\}}$. For any norm $\|\cdot\|$, a property we will make frequent use of is the triangle inequality $\|x + y\| \leq \|x\| + \|y\|$ $\forall x, y \in \mathbb{R}^d$. Beyond this, we will often make use of the inequality $\|x + y\|^2 \leq 2(\|x\|^2 + \|y\|^2)$ $\forall x, y \in \mathbb{R}^d$. This inequality follows from the triangle inequality and a special case of the very useful *Young's inequality*.

Proposition 2.49 (Young's inequality)**.** *For all* $\alpha, \beta \in \mathbb{R}$ *and all* $\epsilon \in (0,\infty)$*, it holds that*

$$\alpha\beta \leq \frac{\alpha^2}{2\epsilon} + \frac{\epsilon\beta^2}{2}.$$

This general form of Young's inequality is particularly useful in combination with the well-known *Cauchy-Schwarz inequalty* to deal with inner products that are otherwise hard to analze. For $x, y \in \mathbb{R}^d$, we let $\langle x, y\rangle := x^\top y$ denote the standard inner product on $\mathbb{R}^d$.

Proposition 2.50 (Cauchy-Schwarz inequality)**.** *For all $x, y \in \mathbb{R}^d$, it holds that $|\langle x, y\rangle| \leq \|x\|_2 \cdot \|y\|_2$.*

An important result that connects the two norms $\|\cdot\|_2$ and $\|\cdot\|_F$ is the following.

Proposition 2.51. *For all $A \in \mathbb{R}^{d\times d}$ and $x \in \mathbb{R}^d$, it holds that $\|Ax\|_2 \leq \|A\|_F \cdot \|x\|_2$.*

3 Penalty Methods for Smooth Objectives

Throughout this chapter, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All maps $\mathbb{R}^n \to \mathbb{R}^m$, $n, m \in \mathbb{N}$, are implicitly considered to be measurable with respect to the corresponding Borel σ-algebras on $\mathbb{R}^n$ and $\mathbb{R}^m$ (example 2.24). We consider the unconstrained problems

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

for which we make the following standing assumption.

Standing assumption 1. Throughout this chapter, we make the following assumptions about ($\mathrm{P}^{\mathrm{new}}$) and ($\mathrm{P}^k$):

1. There exists $\mu \in (0, \infty)$ such that f is μ-strongly convex;
2. The sequence $(\gamma_k)_{k \in \mathbb{N}}$ is positive and unbounded;
3. There exists at least one feasible point for ($\mathrm{P}^{\mathrm{new}}$);
4. $(\pi^k)_{k \in \mathbb{N}}$ is a sequence of nonnegative convex functions, such that $\gamma_{k+1}\, \pi^{k+1}(x) \geq \gamma_k\, \pi^k(x)$ for all $x \in \mathbb{R}^d$ and $k \in \mathbb{N}$ (TODO: NOT NEEDED ACTUALLY, assume instead existence of a sequence c_k such that $\pi^k(x) \leq c_k$ for all feasible x, and $\gamma_k c_k \to 0$), and $\lim_{k \to \infty} \gamma_k \pi^k(x) = 0$ for all feasible $x \in \mathbb{R}^d$. Moreover, we assume the existence of a continuous function $\pi^\infty \colon \mathbb{R}^d \to [0, \infty)$ such that $\pi^\infty \leq \pi^k$ for all $k \in \mathbb{N}$ and $\pi^\infty(x) = 0$ if and only if x is feasible for ($\mathrm{P}^{\mathrm{new}}$).

Note that the final assumption is always satisfied if $\pi^k \equiv \pi$ for all $k \in \mathbb{N}$ and some convex $\pi \colon \mathbb{R}^d \to [0, \infty)$ that satisfies $\pi(x) = 0$ if and only if x is feasible (continuity follows from proposition 2.8). Hence, the square hinge penalty from example 1.3 satisfies the assumptions. TODO: Huber-like penalty explanation. An implication of standing assumption 1 is that the optimization problems ($\mathrm{P}^{\mathrm{new}}$) and ($\mathrm{P}^k$) each have unique solutions. This follows from proposition 2.17. We will denote the solution of ($\mathrm{P}^{\mathrm{new}}$) as $x^\star$, and, for $k \in \mathbb{N}$, we write $x_k^\star$ for the solution of (P^k).

3.1 Consistency of Solutions

In this section, we will prove that the standing assumption 1 is enough to ensure convergence of the sequence $(x_k^\star)_{k \in \mathbb{N}}$ to to $x^\star$. This is useful for two purposes. First, it guides us in the choice for reasonable penalty functions $(\pi^k)_{k \in \mathbb{N}}$. Second, having established convergence, we know that $(x_k^\star)_{k \in \mathbb{N}}$ is a bounded sequence. This will be used multiple times in later sections, where we derive convergence *rates* of iterative methods to further guide us towards optimal choices of parameters for these methods, including the penalty parameters $(\gamma_k)_{k \in \mathbb{N}}$.

Theorem 3.1. *It holds that $x_k^\star \to x^\star$ and $f^k(x_k^\star) \to f(x^\star)$ as $k \to \infty$.*

First, some preparation.

Lemma 3.2. *Let $h \colon \mathbb{R}^d \to \mathbb{R}$ be a coercive and continuous function, and let $X \subset \mathbb{R}^d$ be nonempty and closed. Then h attains a minimum over X, i. e. there exists $x^\star \in X$ such that $h(x^\star) = \inf_{x \in X} h(x)$.*

Proof. Let $x_0 \in X$. Since h is coercive, there exists $r > 0$ such that $h(x) \geq h(x_0)$ for all $x \in \mathbb{R}^d$ with $\|x\| > r$, therefore any minimum of h – 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

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

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

Lemma 3.3. *Let $h \colon \mathbb{R}^d \to \mathbb{R}$ be strongly convex and differentiable. Then h is coercive.*

Proof. Let $x^\star$ be the minimizer of h (proposition 2.17). By proposition 2.15, we have

$$h(x) \geq h(x^\star) + \frac{\mu}{2}\|x - x^\star\|^2$$

for all $x \in \mathbb{R}^d$ and some $\mu \in (0, \infty)$. Letting $\|x\| \to \infty$, we see that $h(x) \to \infty$. □

Lemma 3.4. *Let $h \colon \mathbb{R}^d \to \mathbb{R}$ be a coercive function If $\{\, h(u_k) \mid k \in \mathbb{N} \,\}$ is bounded for some sequence $(u_k)_{k \in \mathbb{N}} \subset \mathbb{R}^d$, then $(u_k)_{k \in \mathbb{N}}$ must also be bounded.*

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 $\|u_{k_r}\| \to \infty$ for $r \to \infty$. By coercivity, this would imply that $\lim_{k \to \infty} h(u_k) = \infty$, contradicting our assumption that $\{\, h(u_k) \mid k \in \mathbb{N} \,\}$ is bounded. Hence, $(u_k)_{k \in \mathbb{N}}$ must be bounded. □

Lemma 3.5. *Let $U := \{\, u_k \mid k \in \mathbb{N} \,\}$ 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 $u_k \to u$ for $k \to \infty$.*

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

$$\|u_{k_r} - u\| \geq \epsilon \tag{3.1}$$

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

We will now prove the main theorem.

Proof of Theorem 3.1. Broadly, the argument is based on the proof of proposition 4.8 in [GH22]. A notable difference is the fact that we allow the penalty function to vary over time. From strong convexity of f, convexity of π^k, and $\gamma_k > 0$, it follows that f^k is also strongly convex for all $k \in \mathbb{N}$ (proposition 2.13). Thus, for every $k \in \mathbb{N}$, proposition 2.17 implies that there exists a unique solution $x_k^\star$ to problem (P^k). Let x be any feasible point for ($\mathrm{P}^{\mathrm{new}}$), then, for any $k, r \in \mathbb{N}$,

$$f(x_k^\star) \leq f^k(x_k^\star) \leq f^k(x) \leq f^{k+r}(x) = f(x) + \gamma_{k+r}\pi^{k+r}(x),$$

where the last inequality follows from the fact that $\gamma_{k+r}\pi^{k+r} \geq \gamma_k \pi^k$, which in turn follows from one of our assumptions on $(\pi^k)_{k \in \mathbb{N}}$. Letting $r \to \infty$, we find that

$$f(x_k^\star) \leq f^k(x_k^\star) \leq f^k(x) \leq f(x) \tag{3.2}$$

for all $k \in \mathbb{N}$ and any feasible $x \in \mathbb{R}^d$. In particular, $(f^k(x_k^\star))_{k \in \mathbb{N}}$ and $(f(x_k^\star))_{k \in \mathbb{N}}$ are bounded sequences (boundedness from below follows directly from coercivity). It follows from coercivity of f (lemmas 3.3

and 3.4) that the sequence $(x_k^\star)_{k\in\mathbb{N}}$ is also bounded and therefore it contains a subsequence $(x_{k_r}^\star)_{r\in\mathbb{N}}$ that converges to a point $x_\infty^\star \in \mathbb{R}^d$. For any $k \in \mathbb{N}$, we have

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

This implies that $(f^k(x_k^\star))_{k\in\mathbb{N}}$ is a monotonically increasing sequence. Together with the fact that $(f^k(x_k^\star))_{k\in\mathbb{N}}$ is bounded, we therefore know that $(f^k(x_k^\star))_{k\in\mathbb{N}}$ must converge. In particular, we have

$$\limsup_{k\to\infty} f^k(x_k^\star) - f(x_k^\star) < \infty.$$

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

$$\limsup_{k\to\infty} \gamma_k \pi^k(x_k^\star) = \lim_{k\to\infty} \gamma_k \pi^k(x_k^\star) < \infty,$$

and since $\lim_{k\to\infty} \gamma_k = \infty$, it must therefore hold that

$$\lim_{k\to\infty} \pi^k(x_k^\star) = 0.$$

By our assumption, there exists a continuous function $\pi^\infty \colon \mathbb{R}^d \to [0,\infty)$ such that $\pi^\infty(x) = 0$ if and only if x is feasible, and $\pi^\infty \leq \pi^k$ for all $k \in \mathbb{N}$. Therefore,

$$\pi^\infty(x_\infty^\star) = \lim_{k\to\infty} \pi^\infty(x_k^\star) \leq \lim_{k\to\infty} \pi^k(x_k^\star) = 0,$$

which proves that $x_\infty^\star$ is feasible. To prove optimality of $x_\infty^\star$, let $x^\star$ be the solution to ($\mathrm{P}^{\mathrm{new}}$). Then we have, again from (3.2),

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

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

$$\lim_{k\to\infty} f^k(x_k^\star) = f(x^\star),$$

as desired. Finally, by uniqueness of $x^\star$ we must have $x_\infty^\star = x^\star$, proving that $x^\star$ is a limit point of $(x_k^\star)_{k\in\mathbb{N}}$. Note that the assumptions on $(\gamma_k)_{k\in\mathbb{N}}$ still hold if we replace $(\gamma_k)_{k\in\mathbb{N}}$ by any subsequence $(\gamma_{k_r})_{r\in\mathbb{N}}$, and the same arguments imply that $x^\star$ is also a limit point of the resulting subsequence $(x_{k_r}^\star)_{r\in\mathbb{N}}$. Hence, by lemma 3.5, we do in fact have $\lim_{k\to\infty} x_k^\star = x^\star$. □

3.2 Sequential SGD

We will now analyze a form of stochastic gradient descent to efficiently solve ($\mathrm{P}^{\mathrm{new}}$).

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

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

where

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

$(\xi_i^j)_{i=1,\dots,k, j=1,\dots b_k}$ are i. i. d. samples from the distribution of ξ and $\tilde{\nabla} f^k(x, \xi)$ is a stochastic gradient 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 $x^\star$ be the solution to ($\mathrm{P}^{\mathrm{new}}$). Our goal is to to determine appropriate parameters τ_k, γ_k and $b_k \in \mathbb{N}$, such $\mathbb{E}\|x_k - x^\star\|$ 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 ($\mathrm{P}^{\mathrm{new}}$), but from the surrogate objective (P^k). In addition, this surrogate depends on γ_k, which may need to satisfy $\gamma_k \to \infty$ – that is, the surrogate objective changes between iterations. Further, the squared gradient norm $\mathbb{E}\|\tilde{\nabla} f^k(x, \xi)\|^2$ grows quadratically in γ_k, 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 [Nem+09], for example. Our plan of attack involves first decomposing $\mathbb{E}\|x_k - x^\star\|$ as follows:

$$\mathbb{E}\|x_k - x^\star\| \leq \mathbb{E}\|x_k - x_k^\star\| + \|x_k^\star - x^\star\|, \tag{3.3}$$

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. Now there are two terms we need to bound: $\mathbb{E}\|x_k - x_k^\star\|$ and $\|x_k^\star - x^\star\|$. We refer to the former as the **tracking error** and the latter as the **surrogate error.** We will refer to the following additional assumptions. To analyze the tracking error, we will use tools from online optimization. In particular, we adapt proof techniques from [CDH23] to bound this term. The surrogate error is analyzed in a case-by-case basis, depending on the penalty sequence $(\pi^k)_{k\in\mathbb{N}}$. A central tool here is an infinite-dimensional version of Hoffman's lemma [Hof03].

In the following, all statements involving random variables are understood to hold almost surely, unless stated otherwise.

Assumption 1. 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 $M \in \mathbb{R}^{n\times n}$. TODO: Define this norm in the norm section. TODO: Make sure $b(\xi)$ is regular enough, too.

Standing assumption 2. Along with standing assumption 1, we assume that the following hold in the following sections.

1. We can sample arbitrarily many independent random variables ξ_i^j, $i, j \in \mathbb{N}$, from the distribution of ξ.
2. The objective f is differentiable and L_f-smooth.
3. For all $k \in \mathbb{N}$, the penalty π^k is differentiable and $\|\nabla \pi^k(x) - \nabla \pi^k(y)\| \leq L_{\pi^k}\|x - y\| + G_\pi$ for all $x, y \in \mathbb{R}^d$ and positive constants $L_{\pi^k}, G_\pi \in (0, \infty)$.
4. The stochastic gradients of f, denoted by $\tilde{\nabla} f(\cdot, \xi)$, satisfy the variance bound

   $$\mathbb{V}\mathrm{ar}(\tilde{\nabla} f(x, \xi)) \leq \sigma_f^2(\|x\|^2 + 1)$$

   for some constant $\sigma_f^2 \in (0, \infty)$.
5. For all $k \in \mathbb{N}$, the stochastic gradients of π^k, denoted by $\tilde{\nabla} \pi^k(\cdot, \xi)$, satisfy the variance bound

   $$\mathbb{V}\mathrm{ar}(\tilde{\nabla} \pi^k(x, \xi)) \leq \sigma_{\pi^k}^2(\|x\|^2 + 1)$$

for some constant $\sigma^2_{\pi^k} \in (0, \infty)$.

3.2.1 Bounding the surrogate error

In this section, we will bound the surrogate error $\|x_k^\star - x^\star\|$. We will restrict our analysis to two the special cases of penalty functions introduced in examples 1.3 and 1.4. Throughout, we will take on the viewpoint of semi-infinite programming to analyze (P^{new}). Thus, we reframe problem (P^{new}) as

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} f(x) \\ &\text{s.t.} \quad A(z)x + b(z) \leq 0 \ \text{ for all } z \in \Xi, \end{aligned} \tag{P^{SIP}}$$

where $\Xi \subset \mathbb{R}^m$ is the support of ξ. Note that, as discussed in TODO, Ξ must not be unique in general, in which case the above problem would not be equivalent to (P^{new}). However, Ξ *is* unqiue if we restrict ourselves to random variabes ξ with compact support (TODO: Not enough. I think if you also assume continuity of $A(z)x + b(z)$ in z and that the distribution is "spread out" over Ξ, then it is enough). In that case,

$$A(z)x + b(z) \leq 0 \ \text{ for all } z \in \Xi \quad \iff \quad \mathbb{P}(A(\xi)x + b(\xi) \leq 0) = 1,$$

so the two formulations (P^{new}) and (P^{SIP}) would indeed be equivalent, then. This motivates the following assumptions.

Assumption 2. The random variable ξ is supported on a compact metric space $\Xi \subset \mathbb{R}^m$. The map $z \mapsto A(z)x + b(z)$ is continuous for all x, and ξ has a density function f^ξ such that $\inf_x f^\xi(x) > 0$.

Assumption 3.

Bounding the distance-to-feasibility

The key to bounding $\|x_k^\star - x^\star\|$ involves first bounding the distance-to-feasibility $\operatorname{dist}(x_k^\star, \mathcal{X})$ by the penalty $\pi^k(x_k^\star)$. In the case where Ξ is a set of finite size, this is immediately achieved by use of the well-known Hoffman inequality.

Lemma 3.6 (Hoffman's inequality)**.** *Let $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^m$, and $S(A, b) := \{\, x \in \mathbb{R}^n \mid Ax + b \leq 0 \,\}$. Further, we let $\|\cdot\|_n$, resp. $\|\cdot\|_m$, denote any norm on $\mathbb{R}^n$, resp. $\mathbb{R}^m$, and we define $\operatorname{dist}(x, S(A,b)) := \inf_{y \in S} \|x - y\|_n$. Then, there exists a constant $c \in (0, \infty)$, which depends on A, $\|\cdot\|_n$, and $\|\cdot\|_m$, such that*

$$\operatorname{dist}(x, S(A,b)) \leq c \, \|(Ax + b)_+\|_m,$$

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

Proof. See [Hof03]. □

If Ξ is finite, say $\Xi = \{1, \ldots, n\} \subset \mathbb{N}$, then we can write

$$\mathcal{X} = \bigcap_{i=1}^n \{\, x \in \mathbb{R}^d \mid A(i)x + b(i) \leq 0 \,\}.$$

Define A as the row-wise concatenation of the matrices $(A(i))_{i \in \{1,\ldots,n\}}$, and b as the row-wise concatenation of the vectors $(b(i))_{i \in \{1,\ldots,n\}}$. Then,

$$\mathcal{X} = \{\, x \in \mathbb{R}^d \mid Ax + b \leq 0 \,\}.$$

Now, by the above lemma, there exists a constant $c \in (0, \infty)$, such that for all $p \in \mathbb{N}$

$$\operatorname{dist}(x, \mathcal{X})^p \leq c^p \, \|(Ax + b)_+\|_\infty^p, \tag{3.4}$$

where $\|\cdot\|_\infty$ is the sup-norm TODO. For $i \in \Xi$, let $p_i := \mathbb{P}(\xi = i)$, and set $p_{\min} := \min_{i \in \{1,\dots,n\}} p_i$. Per definition of Ξ, it holds that $p_{\min} > 0$. We therefore obtain

$$\begin{aligned}
\|(Ax + b)_+\|_\infty^p &= \max_{i \in \{1,\dots,m \cdot n\}} ((Ax + b)_+^p)_i \\
&\leq \sum_{i=1}^{m \cdot n} ((Ax + b)_+^p)_i \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} ((A(i)x + b)_+^p)_j \\
&= \sum_{i=1}^{n} \|(A(i)x + b)_+\|_p^p \\
&= \frac{1}{p_{\min}} \sum_{i=1}^{n} p_{\min} \|(A(i)x + b)_+\|_p^p \\
&\leq \frac{1}{p_{\min}} \sum_{i=1}^{n} p_i \|(A(i)x + b)_+\|_p^p \\
&= \frac{1}{p_{\min}} \mathbb{E}(\|(A(\xi)x + b)_+\|_p^p).
\end{aligned}$$

Combining with (3.4), we therefore obtain

$$\operatorname{dist}(x, \mathcal{X})^p \leq \frac{c^p}{p_{\min}} \mathbb{E}(\|(A(\xi)x + b)_+\|_p^p)$$

for all $p \in \mathbb{N}$. In particular, there exist constants $c_1, c_2 \in (0, \infty)$, such that

$$\begin{aligned}
\operatorname{dist}(x_k^\star, \mathcal{X}) &\leq c_1 \, \pi_{\text{hub}}^k(x_k^\star) \text{ and} \\
\operatorname{dist}(x_k^\star, \mathcal{X})^2 &\leq c_2 \, \pi_{\text{hin}}(x_k^\star),
\end{aligned}$$

for all $k \in \mathbb{N}$, where we used the property of $(\pi_{\text{hub}}^k)_{k \in \mathbb{N}}$ proven in TODO.

Extending this result to the case of infinite support Ξ will require us to find an analogue of lemma 3.6. We will proceed in two stages: First, we will determine conditions such that there exists a constant $c_1 \in (0, \infty)$ such that

$$\operatorname{dist}(x, \mathcal{X})^p \leq c_1 \sup_{z \in \Xi} \|(A(z)x + b(z))_+\|_p^p.$$

To achieve this, we will use a result from [BT96]. In the next step, we will then need to constrain the distribution of ξ to ensure that there is "sufficient mass" on every region of the support Ξ (notice the analogy to the finite support case, where we used $\min_{i \in \{1,\dots,n\}} p_i > 0$). Then, measure theoretic calculations will allow us to show

$$\sup_{z \in \Xi} \|(A(z)x + b(z))_+\|_p^p \leq c_2 \, \mathbb{E}(\|(A(\xi)x + b(\xi))_+\|_p^p)$$

for some constant $c_2 \in (0, \infty)$. Putting the two bounds together, we can then arrive at the desired existence of $c \in (0, \infty)$ such that

$$\operatorname{dist}(x, \mathcal{X})^p \leq c \, \mathbb{E}(\|(A(\xi)x + b(\xi))_+\|_p^p)$$

for all $p \in \mathbb{N}$. For this, we will need two additional assumptions for which we first need to define some terms.

Definition 3.7. Let $f\colon \mathbb{R}^d \to \mathbb{R}$, $g\colon \mathbb{R}^d \times Z \to \mathbb{R}^p$ for $Z \subset \mathbb{R}^m$, and define the set $S := \{\, x \in \mathbb{R}^d \mid g(x,z) \leq 0 \text{ for all } z \in Z \,\}$. The optimization problem $\min_{x \in S} f(x)$ is said to satisfiy **Slater's condition** if there exists a point $x \in S$ such that $g(x,z) < 0$ for all $z \in Z$. Such a point is referred to as a **Slater point**.

We can now formulate the following assumption, which will help us later to establish strong duality of a certain optimization problem.

Assumption 4. Problem ($\mathrm{P}^{\mathrm{new}}$) satisfies Slater's condition.

Before we proceed, we will slightly reformulate ($\mathrm{P}^{\mathrm{SIP}}$) to match the setting in [BT96]. For $g\colon \Xi \to \mathbb{R}^p$, we define the norm $\|g\|_Y := \sup_{z \in \Xi} \|g(z)\|_1$ and use this to define the normed space

$$Y := \{\, y\colon \Xi \to \mathbb{R}^p \text{ continuous} \mid \|y\|_Y < \infty \,\}.$$

With some slight abuse of notation, we define the continuous linear map $A\colon \mathbb{R}^d \to Y$, by

$$(Ax)(z) := A(z)x$$

for $x \in \mathbb{R}^d$ and $z \in \Xi$, where $A(z) \in \mathbb{R}^{d \times p}$ is from the constraint in ($\mathrm{P}^{\mathrm{SIP}}$). Further, we define the closed convex cone

$$K := \{\, y \in Y \mid y(z) \leq 0 \text{ for all } z \in \Xi \,\}.$$

The feasible set in ($\mathrm{P}^{\mathrm{SIP}}$) can now be written as

$$\mathcal{X} = \{\, x \in \mathbb{R}^d \mid Ax + b \in K \,\},$$

where b is given in ($\mathrm{P}^{\mathrm{SIP}}$).

Theorem 3.8. *Let X and Y be normed spaces. Let $h\colon X \to \mathbb{R}$ be a convex function, whose conjugate $h^\star$ is nonnegative everywhere, let K be a nonempty closed convex cone in Y, let $A\colon X \to Y$ be a continuous linear operator, $-b \in \mathrm{im}(A) - K$, and let $x \in X$. Suppose that the following two conditions are satisfied:*
TODO: Define what $\langle\rangle$ is in this case.

1. $\inf_{Ax'+b \in K} h(x - x') = \sup_{y^\star \in K^\circ} \{\langle y^\star, Ax + b\rangle - h^\star(A^\star y^\star)\}$.
2. *There exists a cone $W \subset K^\circ$, independent of x, such that the above supremum is unchanged when K° is replaced by W, i. e.* $\sup_{y^\star \in K^\circ} \{\langle y^\star, Ax + b\rangle - h^\star(A^\star y^\star)\} = \sup_{y^\star \in W} \{\langle y^\star, Ax + b\rangle - h^\star(A^\star y^\star)\}$.

Then, there exists a constant $c \in (0, \infty)$, such that

$$\inf_{Ax'+b \in K} h(x - x') \leq c \operatorname{dist}_Y(Ax + b, K),$$

where $\|\cdot\|_Y$ is the norm on Y and $\operatorname{dist}_Y(y, K) := \inf_{y' \in K} \|y - y'\|_Y$ for all $y \in Y$.

Proof. See [BT96]. □

In the context of the above theorem, we consider $h(x) := \|x\|$. If we assume that the conditions of the above theorem are satisfied, we would then have

$$\inf_{Ax'+b \in K} \|x - x'\| \leq c \operatorname{dist}_Y(Ax + b, K),$$

for some constant $c \in (0, \infty)$. Note that $Ax' + b \in K \iff x' \in \mathcal{X}$, so the term on the left-hand side is just the (euclidian) distance from x to the feasible set. On the other hand, by definition of K and $\|\cdot\|_Y$, we have

$$\operatorname{dist}_Y(Ax + b, K) = \inf_{y \in K} \sup_{z \in \Xi} \|A(z)x + b(z) - y(z)\|_1 = \sup_{z \in \Xi} \|(A(z)x + b(z))_+\|_1,$$

where the second equality follows from the following argument: TODO... Thus, if the assumptions of theorem 3.8 hold, we can conclude

$$\operatorname{dist}(x, \mathcal{X}) \leq c \sup_{z \in \Xi} \|(A(z)x + b(z))_+\|_1.$$

We will now prove that Slater's condition is enough to guarantee that the assumptions of theorem 3.8 do indeed hold.

Lemma 3.9. *In the situation of theorem 3.8, let $h(x) := \|x\|$. If assumptions 2 and 4 hold, then*

1. $\inf_{Ax' + b \in K} h(x - x') = \sup_{y^\star \in K^\circ} \{\langle y^\star, Ax + b\rangle - h^\star(A^\star y^\star)\}$.
2. *There exists a cone $W \subset K^\circ$, independent of x, such that the above supremum is unchanged when K° is replaced by W, i. e. $\sup_{y^\star \in K^\circ} \{\langle y^\star, Ax + b\rangle - h^\star(A^\star y^\star)\} = \sup_{y^\star \in W} \{\langle y^\star, Ax + b\rangle - h^\star(A^\star y^\star)\}$.*

Proof. The first statement is another way of saying that strong duality holds. By Theorem 2.3 in [Sha09], strong duality holds if a) $\inf_{Ax' + b \in K} \|x - x'\| < \infty$, b) Ξ is a compact metric space, and c) $(x, z) \mapsto A(z)x + b(z)$ is continuous on $\mathbb{R}^d \times \Xi$. The former holds since K is assumed to be nonempty, while the remaining two conditions are implied by assumptions 2 and 4.

For the second statement, TODO... □

With lemma 3.9 and theorem 3.8 in hand, we now have easy-to-verify condtions under which

$$\operatorname{dist}(x_k^\star, \mathcal{X}) \leq c \sup_{z \in \Xi} \|(A(z)x_k^\star + b(z))_+\|_1$$

holds for some $c \in (0, \infty)$ and all $k \in \mathbb{N}$. It is now straight-forward to turn this into a lower-bound of the penalties π_{hub}^k and π_{hin}. Since the term on the right-hand side is finite (by continuity and compactness of Ξ) and uniformly bounded over all $k \in \mathbb{N}$ (by theorem 3.1), there must exist some constant c', independent of $k \in \mathbb{N}$, such that

$$\sup_{z \in \Xi} \|(A(z)x_k^\star + b(z))_+\|_1 \leq c' \, \mathbb{E}\|(A(\xi)x_k^\star + b(\xi))_+\|_1$$

for all $k \in \mathbb{N}$, provided that the expectation is non-zero if and only if the left-hand side is non-zero. This can be ensured by invoking assumptions 2 and 3: If $\mathbb{E}\|(A(\xi)x_k^\star + b(\xi))_+\|_1 = 0$, then $\|(A(z)x_k^\star + b(z))_+\|_1 = 0$ for almost all $z \in \Xi$. Assume there exists a $z \in \Xi$ such that $\|(A(z)x_k^\star + b(z))_+\|_1 > 0$. Then, by continuity and compactness, there must exist a whole neighborhood $\mathcal{N} \subset \Xi$ around z, such that $\|(A(z')x_k^\star + b(z'))_+\|_1 > 0$ for all $z' \in \mathcal{N}$. Further, by assumption on the density of ξ, it must hold that $P^\xi(\mathcal{N}) \geq \delta \operatorname{vol}(\mathcal{N}) > 0$, where $\operatorname{vol}(\mathcal{N})$ is the volume of the set $\mathcal{N}$, as measured by the Lebesgue measure. However, this would then imply

$$\mathbb{E}\|(A(\xi)x_k^\star + b(\xi))_+\|_1 \geq \int_{\mathcal{N}} \|(A(z)x_k^\star + b(z))_+\|_1 \, \mathbb{P}^\xi(dz) > 0,$$

which is a contradiction. On the other hand, if the left-hand side of the above inequality is zero, then so is the right-hand side, proving the claim that

$$\sup_{z \in \Xi} \|(A(z)x_k^\star + b(z))_+\|_1 = 0 \iff \mathbb{E}\|(A(z)x_k^\star + b(z))_+\|_1 = 0.$$

We summarize the preceeding discussion with the following result.

Lemma 3.10. *If assumptions 2 to 4 hold, then there exists a constant $c \in (0, \infty)$, such that*

$$\operatorname{dist}(x_k^\star, \mathcal{X}) \leq c\, \mathbb{E}\|(A(\xi)x_k^\star + b(\xi))_+\|_1$$

for all $k \in \mathbb{N}$. In particular, we have

$$\begin{aligned} \operatorname{dist}(x_k^\star, \mathcal{X}) &\leq \pi_{\text{hub}}^k(x_k^\star) \quad \text{and} \\ \operatorname{dist}(x_k^\star, \mathcal{X})^2 &\leq \pi_{\text{hin}}(x_k^\star) \end{aligned}$$

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

Proof. See the preceeding discussion. □

Theorem 3.11. *Let assumptions 2 to 4 hold. Then, for $\pi^k \equiv \pi_{\text{hub}}^k$, it holds that TODO: Need precise bound on $\pi_{\text{hub}}^k(x^\star)$, or just reformulate for general π^k that upper bounds relu.*

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 + \frac{\mu}{2}\|x^\star - x_k^\star\|^2 + (\gamma_k - L)\operatorname{dist}(x_k^\star, \mathcal{X}) \leq \gamma_k c_k,$$

for all $k \in \mathbb{N}$ and a constant $L \in (0, \infty)$. In particular,

$$\|x^\star - x_k^\star\|^2 = \mathcal{O}(\gamma_k c_k) \quad \text{and} \quad \operatorname{dist}(x_k^\star, \mathcal{X}) = \mathcal{O}(c_k).$$

Proof. Let $k \in \mathbb{N}$. By optimality of $x_k^\star$ for f^k, we have $\nabla f^k(x_k^\star) = 0$. Hence, by strong convexity, we obtain

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 \leq f^k(x^\star) - f^k(x_k^\star) \leq f(x^\star) - f(x_k^\star) + \gamma_k \pi^k(x^\star) - \gamma_k \pi^k(x_k^\star).$$

We can write

$$\begin{aligned} f(x^\star) - f(x_k^\star) &= f(x^\star) - f(p_k) + f(p_k) - f(x_k^\star) \\ &\leq -\frac{\mu}{2}\|x^\star - p_k\|^2 - \langle \nabla f(x^\star), p_k - x^\star \rangle + f(p_k) - f(x_k^\star). \end{aligned}$$

Since $p_k \in \mathcal{X}$ and $x^\star$ is optimal for f on $\mathcal{X}$, it holds that

$$\langle \nabla f(x^\star), p_k - x^\star \rangle \geq 0,$$

and thus

$$f(x^\star) - f(x_k^\star) \leq -\frac{\mu}{2}\|x^\star - p_k\|^2 + f(p_k) - f(x_k^\star).$$

Since $(x_k^\star)_{k\in\mathbb{N}}$ is bounded, so is $(p_k)_{k\in\mathbb{N}}$. By continuity of ∇f, f is locally Lipschitz on any bounded set containing $(x_k^\star)_{k\in\mathbb{N}}$ and (p_k). Thus, there exists a constant $L \in (0, \infty)$, such that

$$f(x^\star) - f(x_k^\star) \leq -\frac{\mu}{2}\|x^\star - p_k\|^2 + L \operatorname{dist}(x_k^\star, \mathcal{X}).$$

Plugging this into (3.2.1), we obtain

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 \leq -\frac{\mu}{2}\|x^\star - p_k\|^2 + L \operatorname{dist}(x_k^\star, \mathcal{X}) + \gamma_k \pi^k(x^\star) - \gamma_k \pi^k(x_k^\star).$$

By our lower bound on $\pi^k(x_k^\star)$ from lemma 3.10, we have

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 \leq -\frac{\mu}{2}\|x^\star - p_k\|^2 + (L - \gamma_k)\operatorname{dist}(x_k^\star, \mathcal{X}) + \gamma_k \pi^k(x^\star).$$

Using $\pi^k(x^\star) \leq c_k$ and rearranging, we arrive at the desired result. □

3.2.2 Bounding the tracking error

The following analysis is an adaptation of techniques used in [CDH23]. 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 3.12. For any $k \in \mathbb{N}$, we define $A_k := \|x_k - x_k^\star\|^2$, $a_k := \mathbb{E}(A_k)$, and $\Delta_k := \|x_k^\star - x_{k+1}^\star\|$. Further, we define $\xi_k^{[b_k]} := (\xi_k^1, \dots, \xi_k^{b_k}) \in \mathbb{R}^{m \times b_k}$ and $\mathbb{E}_k(X) := \mathbb{E}\big(X \,|\, \xi_{k-1}^{[b_{k-1}]}, \dots, \xi_1^{[b_1]}\big)$. If **??** holds, then f^k is $(L + \gamma_k L_\pi)$-smooth. In that case, we define $L_k := L + \gamma_k L_\pi$. TODO: Notation for the stochastic gradients. Redefine $\tilde{\nabla} f^k$ as the linear combination of $\tilde{\nabla} f$ and $\tilde{\nabla} \pi^k$.

Theorem 3.13. *For all $k \in \mathbb{N}$, the iterates $(x_k)_{k\in\mathbb{N}}$ of algorithm 2 satisfy*

$$a_{k+1} \leq (1 - \tilde{\rho}_k) a_k + \left(\frac{\sigma_f^2 + \gamma_k^2 \sigma_{\pi^k}^2}{b_k} + \gamma_k^2 G_\pi^2 + 2 \right) M^2 \tau_k^2 + (1 + \eta_k^{-1}) \Delta_k^2,$$

where $\rho_k := \mu\tau_k/2 - 8\big((\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)/b_k + 2\gamma_k^2 L_{\pi^k}^2\big)\tau_k^2$, $\eta_k := \min(1, \mu\tau_k/2)$, and $M^2 \in (0, \infty)$.

Remark 3.14. 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 lemmata in the proof of theorem 3.13.

Lemma 3.15. *For all $k \in \mathbb{N}$, it holds that*

$$\|\nabla \pi^k(x_k^\star)\| \leq \frac{G_f}{\gamma_k},$$

where $G_f := \sup_{k\in\mathbb{N}} \|\nabla f(x_k^\star)\| < \infty$.

Proof. By optimality of $x_k^\star$ for f^k,

$$0 = \nabla f^k(x_k^\star) = \nabla f(x_k^\star) + \gamma_k \nabla \pi^k(x_k^\star),$$

and rearranging yields

$$\|\nabla \pi^k(x_k^\star)\| = \frac{\|\nabla f(x_k^\star)\|}{\gamma_k}.$$

Continuity of ∇f and convergence of $x_k^\star$ (theorem 3.1) imply that $G_f := \sup_{k\in\mathbb{N}} \nabla f(x_k^\star) < \infty$, proving the claim. □

Lemma 3.16. *For all $k \in \mathbb{N}$, the iterates $(x_k)_{k\in\mathbb{N}}$ of algorithm 2 satisfy*

$$\mathbb{E}_k \|\tilde{\nabla} f^k(x_k)\|^2 \leq \left(\frac{4(\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)}{b_k} + 8\gamma_k^2 L_{\pi^k}^2 \right) A_k + \left(\frac{\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2}{b_k} + \gamma_k^2 G_\pi^2 + 2 \right) \frac{M^2}{2},$$

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

Proof. Let $k \in \mathbb{N}$. Using the relation $(\alpha + \beta)^2 \leq 2(\alpha^2 + \beta^2)$ $\forall \alpha, \beta \in \mathbb{R}$, we have

$$\mathbb{E}_k \|\tilde{\nabla} f^k(x_k)\|^2 \leq 2\mathbb{E}_k \|\tilde{\nabla} f(x_k)\|^2 + 2\gamma_k^2 \mathbb{E}_k \|\tilde{\nabla} \pi^k(x_k)\|^2. \tag{3.5}$$

By proposition 2.39 and the definition of stochastic gradients, we have

$$\begin{aligned}\mathbb{E}_k\|\tilde{\nabla} f(x_k)\|^2 &= \mathbb{V}\mathrm{ar}_k(\tilde{\nabla} f(x_k)) + \|\nabla f(x_k)\|^2 \quad \text{and} \\ \mathbb{E}_k\|\tilde{\nabla}\pi^k(x_k)\|^2 &= \mathbb{V}\mathrm{ar}_k(\tilde{\nabla}\pi^k(x_k)) + \|\nabla\pi^k(x_k)\|^2.\end{aligned} \tag{3.6}$$

By proposition 2.38, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}_k(\tilde{\nabla} f(x_k)) &= \frac{1}{b_k}\mathbb{V}\mathrm{ar}_k(\tilde{\nabla} f(x_k,\xi)) \leq \frac{\sigma_f^2}{b_k}(\|x_k\|^2+1) \\ \mathbb{V}\mathrm{ar}_k(\tilde{\nabla}\pi^k(x_k)) &= \frac{1}{b_k}\mathbb{V}\mathrm{ar}_k(\tilde{\nabla}\pi^k(x_k,\xi)) \leq \frac{\sigma_{\pi^k}^2}{b_k}(\|x_k\|^2+1).\end{aligned}$$

Hence, combining with (3.5) and (3.6), we get

$$\begin{aligned}\mathbb{E}_k\|\tilde{\nabla} f^k(x_k)\|^2 &\leq \frac{2\sigma_f^2}{b_k}(\|x_k\|^2+1) + 2\|\nabla f(x_k)\|^2 + \frac{2\gamma_k^2\sigma_{\pi^k}^2}{b_k}(\|x_k\|^2+1) + 2\gamma_k^2\|\nabla\pi^k(x_k)\|^2 \\ &= \frac{2(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2)}{b_k}(\|x_k\|^2+1) + 2\|\nabla f(x_k)\|^2 + 2\gamma_k^2\|\nabla\pi^k(x_k)\|^2.\end{aligned}$$

Using the relation $(\alpha+\beta)^2 \leq 2(\alpha^2+\beta^2)\ \forall\alpha,\beta\in\mathbb{R}$ again, we have

$$\|x_k\|^2 = \|x_k - x_k^\star + x_k^\star\|^2 \leq 2A_k + 2\|x_k^\star\|^2$$

and

$$\|\nabla f(x_k)\|^2 = \|\nabla f(x_k) - \nabla f(x_k^\star) + \nabla f(x_k^\star)\|^2 \leq 2L_f^2 A_k + 2\|\nabla f(x_k^\star)\|^2,$$

where we additionally used the smoothness of f. Hence

$$\begin{aligned}\mathbb{E}_k\|\tilde{\nabla} f^k(x_k)\|^2 &\leq \frac{2(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2)}{b_k}(2A_k + 2\|x_k^\star\|^2+1) + 2(2L_f A_k + 2\|\nabla f(x_k^\star)\|^2) + 2\gamma_k^2\|\nabla\pi^k(x_k)\|^2 \\ &= \frac{4(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)}{b_k}A_k + \frac{2(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2)}{b_k}(2\|x_k^\star\|^2+1) + 4\|\nabla f(x_k^\star)\|^2 + 2\gamma_k^2\|\nabla\pi^k(x_k)\|^2.\end{aligned}$$

By theorem 3.1, $(x_k^\star)_{k\in\mathbb{N}}$ converges, and thus continuity of ∇f implies $\sup_{k\in\mathbb{N}}\nabla\|f(x_k^\star)\|^2 < \infty$. Setting $\tilde{M}^2 := \sup_{k\in\mathbb{N}}\max(4\|x_k^\star\|^2+2,\ 4\nabla\|f(x_k^\star)\|^2)$, we obtain

$$\mathbb{E}_k\|\tilde{\nabla} f^k(x_k)\|^2 \leq \frac{4(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)}{b_k}A_k + \left(\frac{\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2}{b_k}+1\right)\tilde{M}^2 + 2\gamma_k^2\|\nabla\pi^k(x_k)\|^2. \tag{3.7}$$

Next, we have

$$\begin{aligned}\|\nabla\pi^k(x_k)\|^2 &\leq 2\|\nabla\pi^k(x_k) - \nabla\pi^k(x_k^\star)\|^2 + 2\|\nabla\pi^k(x_k^\star)\|^2 \\ &\leq 2(L_{\pi^k}\|x_k - x_k^\star\| + G_\pi)^2 + \frac{2\,G_f^2}{\gamma_k^2} \\ &\leq 4L_{\pi^k}^2 A_k + 4G_\pi^2 + \frac{2\,G_f^2}{\gamma_k^2},\end{aligned}$$

where we used our assumption on $\nabla\pi^k$ and lemma 3.15 in the second step. Finally, setting $M^2 :=$

$1/2\max(\tilde{M}^2, 8, 4G_f^2)$ and combining the above with (3.7), we arrive at

$$\begin{aligned}\mathbb{E}_k\|\tilde{\nabla} f^k(x_k)\|^2 &\leq \frac{4(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)}{b_k}A_k + \left(\frac{\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2}{b_k}+1\right)M^2 + 2\gamma_k^2\left(4L_{\pi^k}^2A_k + 4G_\pi^2 + \frac{2\,G_f^2}{\gamma_k^2}\right)\\ &= \left(\frac{4(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)}{b_k}+8\gamma_k^2L_{\pi^k}^2\right)A_k + \left(\frac{\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2}{b_k}+1\right)\tilde{M}^2 + 8\gamma_k^2G_\pi^2 + 4G_f^2\\ &\leq \left(\frac{4(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)}{b_k}+8\gamma_k^2L_{\pi^k}^2\right)A_k + \left(\frac{\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2}{b_k}+\gamma_k^2G_\pi^2+2\right)\frac{M^2}{2},\end{aligned}$$

as desired. □

Lemma 3.17. *For all $k\in\mathbb{N}$, the iterates of algorithm 2 satisfy*

$$\mathbb{E}\|x_{k+1}-x_k^\star\|^2 \leq (1-q_k)\,a_k + \Big(\frac{\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2}{b_k}+\gamma_k^2G_\pi^2+2\Big)\frac{M^2}{2}\tau_k^2$$

for all $k\in\mathbb{N}$, where $q_k := \mu\tau_k - 4\Big((\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)/b_k + 2\gamma_k^2L_{\pi^k}^2\Big)\tau_k^2$ and $M^2\in(0,\infty)$.

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

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

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

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

Strong convexity of f^k yields

$$\mathbb{E}_k\|x_{k+1}-x_k^\star\|^2 \leq (1-\mu\tau_k)A_k + \tau_k^2\,\mathbb{E}_k\|\tilde{\nabla}f^k(x_k)\|^2. \tag{3.8}$$

By lemma 3.16, we have

$$\mathbb{E}_k\|\tilde{\nabla} f^k(x_k)\|^2 \leq \left(\frac{4(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)}{b_k}+8\gamma_k^2L_{\pi^k}^2\right)A_k + \left(\frac{\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2}{b_k}+\gamma_k^2G_\pi^2+2\right)\frac{M^2}{2}$$

for some constant $M^2\in(0,\infty)$. Plugging this into (3.8), we get

$$\mathbb{E}_k\|x_{k+1}-x_k^\star\|^2 \leq \left(1-\mu\tau_k+\Big(\frac{4(\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2+L_f^2)}{b_k}+8\gamma_k^2L_{\pi^k}^2\Big)\tau_k^2\right)A_k + \Big(\frac{\sigma_f^2+\gamma_k^2\sigma_{\pi^k}^2}{b_k}+\gamma_k^2G_\pi^2+2\Big)\frac{M^2}{2}\tau_k^2.$$

Now, taking expectations of both sides, proposition 2.42 yields the claim. □

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

Proof of Theorem 3.13. Let $k\in\mathbb{N}$. First, we have

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

We can apply the Cauchy-Schwarz and Young inequalities to the inner product term, and obtain

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

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

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

Using lemma 3.17, we can bound the first term:

$$(1+\eta_k)\mathbb{E}\|x_{k+1} - x_k^\star\|^2 \leq (1+\eta_k)(1-q_k)a_k + (1+\eta_k)\left(\frac{\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2}{b_k} + \gamma_k^2 G_\pi^2 + 2\right)\frac{M^2}{2}\tau_k^2\,, \tag{3.10}$$

where

$$q_k = \mu\tau_k - \left(\frac{4(\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)}{b_k} + 8\gamma_k^2 L_{\pi^k}^2\right)\tau_k^2.$$

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

$$\begin{aligned}
(1+\eta_k)(1-q_k) &= (1+\eta_k)\left(1 - \mu\tau_k + \left(\frac{4(\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)}{b_k} + 8\gamma_k^2 L_{\pi^k}^2\right)\tau_k^2\right) \\
&= 1 + \eta_k - (1+\eta_k)\mu\tau_k + (1+\eta_k)\left(\frac{4(\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)}{b_k} + 8\gamma_k^2 L_{\pi^k}^2\right)\tau_k^2 \\
&\leq 1 + \frac{\mu}{2}\tau_k - \mu\tau_k + 2\left(\frac{4(\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)}{b_k} + 8\gamma_k^2 L_{\pi^k}^2\right)\tau_k^2 \\
&= 1 - \frac{\mu}{2}\tau_k + 8\left(\frac{(\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)}{b_k} + 2\gamma_k^2 L_{\pi^k}^2\right)\tau_k^2\,,
\end{aligned}$$

where we used the definition of η_k and $1 \leq 1 + \eta_k \leq 2$. Plugging this into (3.10) and using $1 + \eta_k \leq 2$ again, we arrive at the bound

$$(1+\eta_k)\mathbb{E}\|x_{k+1} - x_k^\star\|^2 \leq (1-\rho_k)\,a_k + \left(\frac{\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2}{b_k} + \gamma_k^2 G_\pi^2 + 2\right)M^2\tau_k^2,$$

with

$$\rho_k := \frac{\mu}{2}\tau_k - 8\left(\frac{(\sigma_f^2 + \gamma_k^2\sigma_{\pi^k}^2 + L_f^2)}{b_k} + 2\gamma_k^2 L_{\pi^k}^2\right)\tau_k^2.$$

Together with (3.9), we obtain the claim. □

3.2.3 Convergence rates

TODO 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 ($\mathrm{P}^{\mathrm{new}}$).

Assumption 5. The random matrix $A(\xi)$ satisfies $\mathbb{E}\|A(\xi)\|_{\mathrm{F}}^4 < \infty$.

Theorem 3.18. *In the situations of* ($\mathrm{P}^{\mathrm{new}}$) *and* ($\mathrm{P}^k$), *let* $\pi_{\mathrm{hin}}(x) := \mathbb{E}\|(0, A(\xi)x - c)_+\|^2$ *and assume that assumption 5 and* **??** *(TODO: Need assumptions of* **??***) 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* $b_k = 1 + k^{2(1/3-\epsilon)}$ *(TODO: I think better to just leave out*

ϵ), *converges, and yields iterates* $(x_k)_{k\in\mathbb{N}}$ *that satisfy*

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

where $x^\star$ *denotes the solution to* (**??**). *Furthermore, it holds that*

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

Lemma 3.19. *In the situations of* ($\mathrm{P}^{\mathrm{new}}$) *and* ($\mathrm{P}^k$), *assume that* f *and* π *are differentiable and* ∇f *is continuous. Then we have*

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

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

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

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

which implies

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

by optimality of $x_k^\star$ for f^k. We can apply strong convexity of f and proposition 2.15 to obtain

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

Convexity of π implies that $\langle x-y, \pi(x)-\pi(y)\rangle \geq 0$ for all $x, y \in \mathbb{R}^d$, by proposition 2.11. Hence, by positivity of γ_k, we have

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

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

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

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

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

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

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

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

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

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

Lemma 3.20 (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 [Chu54]. □

We can now prove the main theorem of this section.

Proof of Theorem 3.18. Let $k \in \mathbb{N}$. By the triangle inequality,

$$\mathbb{E}\|x_k - x^\star\| \leq \mathbb{E}\|x_k - x_k^\star\| + \|x_k^\star - x^\star\|. \tag{3.12}$$

First, we analyze $\mathbb{E}\|x_k - x_k^\star\|$. We want to use theorem 3.13, but for this we first need to show that ???????? hold. By Jensen's inequality (proposition 2.36), assumption 5 implies assumption 1. Hence, by ??, ?? is satisfied by j. Further, by ??, π satisfies ??. ?? was verified in ??. What's left to show is that ?? 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

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

We have

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

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

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

Note that all expectations are finite, by assumption 5. 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 5 and Jensen's inequality (proposition 2.36). Hence, ?? is satisfied.

With our choices for γ_k and b_k, theorem 3.13 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

$$\tilde{\rho}_k \geq \frac{\mu \tau_k}{4}. \tag{3.13}$$

By lemma 3.19, 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 3.20, we have $a_k = \mathcal{O}(k^{-2/3})$ and an application of Jensen's inequality (proposition 2.36) yields $\mathbb{E}\|x_k - x_k^\star\| = \mathcal{O}(k^{-1/3})$.

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

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

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

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

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

Remark 3.21. 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 ith iterate of algorithm 2.

Lemma 3.22. *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 $1/k\sum_{i=1}^{k_0-1} a_i = \mathcal{O}(k^{-1})$ and $a \in (0,1)$, we obtain $1/k\sum_{i=1}^{k}\alpha_i = \mathcal{O}(k^{-a})$. □

Theorem 3.23. *In the situations of* (**??**) *and* (**??**)*, let* $\pi(x) = \mathbb{E}\|(A(\xi)x-b)_+\|^2$*, and assume that* **????????** *and assumption 5 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*

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

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

Proof. Smoothness of j and π is verified in **??**. Let $k \in \mathbb{N}$. We have

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

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

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

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

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

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 3.18, 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 3.22, 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 $\nabla j^i(x_i^\star) = 0$ 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 3.18, we obtain

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

for all $i \in \mathbb{N}$. An application of lemma 3.22 now yields

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

and combining with (3.16), we get

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

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

$$j(x^\star) - j^k(x_k^\star) = j^k(x^\star) - j^k(x_k^\star) \leq L_k \left\|x^\star - x_k^\star\right\|^2$$

and an application of **??** yields

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

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

3.3 Exponential 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 (P^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 [CDH23], who in turn adapt averaging techniques used by Ghadimi and Lan in [GL12]. The analysis hinges on the following fundamental lemma.

Lemma 3.24 (Averaging lemma)**.** *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ 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*

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

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 $t \in \mathbb{N}_0$:

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

Proof. See lemma 42 in [CDH23]. □

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

Lemma 3.25. *Let ?????? hold. Then, algorithm 3 with step sizes $\tau_k \in (0, 1/L_k)$ yields iterates $(x_k)_{k\in\mathbb{N}}$, such that*

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

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,

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

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 2.48, 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 2.15, 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 2.15 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 multipliying 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,

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

thus

$$\mathbb{E} \|z_k\|^2 = \mathbb{E}(\mathbb{V}\mathrm{ar}_k(\tilde{G}^k(x_k))), \tag{3.21}$$

by proposition 2.42. **??** and proposition 2.45 imply, for all $x \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_k(\tilde{G}^k(x_k)) &= \frac{1}{b_k}\mathbb{V}\mathrm{ar}_k(G^k(x_k)) \\
&\leq \frac{C}{b_k}\Big(\|x_k\|^2 + \|x_k\|^2\gamma_k^2 + \gamma_k^2 + 1\Big) \\
&\leq \frac{C}{b_k}\Big(2\|x_k - x\|^2 + 2\|x\|^2 + 2\|x_k - x\|^2\gamma_k^2 + 2\|x\|^2\gamma_k^2 + \gamma_k^2 + 1\Big) \\
&= \frac{C}{b_k}\Big(2(1+\gamma_k^2)\|x_k - x\|^2 + 2(1+\gamma_k^2)\|x\|^2 + \gamma_k^2 + 1\Big) \\
&= \frac{C(1+\gamma_k^2)}{b_k}\Big(2\|x_k - x\|^2 + 2\|x\|^2 + 1\Big) \\
&= \frac{C(1+\gamma_k^2)}{b_k}\Big(2\|x_k - x\|^2 + M_x^2\Big),
\end{aligned}$$

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

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

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

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

for all $x \in \mathbb{R}^d$. □

In [CDH23], the authors make an assumption, which in our setting would essentially boil down to imposing global boundedness on $\|\nabla\pi(x)\|$. For our purposes, however, this would be too strong, which motivates the following relaxed assumption.

Assumption 6. 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 6 holds if π is diffierentiable and Lipschitz continuous or has Lipschitz continuous gradients.

Lemma 3.26. *In the situation of* (P^k)*, assume that assumption 6 holds. Then, for all* $t, i \in \mathbb{N}$, $u, v \in \mathbb{R}^d$, *we have*

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

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 $t, i \in \mathbb{N}$,

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

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

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

We can use assumption 6 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

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

Using this bound in (3.22), we arrive at

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

as desired. □

Lemma 3.27. *In the situation of* (P^k)*, assume that* **????????????** *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 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*

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

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 **??** implies assumption 6. Hence, we can

apply lemmas 3.25 and 3.26 and have, for all $\epsilon > 0$,

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

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

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

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 3.18 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 $x_k^\star$ is optimal for f^k, we have

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

which implies $\|\nabla\pi(x_k^\star)\| = \|\nabla f(x_k^\star)\|/\gamma_k \leq G/\gamma_k$, where $G := \sup_{k\in\mathbb{N}} \|\nabla f(x_k^\star)\|$ and $G < \infty$ due to convergence of $(x_k^\star)_{k\in\mathbb{N}}$ (theorem 3.1) and continuity of ∇f. Since, $\lim_{k\to\infty} \gamma_k = \infty$ 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

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

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,

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

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

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

for all $k, t \geq \max(k_0, k_1)$. □

Corollary 3.28. *In the situation of lemma 3.27, 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*

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

where

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

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

Proof. Let $h(x) := f^k(x) - f^k(x_k^\star)$ for $x \in \mathbb{R}^d$, $\rho_{t+1} := 2\tau_t$, $c_1 = \mu/4$, $c_2 = -\mu/8$, $V_i := \mathbb{E}\|x_k^\star - x_i\|^2$, $i \in \{t, t+1\}$, and $\omega_{t+1} := (\mu/2) M^2 \tau_t$. Then, by lemma 3.27, 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 3.24 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. □

4 Prox-SGD for Simple Nonsmooth Objectives

5 Numerical Examples

5.1 Inventory Control

5.2 Support Vector Machines

5.3 Sparse SVM

6 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

[Ber97] Dimitri P Bertsekas. "Nonlinear programming". In: *Journal of the Operational Research Society* 48.3 (1997).

[BH19] Joseph K Blitzstein and Jessica Hwang. *Introduction to probability*. Chapman and Hall/CRC, 2019.

[BT96] James V Burke and Paul Tseng. "A unified analysis of Hoffman's bound via Fenchel duality". In: *SIAM Journal on Optimization* 6.2 (1996).

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CDH23] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. "Stochastic Optimization under Distributional Drift". In: *Journal of Machine Learning Research* 24.147 (2023). URL: http://jmlr.org/papers/v24/21-1410.html.

[Chu54] Kai Lai Chung. "On a stochastic approximation method". In: *The Annals of Mathematical Statistics* (1954).

[Duc18] John C Duchi. "Introductory lectures on stochastic optimization". In: *The mathematics of data* 25 (2018).

[Dur19] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

[Eva13] Lawrence Evans. "An Introduction to Mathematical Optimal Control Theory Version 0.2". In: (Feb. 2013).

[Fer+19] Olivier Fercoq et al. "Almost surely constrained convex optimization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1910–1919.

[GG23] Guillaume Garrigos and Robert M Gower. "Handbook of convergence theorems for (stochastic) gradient methods". In: *arXiv preprint arXiv:2301.11235* (2023).

[GH22] Caroline Geiersbach and Michael Hintermüller. "Optimality conditions and Moreau–Yosida regularization for almost sure state constraints". In: *ESAIM: Control, Optimisation and Calculus of Variations* 28 (2022).

[GL12] 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).

[Has+04] Trevor Hastie et al. "The entire regularization path for the support vector machine". In: *Journal of Machine Learning Research* 5.Oct (2004).

[Hof03] Alan J Hoffman. "On approximate solutions of systems of linear inequalities". In: *Selected Papers Of Alan J Hoffman: With Commentary*. World Scientific, 2003.

[Nem+09] A. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (2009). DOI: 10.1137/070704277. eprint: https://doi.org/10.1137/070704277. URL: https://doi.org/10.1137/070704277.

[NT20] Angelia Nedić and Tatiana Tatarenko. “Convergence rate of a penalty method for strongly convex problems with linear constraints”. In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020.

[RM51] Herbert Robbins and Sutton Monro. “A stochastic approximation method”. In: *The annals of mathematical statistics* (1951).

[Sha09] Alexander Shapiro. “Semi-infinite programming, duality, discretization and optimality conditions”. In: *Optimization* 58.2 (2009), pp. 133–161.

[WN+99] Stephen Wright, Jorge Nocedal, et al. “Numerical optimization”. In: *Springer Science* 35.67-68 (1999).