Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

A Penalty Method for Almost Surely Constrained Stochastic Optimization

Master Thesis

submitted on: November 8, 2024

Department of Mathematics
Mathematical Statistics and Stochastic Processes

Name:	Amir Miri Lavasani
Student ID:	7310114
Study Program:	Mathematics
First Reviewer:	Johannes Lederer
Second Reviewer:	Caroline Geiersbach

Abstract

Contents

List of Figures

1

Introduction

1.1. Problem statement and objective

Convex optimization is concerned with problems of the form

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

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\mathcal{X} \subset \mathbb{R}^n$ 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 $x \in \mathbb{R}^n$ is referred to as a *decision variable*. If there exists $x^\star \in \mathcal{X}$ such that $f(x^\star) \leq f(x)$ for all $x \in \mathcal{X}$, then $x^\star$ is called *solution*. The value $f(x^\star)$ is called *optimal value*.

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. In this thesis, we consider composite stochastic optimization problems of the form

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \{\, f(x) := \mathbb{E}(F_\xi(x)) + r(x) \,\} \\ &\text{s.t.} \quad A(\xi)x - b(\xi) \leq 0 \quad \text{almost surely (a.s.)}, \end{aligned} \tag{P}$$

where ξ is a random variable, and for any realization z of ξ, $A(z) \in \mathbb{R}^{m \times n}$ and $b(z) \in \mathbb{R}^m$. We assume that $\mathbb{E}(F_\xi)\colon \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex and L-smooth, and that $r\colon \mathbb{R}^n \to \mathbb{R}$ is convex (but possibly nonsmooth). Examples of these problems appear in numerous areas of applied mathematics. One such domain is optimal control, where the randomness often arises from some continuous uncertainty in some variables, which leads to an infinite amount of constraints. For example, such uncertainty could come from an unknown future demand that is subject to gaussian noise (see **??**). In that case, optimization algorithms for solving (P) need a way to deal with the constraints one-by-one or in batches, as a simlutaneous treatment of all constraints, like in the classical projected gradient method [1], would be impossible to implement. This holds true even if the number of constraints is not infinite but merely very large, as is the case in modern machine learning, where the random variable ξ models data points in a data set of size $N \in \mathbb{N}$, and the decision variable x represents parameters of some statistical model. Problem (P) then takes on the specific form

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^{n} F_i(x) + r(x) \right\}$$
$$\text{s.t.} \quad A_i x - b_i \leq 0 \quad \text{for all } i \in \{ 1, \dots, N \}.$$

A classical approach to solve (P) is stochastic subgradient descent (SGD) [2]: We start from an initial point $x_0 \in \mathbb{R}^n$. In iteration $k \in \mathbb{N}$, we pick a *step size* $\eta_k \in (0, \infty)$ and a *stochastic subgradient* g_k of f. Then we set $x_{k+1} = \Pi_{\mathcal{X}}(x_k - \eta_k g_k)$, and repeat. The map $\Pi_{\mathcal{X}} \colon \mathbb{R}^n \to \mathbb{R}^n$ is the *projection map onto* $\mathcal{X}$, which is defined as $\Pi_{\mathcal{X}}(x) := \arg\min_{y \in \mathcal{X}} \|x - y\|$, and ensures that the iterates stay in the feasible set $\mathcal{X}$. If $x^\star \in \mathcal{X}$ solves (P), then one can show that, under suitable choice of step sizes, $\|x_k - x^\star\|^2$ converges to the solution of (P) with rate $O(1/k)$ for strongly convex objectives. However, Nemirovski et al. [1] showed that this convergence is highly dependent on knowing the strong convexity constant, and proposed a more robust version of the algorithm that utilizes suitable averages of the original iterates. They proved that the resulting iterates $(\bar{x}_k)_{k \in \mathbb{N}}$ yield convergence of the function values $f(\bar{x}_k)$ to that of $f(x^\star)$ with rate $O(1/\sqrt{k})$ for objectives f which need only be convex. Unfortunately, we can not apply SGD to our problem, because the complexity of $\mathcal{X}$ makes computing the projection $\Pi_{\mathcal{X}}(x_k - \eta_K g_k)$ infeasible in our case.

A classic idea to deal with complex feasibility sets is to use *penalty functions*. In this approach the constrained problem (P) gets approximated by an unconstrained problem, by introducing a convex function $\pi \colon \mathbb{R}^n \to \mathbb{R}$ that penalizes points that are not feasible: $\pi(x) > 0$ for infeasible x and $\pi(x) = 0$ for feasible x. The resulting approximation to (P) then takes the form

$$\min_{x \in \mathbb{R}^n} \left\{ f_\gamma(x) := f(x) + \gamma \pi(x) \right\}, \tag{1.1}$$

where $\gamma \in (0, \infty)$ is a constant that is used to control the influence of the penalty function π on the objective. For this unconstrained problem, the projection map is simply the identiy map, so one can easily apply SGD to solve (1.1). The larger γ, the closer the solution to (1.1) is to being feasible. Under suitable choice for π, one can show that in the limit $\gamma \to \infty$ the sequence of solutions $(x^\star_\gamma)_{\gamma\in(0,\infty)}$ to (1.1) converges to the solution $x^\star$ of (P): $\lim_{\gamma\to\infty} x^\star_\gamma = x^\star$. For certain choices of π, one can even show that there exists some finite $\gamma \in (0,\infty)$, such that $x^\star_\gamma = x^\star$. Such penalties are called *exact penalties*. A standard example is the Hinge penalty π_{ℓ_1}, defined by $\pi_{\ell_1}(x) := \mathbb{E}(\|(A(\xi)x - b(\xi))_+\|_1)$, where $\|\cdot\|_1$ is the ℓ_1-norm, and $(y)_+$ applies $\mathbb{R} \ni t \mapsto \max(t, 0)$ to every element of $y \in \mathbb{R}^m$. While the defining property of exact penalties is very desirable, they, like the Hinge penalty, all suffer from necessarily being nonsmooth, which is known to slow down stochastic subgradient descent. On the other hand, smooth penalties like the squared hinge penalty $\pi_{\ell_2}(x) := \mathbb{E}(\|(A(\xi)x - b(\xi))_+\|_2^2)$ with $\|\cdot\|_2$ the ℓ_2-norm, often need γ to grow very large to get solutions that are reasonably close to the feasible set. This has the unfortunate side effect that the gradient norm $\|\nabla f_\gamma\|_2$ can grow very large, which makes stochastic gradient descent iterates often very unstable in practice. Additionally, one is forced to use step sizes that decay to 0 quickly to counter the large gradient norms, which slows down convergence.

A solution to the drawbacks of classical penalty methods was introduced by Nedić et al. [3], where the authors analyzed problems similar to ours of the form

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\text{s.t.} \quad a_i^\top x - b_i \le 0 \quad \text{for all } i \in \{1,\dots,m\}. \end{aligned} \tag{1.2}$$

Instead of a fixed penalty function π however, the authors introduced a sequence of smooth *inexact* penalties $(\pi_k^{\text{hub}})_{k\in\mathbb{N}}$, defined as follows: Let $(\delta_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers. For $k \in \mathbb{N}$, we define

$$\pi_k^{\text{hub}}(x) := \frac{1}{m}\sum_{i=1}^m h_k^{\text{hub}}(x; a_i, b_i),$$

where

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

for $x, a \in \mathbb{R}^n$, and $b \in \mathbb{R}$. The authors then considered the sequence of unconstrained problems

$$\min_{x\in\mathbb{R}^n} \left\{ f_k(x) := f(x) + \gamma_k \pi_k^{\text{hub}}(x) \right\}, \tag{1.3}$$

for $k \in \mathbb{N}$ and $\gamma_k \in (0,\infty)$. Where before we required $\pi(x) = 0$ for all feasible points, the inexact penalties only satisfy $\pi_k^{\text{hub}}(x) \ge 0$ for feasible $x \in \mathbb{R}^n$. The

Algorithm 1 Incremental Gradient Method (Nedić et al. [3])

Require: Initial point $x_0 \in \mathbb{R}^n$, step sizes $(\eta_k)_{k\in\mathbb{N}_0}$, penalty weights $(\gamma_k)_{k\in\mathbb{N}_0}$
1: **for** $k = 0$ to $K-1$ **do**
2: Uniformly sample random index $i \in \{1, \ldots, m\}$
3: $g \leftarrow \tilde{\nabla} f(x_k) + \gamma_k \nabla h_k^{\text{hub}}(x_k; a_i, b_i)$
4: $x_{k+1} \leftarrow x_k - \eta_k g$
5: **end for**
6: $S_K \leftarrow \sum_{k=0}^{K} \eta_k^{-1}$
7: $\bar{x}_K \leftarrow S_K^{-1} \sum_{k=0}^{K} \eta_k^{-1} x_k$
8: **return** $\bar{x}_K$

crucial properties of the particular penalty sequence that the authors introduced were that the sequence a) majorizes the hinge penalty, $\pi_k^{\text{hub}}(x) \geq \pi_{\ell_1}(x)$ for all $x \in \mathbb{R}^n$, b) converges to π_{ℓ_1} pointwis, $\lim_{k\to\infty} \pi_k^{\text{hub}}(x) = \pi_{\ell_1}(x)$ for all $x \in \mathbb{R}^n$, and c) has uniformly bounded gradients, $\sup_{k\in\mathbb{N}} \sup_{x\in\mathbb{R}^n} \left\|\nabla \pi_k^{\text{hub}}(x)\right\|_2 < \infty$. By carefully choosing the sequence of parameters that control the convergence to π_{ℓ_1}, denoted by $(\delta_k)_{k\in\mathbb{N}}$, as well as the sequence $(\gamma_k)_{k\in\mathbb{N}}$, the authors were able to show that there exists a $\gamma_K \in (0,\infty)$ large enough such that the distance-to-feasibility of the solution $x_K^\star$ to the corresponding problem (1.3), is independent of γ_K and only controlled by δ_K. Written in mathematical notation, this means that $\operatorname{dist}(x_K^\star, \mathcal{X}) := \inf_{x\in\mathcal{X}} \left\|x - x_K^\star\right\| = \mathcal{O}(\delta_K)$. Thus, this approach manages to combine the highly desirable properties of smooth unconstrained problems with that of exact penalties. The authors then go on to present an iterative stochastic gradient algorithm (see algorithm 1), which proceeds as follows: Start from an initial point $x_0 \in \mathbb{R}^n$. Then, in iteration $k \in \mathbb{N}$: Compute a subgradient of f at x_k, denoted by $\tilde{\nabla} f(x_k)$; sample a random index $i \in \{1, \ldots, m\}$ and calculate the gradient $\nabla h_k^{\text{hub}}(x_k; a_i, b_i)$; finally, update $x_{k+1} := x_k - \eta_k(\tilde{\nabla} f(x_k) + \gamma_k \nabla h_k^{\text{hub}}(x_k; a_i, b_i))$. After a set amount of $K \in \mathbb{N}$ iterations, one computes a certain weighted average of the iterates, which yields the final output $\bar{x}_K$. For strongly convex objectives (that satisfy certain assumptions on the gradients), the authors show that, for any $\epsilon \in (0,\infty)$, one can choose $(\eta_k)_{k\in\mathbb{N}}, (\gamma_k)_{k\in\mathbb{N}}, (\delta_k)_{k\in\mathbb{N}}$ such that the sequece $(\bar{x}_k)_{k\in\mathbb{N}}$ satisfies $\operatorname{dist}(\bar{x}_k, \mathcal{X}) = \mathcal{O}(\log^\epsilon k/k)$ and $|f(\bar{x}_k) - f(x^\star)| = \mathcal{O}(\log^\epsilon k/k)$, where $x^\star$ denotes the solution to (1.2).

In this work, we aim to build on the incremental gradient method (algorithm 1). First, we will extend the method to the more general situation of (P). Namely, our version of (1.3) has the form

$$\min_{x\in\mathbb{R}^n} \left\{ f_k(x) := \mathbb{E}(F_\xi(x)) + r(x) + \gamma_k \pi_k(x) \right\},$$

for $k \in \mathbb{N}$, where, as before, $\gamma_k \in (0,\infty)$. We keep the penalties $\pi_k(x)$ more generic, but assume that there exist smooth real-valued functions $(h_k)_{k\in\mathbb{N}}$ defined on $\mathbb{R}^m$, such that

$$\pi_k(x) = \mathbb{E}(h_k(A(\xi)x - b(\xi))) \text{ and } \pi_k(x) \downarrow_{k\to\infty} \pi_{\ell_1}(x)$$

for all $x \in \mathbb{R}^n$. This more general treatment allows for more flexibility in the design of the method, and includes the recent *softplus penalty* introduced in [4].

As opposed to the setting of algorithm 1, we may not be able to calculate the gradient of our objective, $x \mapsto \mathbb{E}(F_\xi(x)) + r(x)$, because r may be nonsmooth or because calculating the gradient of the expectation functional, $\mathbb{E}(F_\xi(x))$, may be infeasible. Say, for example, because the distribution of ξ is unknown or, in the case of large-scale machine learning, because the size of the dataset is too large to feasibly compute the full gradient of $\mathbb{E}(F_\xi(x))$ (which would require the evaluation of a very large sum) for multiple iterations or even just once. Nedić et al. deal with nonsmoothness by using *subgradients* instead of gradients, which are also defined at nondifferentiable points. However, in our approach we choose to instead use a *proximal operator* [5] to deal with nonsmooth objectives: For $\eta \in (0,\infty)$ and $r : \mathbb{R}^n \to \mathbb{R}$, the proximal operator $\mathrm{prox}_{\eta r} : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$\mathrm{prox}_{\eta r}(x) := \operatorname*{arg\,min}_{u\in\mathbb{R}^n} \left\{ r(u) + \frac{1}{2\eta} \|u - x\|_2^2 \right\}.$$

The proximal map allows us to only work with gradients of the differentiable terms in our objective. To calculate these gradients for objectives involving intractable expectation functionals/large sums, we work with *stochastic gradients* instead of regular (full) gradients. A stochastic gradient of a differentiable function f at some point $x \in \mathbb{R}^n$ is any random variable g such that $\mathbb{E}(g) = \nabla f(x)$. In our case, we can calculate a stochastic gradient of $\mathbb{E}(F_\xi) + \gamma_k \pi_k$ at $x \in \mathbb{R}^n$ based on a sample z from the distribution of ξ as $\nabla F_z(x) + \gamma_k A^\top(z)\nabla h_k(A(z)x - b(z))$. To deal with the variance introduced by using a stochastic gradient instead of the full gradient, we will use a minibatch of samples and average over the resulting stochastic gradients to get a variance-reduced estimate of $\nabla\mathbb{E}(F_\xi(x)) + \gamma_k \nabla \pi_k(x)$. The pseudo-code for the full algorithm is presented in algorithm 2.

Our objective is to analyze algorithm 2 theoretically and support our theoretical findings with numerical examples. TODO: Mention here the outline.

Algorithm 2 Sequential Proximal Stochastic Gradient Descent (SeqProx-SGD)

Require: Initial point $x_0 \in \mathbb{R}^n$, step sizes $(\eta_k)_{k\in\mathbb{N}_0}$, penalty weights $(\gamma_k)_{k\in\mathbb{N}_0}$, minibatch sizes $(\beta_k)_{k\in\mathbb{N}_0}$, smooth penalty functions $(h_k)_{k\in\mathbb{N}_0}$, sample oracle for distribution of ξ

for $k = 0$ to $K-1$ **do**
  Sample minibatch $\{ \xi_1^k, \dots, \xi_{\beta_k}^k \}$ from distribution of ξ
  $g_k \leftarrow \beta_k^{-1} \sum_{i=1}^{\beta_k} \nabla F_{\xi_i^k}(x_k) + \gamma_k A(\xi_i^k)^\top \nabla h_k(A(\xi_i^k)x_k - b(\xi_i^k))$
  $x_{k+1} \leftarrow \operatorname{prox}_{\eta_k r}(x_k - \eta_k g_k)$
end for
$S_K \leftarrow \sum_{k=1}^{K} \eta_k^{-1}$
$\bar{x}_K \leftarrow S_K^{-1} \sum_{k=1}^{K} \eta_k^{-1} x_k$
return $\bar{x}_K$

1.2. Contributions

1.3. Related literature

1.4. Example applications

1.5. Notation

2

Theory Background

2.1. Probability theory

2.2. Optimization theory

2.2.1. Convex optimization

2.2.2. Stochastic optimization

2.2.3. Multifunctions and metric regularity

3

Sequential Proximal SGD Method

In this chapter we will analyze the convergence properties of algorithm 2 applied to the constrained stochastic optimization problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \{ f(x) := \mathbb{E}(F_\xi(x)) + r(x) \} \\ &\text{s. t.} \quad A(\xi)x - b(\xi) \leq 0 \quad \text{a. s.}, \end{aligned} \tag{P}$$

where we implicitly assume the existence of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which the random variable $\xi \colon \Omega \longrightarrow \mathbb{R}^p$, as well as the expected value mapping, are defined. We denote the feasible set for (P) as

$$\mathcal{X} := \{ x \in \mathbb{R}^n \mid A(\xi)x - b(\xi) \leq 0 \ \text{a. s.} \}.$$

Assumption 1. Throughout, we will make the following basic assumptions about problem (P):

1. The function $x \mapsto F_\xi(x)$ is almost surely L-smooth for some $L \in (0, \infty)$, and there exists a point $x \in \mathbb{R}^n$ such that $\mathbb{E} \|\nabla F_\xi(x)\|^2 < \infty$. Further, the expectation $x \mapsto \mathbb{E}(F_\xi(x))$ is μ-strongly convex for some $\mu \in (0, \infty)$.

2. The regularizer r is convex and subdifferentiable with bounded subdifferential on compacta. More precisely, for any compact set $C \subset \mathbb{R}^n$, it holds that

$$\sup_{x \in C} \sup_{\tilde{\nabla} r(x) \in \partial r(x)} \left\| \tilde{\nabla} r(x) \right\| < \infty.$$

TODO: This is not actually needed. Closed, proper, convex (on the domain) are enough.

3. The matrix-valued map $A\colon \mathbb{R}^p \to \mathbb{R}^{m\times n}$, and the vector valued map $b\colon \mathbb{R}^p \to \mathbb{R}^m$, are both (Borel-)measurable.

4. The random matrix $A(\xi)$ has bounded norm Frobenius norm, $\|A(\xi)\|_F < \infty$, almost surely. TODO: Remove this once you fixed the definition of h_k.

5. The sequence $(\gamma_k)_{k\in\mathbb{N}}$ is nondecreasing and unbounded.

6. There exists at least one feasible point $x \in \mathcal{X}$.

Algorithm 2 work with penalty functions $(\pi_k)_{k\in\mathbb{N}}$. These will always take the form $\pi_k(x) := \mathbb{E}(h_k(A(\xi)x - b(\xi)))$, where, for all $k \in \mathbb{N}$, we let $(h_k)_{k\in\mathbb{N}}$ be any sequence of functions $\mathbb{R}^m \to \mathbb{R}$ with the following properties:

1. For all $k \in \mathbb{N}$, h_k is convex and L_{h_k}-smooth for some $L_{h_k} \in (0, \infty)$ (TODO: Think convex differentiable is enough).

2. $h_k(x) \geq \|(x)_+\|_1$ for all $x \in \mathbb{R}^m$ and $k \in \mathbb{N}$.

3. There exists a sequence $(\alpha_k)_{k\in\mathbb{N}} \subset (0, \infty)$ such that $\sup_{k\in\mathbb{N}} \gamma_k\alpha_k < \infty$ and $h_k(x) \leq \alpha_k$ for all $x \in \mathbb{R}^m_{\leq 0}$. In particular, for all feasible points $\tilde{x} \in \mathcal{X}$, it holds that $\pi_k(\tilde{x}) \leq \alpha_k$ for all $k \in \mathbb{N}$.

4. The gradients of $(h_k)_{k\in\mathbb{N}}$ are bounded uniformly:

$$\sup_{k\in\mathbb{N}} \sup_{x\in\mathbb{R}^m} \nabla h_k(x) < \infty.$$

Finally, for all $k \in \mathbb{N}$, we define the sequence of unconstrained optimization problems

$$\min_{x\in\mathbb{R}^n} \left\{ f_k(x) := f(x) + \gamma_k\pi_k(x) \right\}, \tag{P$^\mathrm{k}$}$$

where $\gamma_k \in (0, \infty)$. Since proximal methods seperate the smooth part of the above objective, given by $x \mapsto \mathbb{E}(F_\xi(x)) + \gamma_k\pi_k(x)$, from the nonsmooth part, $x \mapsto r(x)$, it is useful to also define the functions

$$\psi_k(x) := \mathbb{E}(F_\xi(x)) + \gamma_k\pi_k(x),$$

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

Assumption 1 has multiple useful implications, which are captured by the following lemma.

Lemma 3.1. Assumption 1 implies the following:

1. The objective f of (P) is μ-strongly convex and subdifferentiable. Further, the subgradients of f are bounded on compacta: For any compact set $C \in \mathbb{R}^n$, it holds that
$$\sup_{x \in C} \sup_{\tilde{\nabla} f(x) \in \partial f(x)} \left\| \tilde{\nabla} f(x) \right\| < \infty.$$

2. The objectives $(f_k)_{k \in \mathbb{N}}$ of ($\mathrm{P^k}$) are μ-strongly convex and subdifferentiable. Further, for any compact set $C \in \mathbb{R}^n$, it holds that
$$\sup_{x \in C} \sup_{\tilde{\nabla} f_k(x) \in \partial f_k(x)} \left\| \tilde{\nabla} f_k(x) \right\| = \mathcal{O}(\gamma_k),$$
for all $k \in \mathbb{N}$.

3. The functions $(\psi_k)_{k \in \mathbb{N}}$ are μ-strongly convex and L_k-smooth, where $L_k := L + \gamma_k L_{h_k}$. Further, for any compact set $C \subset \mathbb{R}^n$, it holds that
$$\sup_{x \in C} \left\| \nabla \psi_k(x) \right\| = \mathcal{O}(\gamma_k).$$

4. The gradients of $(\pi_k)_{k \in \mathbb{N}}$ are uniformly bounded: There exists $G \in (0, \infty)$ such that
$$\sup_{k \in \mathbb{N}} \sup_{x \in \mathbb{R}^n} \|\nabla \pi_k(x)\| \leq G.$$

5. There exists a unique solution $x^\star \in \mathcal{X}$ for (P) and, for all $k \in \mathbb{N}$, there exists a unique solution $x_k^\star \in \mathbb{R}^n$ for ($\mathrm{P^k}$).

Proof. □

The fact that $x_k^\star$ must not be feasible introduces difficulties that prevent the use of standard arguments from the SGD literature to analyse convergence. Our proof methods combine approaches from recent works, mainly the already mentioned [3], as well as [6]. The latter paper investiages *stochastic optimization problems under distributional drift*. While these kinds of problems are not exactly the same as the ones we are working with, the two settings do indeed exhibit striking resemblences. Namely, Cutler et al. [6] investigated a sequence of time-dependent composite problems of the form
$$\min_{x \in \mathbb{R}^n} g_t(x) + r_t(x),$$
where, for all $t \in \mathbb{N}$, g_t is smooth strongly convex, and r_t is convex. Comparing to

our problem (P^k), rewritten as

$$\min_{x\in\mathbb{R}^n} \left\{ f_k(x) = \psi_k(x) + r(x) \right\}, \tag{P^k}$$

we see that the two settings almost match, except for one important difference: Cutler et al. [6] assume that the smoothness constant (as well as the strong convexity constant) of g_t is independent of $t \in \mathbb{N}$ (TODO: Change this when the smoothness assumption on h_k is dropped). In our case however, the smoothness constant of ψ_k depends on $\gamma_k h_k$, and will generally grow over time. This will lead to a constraint on the decay rate of our step size schedule $(\eta_k)_{k\in\mathbb{N}}$. Namely, we will need $\eta_k \in (0, 1/L_k)$ with L_k the smoothness constant of ψ_k, for all $k \in \mathbb{N}$. Nevertheless, adapting the proof technique to our case will not be very complicated, and will help us establish a useful inequality. This inequality can then be used to adapt the proof strategy from Nedić et al. [3] (where the authors did not use proximal maps in their algorithm) to our proximal method.

3.1. Almost sure convergence

In this section we will establish conditions under which we can guarantee almost sure convergence of the sequence of iterates $(x_k)_{k\in\mathbb{N}}$. The proof will also yield convergence in expectation of $(x_k)_{k\in\mathbb{N}}$ to $x^\star$ along a subsequence. The main use of the almost sure convergence result will be that we will have conditions on the stepsizes $(\eta_k)_{k\in\mathbb{N}}$ and the penalty parameters $(\gamma_k)_{k\in\mathbb{N}}$ to ensure that $(x_k)_{k\in\mathbb{N}}$ is bounded with probability one. This, together with some of the results we prove along the way, will come in very handy in the subsequent analysis of the quantitative convergence rates of our methods.

The proof for almost sure convergence hinges on two technical lemmata. The first is the well-known Robbins-Siegmund lemma, which provides a general sufficient condition to guarantee almost sure convergence of so-called "almost supermartingales".

Lemma 3.2 (Robbins-Siegmund). Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be an increasing sequence of σ-algebras and v_n, a_n, b_n, c_n be nonnegative $\mathcal{F}_n$-measurable random variables. If, for all $n \in \mathbb{N}$,

$$\mathbb{E}(v_{n+1} \mid \mathcal{F}_n) \le v_n(1 + a_n) + b_n - c_n, \tag{3.1}$$

and $\sum_{n=1}^\infty a_n < \infty$, $\sum_{n=1}^\infty b_n < \infty$ a. s., then with probability one, $(v_n)_{n\in\mathbb{N}}$ is convergent and it holds that $\sum_{n=1}^\infty c_n < \infty$.

Proof. See [7]. □

Our goal for the rest of this section is to derive a recursive inequality for the sequence $\|x_k - x^\star\|^2$, which resembles (3.1). In order to do this, we will first analyze the behavior of the sequence of solutions $(x_k^\star)_{k\in\mathbb{N}}$ to the sequence of unconstrained problems ($\mathrm{P^k}$). We can then use this to "build a bridge" between $(x_k)_{k\in\mathbb{N}}$ and $x^\star$, by considering their respective relationships to the sequence $(x_k^\star)_{k\in\mathbb{N}}$.

The first step towards analyzing the convergence of $(x_k^\star)$ will be to (locally) bound the distance $\operatorname{dist}(x, \mathcal{X})$ by (a term proportional to) the penalty $\pi_k(x)$. We will rely on an extension of a classic result by Hoffman [8], who analyzed the distance of points $x \in \mathbb{R}^n$ to the set of solutions of linear systems of inequalities $Ax \leq b$ with $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Crucially, there always exists a constant $\tau \in (0, \infty)$, such that

$$\tau \operatorname{dist}(x, S) \leq \|(Ax - b)_+\|_\infty ,$$

where $S := \{\, y \in \mathbb{R}^n \mid Ay \leq b \,\}$. In [9], this result was then extended to the case of infinte systems of linear inequalities, which is the setting that we are interested in. However, before we can apply this inequality, we will first need to impose two assumptions. The first one concerns the joint distribution of $A(\xi)$ and $b(\xi)$. For $i \in \{\, 1, \dots, m \,\}$, $z \in \mathbb{R}^p$, we let $a_i(z) \in \mathbb{R}^n$, resp. $b_i(z) \in \mathbb{R}$, denote the i-th row of $A(z)$, resp. $b(z)$.

Assumption 2. There exists a compact set $C \in \mathbb{R}^{n+1}$ such that $0 \notin C$ and the set

$$\{\, (a_i(\xi), b_i(\xi)) \mid i \in \{\, 1, \dots, m \,\} \,\}$$

is contained in C almost surely. TODO: This implies the bounded second moment assumption on $\mathbb{E}\,\|A(\xi)\|_F^2$.

The second assumption we make concerns the structure of the solution set $\mathcal{X}$ and is a very common assumption in optimization.

Assumption 3. There exists a *Slater point*, i. e. a point $x \in \mathcal{X}$ such that $A(\xi)x < b(\xi)$ almost surely.

We will now formulate and prove the lemma.

Lemma 3.3 (Subfeasibility bound via penalty functions)**.** Assume assumptions 2 and 3 hold, let $B \subset \mathbb{R}^n$ be a bounded subset. Then there exists a constant $\tau \in (0, \infty)$ such that

$$\tau \operatorname{dist}(x, \mathcal{X}) \leq \pi_k(x),$$

for all $x \in B$ and $k \in \mathbb{N}$.

Proof. □

Theorem 3.4 (Convergence of $x_k^\star$)**.** Assume assumptions 2 and 3 hold. Then, for all $k \in \mathbb{N}$, it holds that

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 + \frac{\mu}{2}\|x^\star - \Pi_{\mathcal{X}}(x_k^\star)\|^2 + (\tau\gamma_k - M)\ \mathrm{dist}(x_k^\star, \mathcal{X}) \leq \gamma_k \alpha_k,$$

where $M, \tau \in (0, \infty)$ are constants. In particular,

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

Proof. Let $k \in \mathbb{N}$. By optimality of $x_k^\star$ for f_k, there exists $0 \in \partial f_k(x_k^\star)$. Hence, by strong convexity, we obtain

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 \leq f_k(x^\star) - f_k(x_k^\star) = f(x^\star) - f(x_k^\star) + \gamma_k \pi_k(x^\star) - \gamma_k \pi_k(x_k^\star). \tag{3.2}$$

We can write

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

where we again used strong convexity in the second step. Since $\Pi_{\mathcal{X}}(x_k^\star) \in \mathcal{X}$ and $x^\star$ is optimal for f on $\mathcal{X}$, it holds that

$$\langle \tilde{\nabla} f(x^\star), \Pi_{\mathcal{X}}(x_k^\star) - x^\star \rangle \geq 0,$$

and thus

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

To continue with the proof, we will need to first show that $(x_k^\star)_{k \in \mathbb{N}}$ is bounded: By strong convexity and optimality of $x_k^\star$ for f_k, for any feasible $x \in \mathcal{X}$, it holds that

$$\begin{aligned} \frac{\mu}{2}\left\|x - x_k^\star\right\|^2 &\leq f_k(x) - f_k(x_k^\star) \\ &= f(x) + \gamma_k \pi_k(x) - f_k(x_k^\star) \\ &\leq f(x) + \gamma_k \alpha_k - f_k(x_k^\star) \\ &\leq f(x) + \gamma_k \alpha_k - f_{\min}, \end{aligned}$$

where $f_{\min}$ is the minimum value of f, which exists by strong convexity. Noting that $\lim_{k \to \infty} \gamma_k \alpha_k = 0$ per one of our assumptions, it follows that $(x_k^\star)_{k \in \mathbb{N}}$ must be a bounded sequence. It follows that $(\Pi_{\mathcal{X}}(x_k^\star))_{k\in\mathbb{N}}$ is bounded as well (continuous functions map compact sets to compact sets), and therefore there exists a constant $M \in (0,\infty)$ such that $M = \sup_{k\in\mathbb{N}} \left\|\tilde{\nabla} f(\Pi_{\mathcal{X}}(x_k^\star))\right\|$. Hence, by convexity of f and Cauchy-Schwarz,

$$f(\Pi_{\mathcal{X}}(x_k^\star)) - f(x_k^\star) \le \langle \tilde{\nabla} f(\Pi_{\mathcal{X}}(x_k^\star)), \Pi_{\mathcal{X}}(x_k^\star) - x_k^\star\rangle \le M \operatorname{dist}(x_k^\star).$$

We now obain

$$f(x^\star) - f(x_k^\star) \le -\frac{\mu}{2}\|x^\star - \Pi_{\mathcal{X}}(x_k^\star)\|^2 + M \operatorname{dist}(x_k^\star, \mathcal{X}).$$

Plugging this into (3.2), we obtain

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 \le -\frac{\mu}{2}\|x^\star - \Pi_{\mathcal{X}}(x_k^\star)\|^2 + M \operatorname{dist}(x_k^\star, \mathcal{X}) + \gamma_k \pi_k(x^\star) - \gamma_k \pi_k(x_k^\star).$$

Now, using our lower bound on $\pi_k(x_k^\star)$ from lemma 3.3, and combining terms, we arrive at the inequality

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 \le -\frac{\mu}{2}\|x^\star - \Pi_{\mathcal{X}}(x_k^\star)\|^2 + (M - \gamma_k\tau) \operatorname{dist}(x_k^\star, \mathcal{X}) + \gamma_k \pi_k(x^\star).$$

Using $\pi_k(x^\star) \le \alpha_k$ and rearranging, we obtain

$$\frac{\mu}{2}\|x^\star - x_k^\star\|^2 + \frac{\mu}{2}\|x^\star - \Pi_{\mathcal{X}}(x_k^\star)\|^2 + (\gamma_k\tau - M) \operatorname{dist}(x_k^\star, \mathcal{X}) \le \gamma_k \alpha_k.$$

The asymptotic rate for $\|x^\star - x_k^\star\|^2$ now follows. For the bound on $\operatorname{dist}(x_k^\star, \mathcal{X})$, we let K be large enough such that $\gamma_k\tau > M$ for all $k \ge K$. Dividing by γ_k on both sides and using the nonnegativity of the other terms on the left-hand side, we get

$$c \cdot \operatorname{dist}(x_k^\star, \mathcal{X}) \le \frac{\gamma_k\tau - M}{\gamma_k} \operatorname{dist}(x_k^\star, \mathcal{X}) \le \alpha_k,$$

for all $k \ge K$ and some constant $c \in (0,1)$, as desired. □

Having established the convergence of the sequence $(x_k^\star)_{k\in\mathbb{N}}$ to $x^\star$, we will now shift our attention to the iterates $(x_k)_{k\in\mathbb{N}}$ of algorithm 2. We begin with the following fundamental recursive inequality.

Lemma 3.5 (One-step improvement)**.** Let $\rho \in (0,1)$ and $\eta_k \in (0, \rho L^{-1}]$ for all $k \in \mathbb{N}$. Then the iterates $(x_k)_{k\in\mathbb{N}}$ generated by algorithm 2 with step size schedule $(\eta_k)_{k\in\mathbb{N}}$

satisfy

$$2\eta_k \mathbb{E}_k(f_k(x_{k+1}) - f_k(x)) \le (1 - \mu\eta_k) \|x - x_k\|^2 - \mathbb{E}_k \|x - x_{k+1}\|^2 + \frac{2\eta_k^2}{1-\rho} \operatorname{Var}_k(g_k) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho}.$$

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

Proof. For $k \in \mathbb{N}$ we denote by g_k the stochastic gradient of ψ_k at x_k that is used in iteration k of algorithm 2. We also let $\psi(x) := \mathbb{E}(F_\xi(x))$ for $x \in \mathbb{R}^n$, so that $\psi_k = \psi + \gamma_k \pi_k$ and $\nabla\psi_k = \nabla\psi + \gamma_k \nabla\pi_k$. By L-smoothness of ψ (assumption 1), we have

$$\begin{aligned}
f_k(x_{k+1}) &= \psi_k(x_{k+1}) + r(x_{k+1}) \\
&= \psi(x_{k+1}) + \gamma_k \pi_k(x_{k+1}) + r(x_{k+1}) \\
&\le \psi(x_k) + \langle \nabla\psi(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 + \gamma_k \pi_k(x_{k+1}) + r(x_{k+1}) \\
&= \psi(x_k) + \langle \nabla\psi_k(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 + \gamma_k \pi_k(x_{k+1}) + r(x_{k+1}) \\
&\quad + \gamma_k \langle \nabla\pi_k(x_k), x_k - x_{k+1} \rangle.
\end{aligned}$$

By convexity of π_k, we further have

$$\pi_k(x_{k+1}) \le \pi_k(x_k) + \langle \nabla\pi_k(x_{k+1}), x_{k+1} - x_k \rangle,$$

and an application of Cauchy-Schwarz and Young's inequality yields

$$\pi_k(x_{k+1}) \le \pi_k(x_k) + \frac{\epsilon_k^{-1}}{2} \|\nabla\pi_k(x_{k+1})\|^2 + \frac{\epsilon_k}{2} \|x_{k+1} - x_k\|^2,$$

for any $\epsilon_k \in (0, \infty)$. Note that the gradients $(\nabla\pi_k)_{k\in\mathbb{N}}$ are bounded uniformly (lemma 3.1), hence there exists $G \in (0, \infty)$ such that

$$\pi_k(x_{k+1}) \le \pi_k(x_k) + \frac{\epsilon_k}{2} G^2 + \frac{\epsilon_k^{-1}}{2} \|x_{k+1} - x_k\|^2,$$

for any $\epsilon_k \in (0, \infty)$. With this, we can further bound $f_k(x_{k+1})$ by

$$
\begin{aligned}
f_k(x_{k+1}) &\leq \psi(x_k) + \langle \nabla \psi_k(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 + r(x_{k+1}) \\
&\qquad + \gamma_k \Big(\pi_k(x_k) + \frac{\epsilon_k}{2} G^2 + \frac{\epsilon_k^{-1}}{2} \|x_{k+1} - x_k\|^2 \Big) + \gamma_k \langle \nabla \pi_k(x_k), x_k - x_{k+1} \rangle. \\
&= \psi_k(x_k) + \langle \nabla \psi_k(x_k), x_{k+1} - x_k \rangle + \frac{L + \gamma_k \epsilon_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + r(x_{k+1}) \\
&\qquad + \frac{\gamma_k \epsilon_k G^2}{2} + \gamma_k \langle \nabla \pi_k(x_k), x_k - x_{k+1} \rangle.
\end{aligned}
$$

By another application of Cauchy-Schwarz and Young's inequality, we have for any $\epsilon_k \in (0, \infty)$

$$
\langle \nabla \pi_k(x_k), x_k - x_{k+1} \rangle \leq \frac{\epsilon_k}{2} G^2 + \frac{\epsilon_k^{-1}}{2} \|x_{k+1} - x_k\|^2 ,
$$

where again used the bound $\sup_{k \in \mathbb{N}} \sup_{x \in \mathbb{R}^n} \nabla \pi_k(x) \leq G$, and therefore we obtain

$$
\begin{aligned}
f_k(x_{k+1}) &\leq \psi_k(x_k) + \langle \nabla \psi_k(x_k), x_{k+1} - x_k \rangle + \frac{L + 2\gamma_k \epsilon_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + r(x_{k+1}) \\
&\qquad + \gamma_k \epsilon_k G^2
\end{aligned}
$$

for any $\epsilon_k \in (0, \infty)$. We let $z_k := \nabla \psi_k(x_k) - g_k$ be the error in the k-th stochastic gradient. By adding and subtracting $\langle z_k, x_{k+1} - x_k \rangle$ in the above inequality, we get

$$
\begin{aligned}
f_k(x_{k+1}) &\leq \psi_k(x_k) + \langle g_k, x_{k+1} - x_k \rangle + \frac{L + 2\gamma_k \epsilon_k^{-1}}{2} \|x_{k+1} - x_k\|^2 + r(x_{k+1}) \\
&\qquad + \gamma_k \epsilon_k G^2 + \langle z_k, x_{k+1} - x_k \rangle.
\end{aligned}
$$

By yet another application of Cauchy-Schwarz and Young's inequality, we have

$$
\langle z_k, x_{k+1} - x_k \rangle \leq \frac{\delta_k}{2} \|z_k\|^2 + \frac{\delta_k^{-1}}{2} \|x_{k+1} - x_k\|^2 ,
$$

for all $\delta_k \in (0, \infty)$, and therefore

$$\begin{aligned}
f_k(x_{k+1}) &\le \psi_k(x_k) + \langle g_k, x_{k+1} - x_k\rangle + \frac{L + 2\gamma_k\epsilon_k^{-1} + \delta_k^{-1}}{2}\|x_{k+1} - x_k\|^2 + r(x_{k+1}) \\
&\quad + \gamma_k\epsilon_k G^2 + \frac{\delta_k}{2}\|z_k\|^2, \\
&= \psi_k(x_k) + r(x_{k+1}) + \langle g_k, x_{k+1} - x_k\rangle + \frac{1}{2\eta_k}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{L + 2\gamma_k\epsilon_k^{-1} + \delta_k^{-1} - \eta_k^{-1}}{2}\|x_{k+1} - x_k\|^2 + \gamma_k\epsilon_k G^2 + \frac{\delta_k}{2}\|z_k\|^2,
\end{aligned}$$

where in the last step we added and subtracted $(2\eta_k)^{-1}\|x_{k+1} - x_k\|^2$ and moved $r(x_{k+1})$ further forward. From the definition of the proximal operator, it follows that

$$\begin{aligned}
x_{k+1} &= \operatorname{prox}_{\eta_k r}(x_k - \eta_k g_k) \\
&= \arg\min_{x\in\mathbb{R}^n}\left\{ r(x) + \frac{1}{2\eta_k}\left\|x - (x_k - \eta_k g_k)\right\|^2 \right\} \\
&= \arg\min_{x\in\mathbb{R}^n}\left\{ r(x) + \langle g_k, x - x_k\rangle + \frac{1}{2\eta_k}\|x - x_k\|^2 \right\},
\end{aligned}$$

where the last step follows from expanding the square and dropping the constant term $\eta_k^2\left\|g_k\right\|^2$ from the minimization. The function $x \mapsto r(x) + \langle g_k, x - x_k\rangle + (2\eta_k)^{-1}\|x - x_k\|^2$ is $(2\eta_k)^{-1}$-strongly convex and minimized by x_{k+1}. Thus, comparing with our previous bound on $f_k(x_{k+1})$, we can conclude

$$\begin{aligned}
f_k(x_{k+1}) &\le \psi_k(x_k) + r(x) + \langle g_k, x - x_k\rangle + \frac{1}{2\eta_k}\|x - x_k\|^2 - \frac{1}{2\eta_k}\|x - x_{k+1}\|^2 \\
&\quad + \frac{L + 2\gamma_k\epsilon_k^{-1} + \delta_k^{-1} - \eta_k^{-1}}{2}\|x_{k+1} - x_k\|^2 + \gamma_k\epsilon_k G^2 + \frac{\delta_k}{2}\|z_k\|^2,
\end{aligned}$$

for all $x \in \mathbb{R}^n$. By μ-strong convexity of ψ_k, we have

$$\begin{aligned}
\psi_k(x_k) + r(x) + \langle g_k, x - x_k\rangle &= \psi_k(x_k) + r(x) + \langle \nabla\psi_k(x_k), x - x_k\rangle + \langle z_k, x_k - x\rangle \\
&\le \psi_k(x) - \frac{\mu}{2}\|x - x_k\|^2 + r(x) + \langle z_k, x_k - x\rangle \\
&= f_k(x) - \frac{\mu}{2}\|x - x_k\|^2 + \langle z_k, x_k - x\rangle,
\end{aligned}$$

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

$$
\begin{aligned}
f_k(x_{k+1}) \le f_k(x) &+ \Big(\frac{1}{2\eta_k} - \frac{\mu}{2}\Big)\|x - x_k\|^2 + \langle z_k, x_k - x\rangle - \frac{1}{2\eta_k}\|x - x_{k+1}\|^2 \\
&+ \frac{L + 2\gamma_k\epsilon_k^{-1} + \delta_k^{-1} - \eta_k^{-1}}{2}\|x_{k+1} - x_k\|^2 + \gamma_k\epsilon_k G^2 + \frac{\delta_k}{2}\|z_k\|^2,
\end{aligned}
\tag{3.3}
$$

Fix $\alpha \in (0, 1 - \rho)$ and define $\epsilon_k := 2\alpha^{-1}\eta_k\gamma_k$. Then

$$
(L + 2\gamma_k\epsilon_k^{-1})\eta_k = \Big(L + 2\gamma_k\frac{\alpha}{2\eta_k\gamma_k}\Big)\eta_k = L\eta_k + \alpha \le \rho + \alpha < 1,
$$

where we used that $\eta_k \le \rho L^{-1}$, which holds per assumption. Choosing

$$
\delta_k := \frac{\eta_k}{1 - (L\eta_k + \alpha)} \in (0, \infty)
$$

therefore yields

$$
L + 2\gamma_k\epsilon_k^{-1} + \delta_k^{-1} - \eta_k^{-1} = L + \frac{\alpha}{\eta_k} - \frac{1 - (L\eta_k + \alpha)}{\eta_k} - \frac{1}{\eta_k} = 0.
$$

Hence, we can drop the $\|x_{k+1} - x_k\|^2$ term from (3.3) and get

$$
\begin{aligned}
f_k(x_{k+1}) \le f_k(x) &+ \Big(\frac{1}{2\eta_k} - \frac{\mu}{2}\Big)\|x - x_k\|^2 + \langle z_k, x_k - x\rangle - \frac{1}{2\eta_k}\|x - x_{k+1}\|^2 \\
&+ 2\alpha^{-1}\eta_k\gamma_k^2 G^2 + \frac{\eta_k}{2(1 - (L\eta_k + \alpha))}\|z_k\|^2.
\end{aligned}
$$

Subtracting by $f_k(x)$ and multiplying both sides by $2\eta_k$ yields

$$
\begin{aligned}
2\eta_k(f_k(x_{k+1}) - f_k(x)) \le (1 - \mu\eta_k)\|x - x_k\|^2 &+ 2\eta_k\langle z_k, x_k - x\rangle - \|x - x_{k+1}\|^2 \\
&+ 4\alpha^{-1}\eta_k^2\gamma_k^2 G^2 + \frac{\eta_k^2}{1 - (L\eta_k + \alpha)}\|z_k\|^2.
\end{aligned}
$$

For the specifc choice $\alpha := (1 - \rho)/2$, it holds that $L\eta_k + \alpha \le \rho + (1 - \rho)/2 = (1 + \rho)/2$. Hence, we arrive at

$$
\begin{aligned}
2\eta_k(f_k(x_{k+1}) - f_k(x)) \le (1 - \mu\eta_k)\|x - x_k\|^2 &+ 2\eta_k\langle z_k, x_k - x\rangle - \|x - x_{k+1}\|^2 \\
&+ \frac{8G^2\eta_k^2\gamma_k^2}{1 - \rho} + \frac{2\eta_k^2}{1 - \rho}\|z_k\|^2.
\end{aligned}
\tag{3.4}
$$

By properties of conditional expectation and the definition of stochastic gradients, we know that $\mathbb{E}_k\langle z_k, x_k - x\rangle = 0$, and thus, applying conditional expectations to both sides of the above inequality, we obtain

$$\begin{aligned}2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x)) \leq (1-\mu\eta_k)\left\|x - x_k\right\|^2 - \mathbb{E}_k\left\|x - x_{k+1}\right\|^2 \\ + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho} + \frac{2\eta_k^2}{1-\rho}\mathbb{E}_k\left\|z_k\right\|^2.\end{aligned}$$

for all $x \in \mathbb{R}^n$. Since $\mathbb{E}_k\left\|z_k\right\|^2 = \mathbb{E}_k\left\|g_k - \nabla\psi_k(x_k)\right\|^2$ and $\mathbb{E}_k(g_k) = \nabla\psi_k(x_k)$ per our definition of stochastic gradients, it holds that $\mathbb{E}_k\left\|z_k\right\|^2 = \mathrm{Var}_k(g_k)$. We have thus arrived at the desired result. □

After rearranging the inequality from lemma 3.5, setting $x = x^\star$, and dropping the $-\mu\eta_k$ term, we get

$$\mathbb{E}_k\left\|x^\star - x_{k+1}\right\|^2 \leq \left\|x^\star - x_k\right\|^2 + \frac{2\eta_k^2}{1-\rho}\mathrm{Var}_k(g_k) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho} - 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x^\star)). \tag{3.5}$$

This is not yet quite in the form needed to apply lemma 3.2. For one, we need to bound the noise term $\mathrm{Var}_k(g_k)$. Second, we cannot in general guarantee that $\mathbb{E}_k(f_k(x_{k+1}) - f_k(x^\star)) \geq 0$. However, as we show in the next lemma, we can find a lower bound that involves a nonnegative term plus a small negative term.

Lemma 3.6. Assume that assumptions 2 and 3 hold, and let $\eta_k \in (0, \rho L^{-1}]$ for all $k \in \mathbb{N}$ and some $\rho \in (0,1)$. Then the iterates $(x_k)_{k\in\mathbb{N}}$ generated by algorithm 2 with step size schedule $(\eta_k)_{k\in\mathbb{N}}$ satisfy

$$\begin{aligned}\mathbb{E}_k\left\|x^\star - x_{k+1}\right\|^2 \leq (1-\mu\eta_k)\left\|x^\star - x_k\right\|^2 + \frac{2\eta_k^2}{1-\rho}\mathrm{Var}_k(g_k) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho} \\ + \mathcal{O}(\eta_k\gamma_k\alpha_k) - 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)),\end{aligned}$$

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

Proof. Let $k \in \mathbb{N}$. From lemma 3.5, we have

$$\begin{aligned}2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x^\star)) \leq (1-\mu\eta_k)\left\|x^\star - x_k\right\|^2 - \mathbb{E}_k\left\|x^\star - x_{k+1}\right\|^2 + \frac{2\eta_k^2}{1-\rho}\mathrm{Var}_k(g_k) \\ + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho}.\end{aligned}$$

Write

$$f_k(x_{k+1}) - f_k(x^\star) = (f_k(x_{k+1}) - f_k(x_k^\star)) + (f_k(x_k^\star) - f_k(\Pi_{\mathcal{X}}(x_k^\star))) \\ + (f_k(\Pi_{\mathcal{X}}(x_k^\star)) - f_k(x^\star)). \tag{3.6}$$

By definition, it holds that

$$\begin{aligned} f_k(\Pi_{\mathcal{X}}(x_k^\star)) - f_k(x^\star) &= f(\Pi_{\mathcal{X}}(x_k^\star)) + \gamma_k \pi_k(\Pi_{\mathcal{X}}(x_k^\star)) - f(x^\star) - \gamma_k \pi_k(x^\star) \\ &\geq f(\Pi_{\mathcal{X}}(x_k^\star)) - f(x^\star) - \gamma_k \pi_k(x^\star) \\ &\geq -\gamma_k \pi_k(x^\star), \end{aligned}$$

where the last step follows from the fact that $x^\star$ minimizes f on $\mathcal{X}$ and $\Pi_{\mathcal{X}}(x_k^\star) \in \mathcal{X}$. Further, feasibility implies $\pi_k(x^\star) \leq \alpha_k$. Hence, combining with (3.6), we have

$$f_k(x_{k+1}) - f_k(x^\star) \geq (f_k(x_{k+1}) - f_k(x_k^\star)) + f_k(x_k^\star) - f_k(\Pi_{\mathcal{X}}(x_k^\star)) - \gamma_k \alpha_k.$$

To analyze $f_k(x_k^\star) - f_k(\Pi_{\mathcal{X}}(x_k^\star))$, we first use convexity and Cauchy-Schwarz to get

$$f_k(x_k^\star) - f_k(\Pi_{\mathcal{X}}(x_k^\star)) \geq \langle \tilde{\nabla} f_k(\Pi_{\mathcal{X}}(x_k^\star)), x_k^\star - \Pi_{\mathcal{X}}(x_k^\star) \rangle \geq - \left\| \tilde{\nabla} f_k(\Pi_{\mathcal{X}}(x_k^\star)) \right\| \operatorname{dist}(x_k^\star, \mathcal{X}).$$

The sequence $(x_k^\star)_{k\in\mathbb{N}}$ converges to $x^\star$, which implies, by continuity of the projection map, $\lim_{k\to\infty} \Pi_{\mathcal{X}}(x_k^\star) = \Pi_{\mathcal{X}}(x^\star) = x^\star$. In particular, $(\Pi_{\mathcal{X}}(x_k^\star))_{k\in\mathbb{N}}$ is bounded, so by lemma 3.1, it follows that $\sup_{k\in\mathbb{N}} \left\| \tilde{\nabla} f_k(\Pi_{\mathcal{X}}(x_k^\star)) \right\| \leq c\gamma_k$ for some $c \in (0, \infty)$. Hence,

$$f_k(x_k^\star) - f_k(\Pi_{\mathcal{X}}(x_k^\star)) \geq -c\gamma_k \operatorname{dist}(x_k^\star, \mathcal{X}).$$

By theorem 3.4, it holds that there exist $K \in \mathbb{N}$ and $c' \in (0, \infty)$ such that, for all $k \geq K$, $\operatorname{dist}(x_k^\star, \mathcal{X}) \leq c' \alpha_k$. Therefore,

$$f_k(x_k^\star) - f_k(\Pi_{\mathcal{X}}(x_k^\star)) \geq -cc' \cdot \gamma_k \alpha_k$$

and

$$\begin{aligned} f_k(x_{k+1}) - f_k(x^\star) &\geq f_k(x_{k+1}) - f_k(x_k^\star) - cc' \cdot \gamma_k \alpha_k - \gamma_k \alpha_k \\ &= f_k(x_{k+1}) - f_k(x_k^\star) - (1 + cc')\, \gamma_k \alpha_k, \end{aligned}$$

for all $k \geq K$. Plugging this into our original estimate, we have

$$2\eta_k\big(\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)) - (1+cc')\gamma_k\alpha_k\big) \leq (1-\mu\eta_k)\left\|x^\star - x_k\right\|^2 - \mathbb{E}_k\left\|x^\star - x_{k+1}\right\|^2 + \frac{2\eta_k^2}{1-\rho}\operatorname{Var}_k(g_k) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho},$$

and thus

$$\mathbb{E}_k\left\|x^\star - x_{k+1}\right\|^2 \leq (1-\mu\eta_k)\left\|x^\star - x_k\right\|^2 + \frac{2\eta_k^2}{1-\rho}\operatorname{Var}_k(g_k) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho} + 2(1+cc')\,\eta_k\gamma_k\alpha_k - 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)),$$

which directly implies the claim. □

Combining the above lemma with (3.5), we now have

$$\mathbb{E}_k\left\|x^\star - x_{k+1}\right\|^2 \leq \left\|x^\star - x_k\right\|^2 + \frac{2\eta_k^2}{1-\rho}\operatorname{Var}_k(g_k) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho} + O(\eta_k\gamma_k\alpha_k) - 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)).$$

The term $2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star))$ is indeed nonnegative by optimality of $x_k^\star$ for f_k. Assuming that $\sum_{k=1}^\infty \eta_k\gamma_k\alpha_k < \infty$ and $\sum_{k=1}^\infty \eta_k^2\gamma_k^2 < \infty$, we are almost ready to apply lemma 3.2. The following lemma will give a bound on the gradient noise $\operatorname{Var}_k(g_k)$.

Lemma 3.7 (Bound on gradient noise)**.** In the situation of algorithm 2, there exists a constant $\sigma^2 \in (0, \infty)$ such that

$$\operatorname{Var}_k(g_k) \leq \frac{2L^2}{\beta_k}\left\|x_k - x^\star\right\|^2 + \frac{1+\gamma_k^2}{\beta_k}\sigma^2,$$

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

Proof. Let $k \in \mathbb{N}$ and define $\tilde{g}_k := \nabla F_\xi(x_k) + \gamma_k A(\xi)^\top \nabla h_k(A(\xi)x_k - b(\xi))$. We have

$$\begin{aligned}
\operatorname{Var}_k(g_k) &= \frac{1}{\beta_k}\operatorname{Var}_k(\tilde{g}_k) \\
&\leq \frac{1}{\beta_k}\mathbb{E}_k\left\|\tilde{g}_k\right\|^2 \\
&\leq \frac{1}{\beta_k}\Big(\|\nabla F_\xi(x_k)\|^2 + \gamma_k^2\,\mathbb{E}_k\left\|A(\xi)^\top\nabla h_k(A(\xi)x_k - b(\xi))\right\|^2\Big).
\end{aligned}$$

Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2\ \forall a, b \in \mathbb{R}$, and the (almost sure) L-

smoothness of $x \mapsto F_\xi(x)$, we have

$$\begin{aligned}
\|\nabla F_\xi(x_k)\|^2 &= \left\|\nabla F_\xi(x_k) - \nabla F_\xi(x^\star) + \nabla F_\xi(x^\star)\right\|^2 \\
&\leq 2\Big(\left\|\nabla F_\xi(x_k) - \nabla F_\xi(x^\star)\right\|^2 + \mathbb{E}_k \left\|\nabla F_\xi(x^\star)\right\|^2 \Big) \\
&\leq 2L^2 \left\|x_k - x^\star\right\|^2 + 2\,\mathbb{E}\left\|\nabla F_\xi(x^\star)\right\|^2,
\end{aligned}$$

where in the last step we also used that x_k is $\mathcal{F}_k$-measurable and ξ is independent of $\mathcal{F}_k$. By one of our assumptions, we can find a point $x \in \mathbb{R}^n$ such that $\mathbb{E}\|F_\xi(x)\|^2 < \infty$. Hence, using smoothness and the inequality $(a+b)^2 \leq 2a^2 + 2b^2 \;\; \forall a, b \in \mathbb{R}$ again, there exists a constant $M^2 \in (0, \infty)$ such that

$$\begin{aligned}
\mathbb{E}\left\|\nabla F_\xi(x^\star)\right\|^2 &= \mathbb{E}\left\|\nabla F_\xi(x^\star) - \nabla F_\xi(x) + \nabla F_\xi(x)\right\|^2 \\
&\leq 2L^2 \left\|x^\star - x\right\|^2 + 2\,\mathbb{E}\|F_\xi(x)\|^2 \\
&\leq \frac{1}{2}M^2.
\end{aligned}$$

Therefore, combining with the previous inequality, we get the bound

$$\|\nabla F_\xi(x_k)\|^2 \leq 2L^2 \left\|x_k - x^\star\right\|^2 + M^2.$$

Now, we can use the fact that $\|Av\| \leq \|A\|_F \|v\|$ for any matrix A and vector v, to find

$$\mathbb{E}_k \left\|A(\xi)^\top \nabla h_k(A(\xi)x_k - b(\xi))\right\|^2 \leq \mathbb{E}_k \Big(\|A(\xi)\|_F^2 \|\nabla h_k(A(\xi)x_k - b(\xi))\|^2 \Big).$$

Since, per assumption, it holds that a) the family $(h_k)_{k\in\mathbb{N}}$ has uniformly bounded gradients, and b) $\mathbb{E}\|A(\xi)\|_F^2 < \infty$, there exists a constant $\tilde{M}^2 \in (0, \infty)$ such that

$$\mathbb{E}_k \left\|A(\xi)^\top \nabla h_k(A(\xi)x_k - b(\xi))\right\|^2 \leq \tilde{M}^2.$$

Putting everything together, we obtain

$$\mathrm{Var}_k(g_k) \leq \frac{1}{\beta_k}\Big(2L^2 \left\|x_k - x^\star\right\|^2 + M^2 + \gamma_k^2 \tilde{M}^2\Big) = \frac{2L^2}{\beta_k}\left\|x_k - x^\star\right\|^2 + \frac{1}{\beta_k}(M^2 + \gamma_k^2 \tilde{M}^2).$$

Setting $\sigma^2 := \max(M^2, \tilde{M}^2)$ yields the desired result. □

We are now ready to prove the main theorem of this section.

Theorem 3.8 (Almost sure convergence)**.** Assume assumptions 2 and 3 hold. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by algorithm 2 with parameters $(\eta_k)_{k\in\mathbb{N}}$, $(\gamma_k)_{k\in\mathbb{N}}$, $(h_k)_{k\in\mathbb{N}}$ that satisfiy

1. $\sum_{k=1}^{\infty} \eta_k = \infty$ and there exists $\rho \in (0,1)$ such that $\eta_k \leq \rho L^{-1}$ for all $k \in \mathbb{N}$.

2. $\sum_{k=1}^{\infty} \eta_k \gamma_k \alpha_k < \infty$.

3. $\sum_{k=1}^{\infty} \eta_k^2 \gamma_k^2 < \infty$.

Then $\|x_k - x^\star\|$ converges almost surely and $\liminf_{k\to\infty} \mathbb{E}\,\|x_k - x^\star\|^2 = 0$. In particular, $(x_k)_{k\in\mathbb{N}}$ is bounded almost surely.

Proof. By lemma 3.6 (dropping the $-\mu\eta_k$ term), we have

$$\begin{aligned}\mathbb{E}_k \left\|x^\star - x_{k+1}\right\|^2 \leq \left\|x^\star - x_k\right\|^2 + \frac{2\eta_k^2}{1-\rho}\operatorname{Var}_k(g_k) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho} \\ + O(\eta_k\gamma_k\alpha_k) - 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)),\end{aligned}$$

Lemma 3.7 lets us bound the variance by

$$\operatorname{Var}_k(g_k) \leq \frac{2L^2}{\beta_k}\left\|x_k - x^\star\right\|^2 + \frac{1+\gamma_k^2}{\beta_k}\sigma^2,$$

for a constant $M^2 \in (0,\infty)$ and all $k \in \mathbb{N}$. Thus, we have

$$\begin{aligned}\mathbb{E}_k \left\|x^\star - x_{k+1}\right\|^2 &\leq \left\|x^\star - x_k\right\|^2 + \frac{2\eta_k^2}{1-\rho}\Big(\frac{2L^2}{\beta_k}\left\|x_k - x^\star\right\|^2 + \frac{1+\gamma_k^2}{\beta_k}\sigma^2\Big) \\ &\quad + O(\eta_k\gamma_k\alpha_k) - 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)) \\ &= \Big(1 + 4L^2(1-\rho)^{-1}\beta_k^{-1}\eta_k^2\Big)\left\|x^\star - x_k\right\|^2 + 2\eta_k^2(1-\rho)^{-1}\frac{1+\gamma_k^2}{\beta_k}\sigma^2 \\ &\quad + O(\eta_k\gamma_k\alpha_k) - 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)),\end{aligned}$$

for all $k \in \mathbb{N}$. Define the nonegative sequences $(a_k)_{k\in\mathbb{N}}, (b_k)_{k\in\mathbb{N}}, (c_k)_{k\in\mathbb{N}}$ by

$$\begin{aligned}a_k &:= 4L^2(1-\rho)^{-1}\beta_k^{-1}\eta_k^2 \\ b_k &:= 2\eta_k^2(1-\rho)^{-1}\frac{1+\gamma_k^2}{\beta_k}\sigma^2 + O(\eta_k\gamma_k\alpha_k) \\ c_k &:= 2\eta_k\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)).\end{aligned}$$

Note that c_k is indeed nonnegative, since $x_k^\star$ minimizes f_k. The above inequality now takes the form

$$\mathbb{E}_k \left\|x^\star - x_{k+1}\right\|^2 \leq (1+a_k)\left\|x^\star - x_k\right\|^2 + b_k - c_k.$$

Our assumptions imply that $\sum_{k=1}^{\infty} a_k < \infty$ and $\sum_{k=1}^{\infty} b_k < \infty$, so we can apply lemma 3.2, which implies that with probability one the sequence $\|x^\star - x_k\|^2$ converges and $\sum_{k=1}^{\infty} \eta_k \mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star)) < \infty$. By the bounded convergence theorem (TODO: add ref), it further holds that

$$\begin{aligned}\infty > \mathbb{E}\Big(\sum_{k=1}^{\infty} \eta_k \mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star))\Big) &= \sum_{k=1}^{\infty} \eta_k \mathbb{E}(\mathbb{E}_k(f_k(x_{k+1}) - f_k(x_k^\star))) \\ &= \sum_{k=1}^{\infty} \eta_k \mathbb{E}(f_k(x_{k+1}) - f_k(x_k^\star)).\end{aligned}$$

Since $\sum_{k=1}^{\infty} \eta_k = \infty$, it must therefore hold that

$$\liminf_{k\to\infty} \mathbb{E}(f_k(x_{k+1}) - f_k(x_k^\star)) = 0.$$

Strong convexity of f_k and optimaliy of $x_k^\star$ for f_k imply

$$f_k(x_{k+1}) - f_k(x_k^\star) \geq \frac{\mu}{2} \left\| x_{k+1} - x_k^\star \right\|^2,$$

and thus $\liminf_{k\to\infty} \mathbb{E} \left\| x_{k+1} - x_k^\star \right\|^2 = 0$, which implies that

$$\liminf_{k\to\infty} \mathbb{E} \left\| x_k - x^\star \right\|^2 = 0,$$

since $x_k^\star$ converges to $x^\star$ (theorem 3.4). □

3.2. Convergence rates in expectation

Assumption 4. The sequence $(h_k)_{k\in\mathbb{N}}$ and the parameters $(\eta_k)_{k\in\mathbb{N}}$, $(\gamma_k)_{k\in\mathbb{N}}$ from algorithm 2 satisfiy

1. $\sum_{k=1}^{\infty} \eta_k = \infty$ and there exist constants $\rho \in (0,1)$ and $K \in \mathbb{N}$ such that $\eta_k \leq \rho L^{-1}$ for all $k \in \mathbb{N}, k \geq K$.

2. $\sum_{k=1}^{\infty} \eta_k \gamma_k \alpha_k < \infty$.

3. $\sum_{k=1}^{\infty} \eta_k^2 \gamma_k^2 < \infty$.

Lemma 3.9. Let assumptions 2 to 4 hold. Then there exists a deterministic sequence $(M_k^2)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ with $M_k^2 = \mathcal{O}(\beta_k^{-1}(1+\gamma_k^2) + \gamma_k^2)$, such that

$$2\eta_k \mathbb{E}(f_k(x_k) - f_k(x)) \leq (1 - \mu\eta_k)\mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + \eta_k^2 M_k^2,$$

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

Proof. Let $k \in \mathbb{N}, k \geq K$. Taking expectations on both sides of the inequality provided by lemma 3.5, we get

$$2\eta_k \mathbb{E}(f_k(x_{k+1}) - f_k(x)) \leq (1 - \mu\eta_k)\mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + \frac{2\eta_k^2}{1-\rho}\mathbb{E}(\mathrm{Var}_k(g_k)) + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho},$$

for all $x \in \mathbb{R}^n$. Recall that lemma 3.7 lets us bound the conditional variance $\mathrm{Var}_k(g_k)$ as follows:

$$\mathrm{Var}_k(g_k) \leq \frac{2L^2}{\beta_k}\left\|x_k - x^\star\right\|^2 + \frac{1+\gamma_k^2}{\beta_k}\sigma^2,$$

for a constant $\sigma^2 \in (0, \infty)$. Assumption 4 and theorem 3.8 imply that there exists a constant $\tilde{\sigma}^2 \in (0, \infty)$ such that $2L^2\|x_k - x^\star\|^2 \leq \tilde{\sigma}^2$ for all $k \in \mathbb{N}$ (almost surely). We can assume without loss of generality that $\sigma^2 \leq \tilde{\sigma}^2$, and since $1 + \gamma_k^2 > 1$, we have

$$\mathrm{Var}_k(g_k) \leq \frac{1}{\beta_k}\tilde{\sigma}^2 + \frac{1+\gamma_k^2}{\beta_k}\tilde{\sigma}^2 \leq \frac{2(1+\gamma_k^2)}{\beta_k}\tilde{\sigma}^2,$$

almost surely. Therefore, we obtain

$$2\eta_k \mathbb{E}(f_k(x_{k+1}) - f_k(x)) \leq (1 - \mu\eta_k)\mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + \frac{4\eta_k^2(1+\gamma_k^2)}{(1-\rho)\beta_k}\tilde{\sigma}^2,$$

for all $x \in \mathbb{R}^n$. By convexity of f_k and Cauchy-Schwarz, we have

$$f_k(x_{k+1}) \geq f_k(x_k) + \langle \tilde{\nabla} f_k(x_k), x_{k+1} - x_k \rangle \geq f_k(x_k) - \left\|\tilde{\nabla} f_k(x_k)\right\| \|x_{k+1} - x_k\|,$$

hence

$$2\eta_k \mathbb{E}(f_k(x_k) - f_k(x)) \leq (1 - \mu\eta_k)\mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + \frac{4\eta_k^2(1+\gamma_k^2)}{(1-\rho)\beta_k}\tilde{\sigma}^2 + 2\eta_k\mathbb{E}\left(\left\|\tilde{\nabla} f_k(x_k)\right\| \|x_{k+1} - x_k\|\right).$$

By definition of x_{k+1} as the solution to $\min_{x \in \mathbb{R}^n}\{r(x) + (2\eta_k)^{-1}\left\|x - (x_k - \eta_k g_k)\right\|^2\}$, with probability one there exists a subgradient $\tilde{\nabla} r(x_{k+1}) \in \partial r(x_{k+1})$, such that

$$\tilde{\nabla} r(x_{k+1}) + \frac{1}{\eta_k}(x_{k+1} - x_k + \eta_k g_k) = 0 \iff x_{k+1} - x_k = -\eta_k(\tilde{\nabla} r(x_{k+1}) + g_k)$$

Therefore,

$$\|x_{k+1} - x_k\| \le \eta_k \left(\left\|\tilde{\nabla} r(x_{k+1})\right\| + \left\|g_k\right\| \right) \text{ (a. s.)}.$$

Again using that $(x_k)_{k\in\mathbb{N}}$ is bounded almost surely, lemma 3.1 implies

$$\left\|g_k\right\| = \left\|\nabla\psi_k(x_k)\right\| = O(\gamma_k), \ \sup_{k\in\mathbb{N}} \left\|\tilde{\nabla} r(x_k)\right\| < \infty, \text{ and } \sup_{k\in\mathbb{N}} \left\|\tilde{\nabla} f_k(x_k)\right\| = O(\gamma_k),$$

and thus $\left\|\tilde{\nabla} f_k(x_k)\right\| \|x_{k+1} - x_k\| = \eta_k O(\gamma_k^2)$ almost surely. It follows that

$$\begin{aligned} 2\eta_k \mathbb{E}(f_k(x_k) - f_k(x)) \le (1 - \mu\eta_k)\mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + \frac{4\eta_k^2(1+\gamma_k^2)}{(1-\rho)\beta_k}\tilde{\sigma}^2 \\ + 2\eta_k^2 O(\gamma_k^2). \end{aligned}$$

Therefore, there exists a sequence $(M_k^2)_{k\in\mathbb{N}} \subset \mathbb{R}_{\ge 0}$ such that $M_k^2 = O(\beta_k^{-1}(1+\gamma_k^2) + \gamma_k^2)$ and

$$2\eta_k \mathbb{E}(f_k(x_k) - f_k(x)) \le (1 - \mu\eta_k)\mathbb{E}\|x - x_k\|^2 - \mathbb{E}\|x - x_{k+1}\|^2 + \eta_k^2 M_k^2,$$

for all $x \in \mathbb{R}^n$, as desired. □

Lemma 3.10. Let $\eta_k := 2/(\mu k)$ for all $k \in \mathbb{N}$, $\eta_0 := 2/\mu$. Then, for any $e \in (0, \infty)$, there exist sequences $(\gamma_k)_{k\in\mathbb{N}_0}$ and $(h_k)_{k\in\mathbb{N}_0}$ such that $\gamma_k = O(\log^e(k))$, $\alpha_k = O(1/k)$ and assumption 4 holds.

Proof. □

Lemma 3.11. Let assumptions 2 and 3 hold and define $\eta_k := 2/(\mu k)$ for all $k \in \mathbb{N}$, $\eta_0 := 2/\mu$. Then there exists a constant $k_0 \in \mathbb{N}$, as well as sequences $(M_k^2)_{k\in\mathbb{N}_0} \subset \mathbb{R}^n$, $(\beta_k)_{k\in\mathbb{N}_0}$, $(\gamma_k)_{k\in\mathbb{N}_0}$, $(h_k)_{k\in\mathbb{N}_0}$ such that the iterates $(\bar{x}_k)_{k\in\mathbb{N}}$ generated by algorithm 2 with parameters $(\eta_k)_{k\in\mathbb{N}_0}$, $(\gamma_k)_{k\in\mathbb{N}_0}$, $(\beta_k)_{k\in\mathbb{N}_0}$ satisfy

$$\mathbb{E}(f(\bar{x}_K) - f(x^\star)) \le \frac{S_{0,k_0}}{S_K}\mathbb{E}(f(\bar{x}_{0,k_0}) - f(x^\star)) + \frac{e_{k_0}}{2S_K} + \frac{\sum_{k=k_0}^{K-1} M_k^2}{2S_K} + \frac{\sum_{k=k_0}^{K-1} \eta_k^{-1}\gamma_k\alpha_k}{S_K},$$

for all $K \in \mathbb{N}$, where $S_K := \sum_{k=0}^{K-1} \eta_k^{-1}$, $e_{k_0} := \eta_{k_0-1}^{-2}\mathbb{E}\left\|x^\star - x_{k_0}\right\|^2$ and $M_k^2 = O(\beta_k^{-1}(1 + \gamma_k^2) + \gamma_k^2)$. Furthermore, there exist constants $k_1 \in \mathbb{N}$ and $\tau \in (0, \infty)$ such that

$$\mathbb{E}(\mathrm{dist}(\bar{x}_K, \mathcal{X})) \le \frac{S_{0,k_1}}{S_K}\mathbb{E}(\mathrm{dist}(\bar{x}_{0,k_1})) + \frac{\tau^{-1}d_{k_1}}{S_K} + \frac{\tau^{-1}\sum_{k=k_1}^{K-1}\gamma_k^{-1}M_k^2}{S_K} + \frac{2\tau^{-1}\sum_{k=k_1}^{K-1}\eta_k^{-1}\alpha_k}{S_K},$$

for all $K \in \mathbb{N}$ with $K \ge k_1$, where $d_{k_1} := \gamma_{k_1}^{-1}\eta_{k_1-1}^{-2}\mathbb{E}(\mathrm{dist}(x_{k_1}, \mathcal{X})^2)$.

Proof. Let $k_0 \in \mathbb{N}$ be large enough such that $2/(\mu k_0) \leq 1/(2L)$ and fix some $k \in \mathbb{N}$ with $k \geq k_0$. Lemma 3.10 implies that there exist parameters $(\gamma_k)_{k\in\mathbb{N}_0}$ and $(h_k)_{k\in\mathbb{N}_0}$ such that assumption 4 holds. Thus we can apply lemma 3.9, which implies

$$2\eta_k\mathbb{E}(f_k(x_k) - f_k(x)) \leq (1 - \mu\eta_k)\mathbb{E}\,\|x - x_k\|^2 - \mathbb{E}\,\|x - x_{k+1}\|^2 + \eta_k^2 M_k^2,$$

for any $x \in \mathbb{R}^n$. Assume now that $x \in \mathcal{X}$. By a property of π_k, the fact that $(x_k)_{k\in\mathbb{N}}$ is bounded almost surely (theorem 3.8), and lemma 3.3, we can deduce that there exists $\tau \in (0, \infty)$ such that

$$\begin{aligned} f_k(x_k) - f_k(x) &= f(x_k) - f(x) + \gamma_k(\pi_k(x_k) - \pi_k(x)) \\ &\geq f(x_k) - f(x) + \gamma_k(\tau \operatorname{dist}(x_k, \mathcal{X}) - \alpha_k) \end{aligned}$$

with probability one. Hence, we have

$$\begin{aligned} 2\eta_k\mathbb{E}(f(x_k) - f(x)) \leq (1 - \mu\eta_k)\mathbb{E}\,\|x - x_k\|^2 - \mathbb{E}\,\|x - x_{k+1}\|^2 + \eta_k^2 M_k^2 + 2\eta_k\gamma_k\alpha_k \\ - 2\tau\eta_k\gamma_k\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})), \end{aligned} \tag{3.7}$$

for all $x \in \mathcal{X}$. We can now closely follow the proof stategy of lemma 13 in [3] to arrive at our desired result. First, we will prove the inequality for $\mathbb{E}(f(x_k) - f(x^\star))$. Dropping the $-\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X}))$ term from (3.7), and multiplying both sides by η_k^{-2}, we get

$$\begin{aligned} 2\eta_k^{-1}\mathbb{E}(f(x_k) - f(x^\star)) \leq \eta_k^{-2}(1 - \mu\eta_k)\mathbb{E}\left\|x^\star - x_k\right\|^2 - \eta_k^{-2}\mathbb{E}\left\|x^\star - x_{k+1}\right\|^2 \\ + M_k^2 + 2\eta_k^{-1}\gamma_k\alpha_k. \end{aligned}$$

For the choice of step size $\eta_k = 2/(\mu k)$, it holds that

$$\frac{1 - \mu\eta_k}{\eta_k^2} = \frac{\mu^2 k^2(1 - 2/k)}{4} = \frac{\mu^2(k^2 - 2k)}{4} = \frac{\mu^2((k-1)^2 - 1)}{4} \leq \frac{\mu^2(k-1)^2}{4} = \eta_{k-1}^{-2}.$$

Setting $e_j := \eta_{j-1}^{-2}\mathbb{E}\left\|x^\star - x_j\right\|^2$ for all $j \in \mathbb{N}_0$, we have

$$2\eta_k^{-1}\mathbb{E}(f(x_k) - f(x^\star)) \leq e_k - e_{k+1} + M_k^2 + 2\eta_k^{-1}\gamma_k\alpha_k.$$

Summing both sides over $k = k_0, \dots, K-1$ for $K \in \mathbb{N}$ yields

$$\begin{aligned} 2\sum_{k=k_0}^{K-1} \eta_k^{-1}\mathbb{E}(f(x_k) - f(x^\star)) &\le e_{k_0} - e_K + \sum_{k=k_0}^{K-1} M_k^2 + 2\sum_{k=k_0}^{K-1} \eta_k^{-1}\gamma_k\alpha_k \\ &\le e_{k_0} + \sum_{k=k_0}^{K-1} M_k^2 + 2\sum_{k=k_0}^{K-1} \eta_k^{-1}\gamma_k\alpha_k, \end{aligned}$$

where we used $e_K \ge 0$ in the second step. We define $S_{t,k} := \sum_{i=t}^{k-1} \eta_i^{-1}$, $S_k := S_{0,k}$, and $\bar{x}_{t,k} := S_{t,k}^{-1} \sum_{i=t}^{k-1} \eta_i^{-1} x_i$ for $t, k \in \mathbb{N}$. Using convexity of f, we get

$$\begin{aligned} \mathbb{E}(f(\bar{x}_{k_0,K}) - f(x^\star)) &\le S_K^{-1} \sum_{k=k_0}^{K-1} \eta_k^{-1}\mathbb{E}(f(x_k) - f(x^\star)) \\ &\le \frac{e_{k_0}}{2S_{k_0,K}} + \frac{\sum_{k=k_0}^{K-1} M_k^2}{2S_{k_0,K}} + \frac{\sum_{k=k_0}^{K-1} \eta_k^{-1}\gamma_k\alpha_k}{S_{k_0,K}}, \end{aligned}$$

for all $K \in \mathbb{N}$, as desired. Note that

$$\bar{x}_K = \frac{S_{0,k_0}}{S_K}\bar{x}_{0,k_0} + \frac{S_{k_0,K}}{S_K}\bar{x}_{k_0,K}$$

and

$$\frac{S_{0,k_0}}{S_K} + \frac{S_{k_0,K}}{S_K} = 1,$$

hence, using convexity of f again, we have

$$\mathbb{E}(f(\bar{x}_K) - f(x^\star)) \le \frac{S_{0,k_0}}{S_K}\mathbb{E}(f(\bar{x}_{0,k_0}) - f(x^\star)) + \frac{S_{k_0,K}}{S_K}\mathbb{E}(f(\bar{x}_{k_0,K}) - f(x^\star)).$$

Combining with the latest bound on $\mathbb{E}(f(\bar{x}_{k_0,K}) - f(x^\star))$ yields

$$\mathbb{E}(f(\bar{x}_K) - f(x^\star)) \le \frac{S_{0,k_0}}{S_K}\mathbb{E}(f(\bar{x}_{0,k_0}) - f(x^\star)) + \frac{e_{k_0}}{2S_K} + \frac{\sum_{k=k_0}^{K-1} M_k^2}{2S_K} + \frac{\sum_{k=k_0}^{K-1} \eta_k^{-1}\gamma_k\alpha_k}{S_K}.$$

Next, we will derive the desired bound for $\operatorname{dist}(x_k, \mathcal{X})$. Again fix $k \ge K_0$. For the choice $x := \Pi_{\mathcal{X}}(x_k)$, inequality (3.7) gives us

$$\begin{aligned} 2\eta_k\mathbb{E}(f(x_k) - f(\Pi_{\mathcal{X}}(x_k))) \le (1 - \mu\eta_k)&\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})^2) - \mathbb{E}\,\|\Pi_{\mathcal{X}}(x_k) - x_{k+1}\|^2 \\ &+ \eta_k^2 M_k^2 + 2\eta_k\gamma_k\alpha_k - 2\tau\eta_k\gamma_k\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})). \end{aligned}$$

Note that $\mathbb{E}\,\|\Pi_{\mathcal{X}}(x_k) - x_{k+1}\|^2 \geq \mathbb{E}(\operatorname{dist}(x_{k+1}, \mathcal{X})^2)$, hence rearranging the above yields

$$\begin{aligned}\mathbb{E}(\operatorname{dist}(x_{k+1}, \mathcal{X})^2) \leq (1 - \mu\eta_k)\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})^2) + \eta_k^2 M_k^2 + 2\eta_k\gamma_k\alpha_k \\ - 2\tau\eta_k\gamma_k\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})) + 2\eta_k\mathbb{E}(f(\Pi_{\mathcal{X}}(x_k)) - f(x_k)).\end{aligned}$$

Using convexity of f, Cauchy-Schwarz, almost sure boundedness of $(x_k)_{k\in\mathbb{N}}$ (theorem 3.8), and the fact that subgradients of f are bounded on compacta (lemma 3.1), there exists a constant $M \in (0, \infty)$, such that

$$f(\Pi_{\mathcal{X}}(x_k)) - f(x_k) \leq \langle \tilde{\nabla} f(\Pi_{\mathcal{X}}(x_k)), \Pi_{\mathcal{X}}(x_k) - x_k\rangle \leq M \operatorname{dist}(x_k, \mathcal{X}).$$

Combining with the previous inequality and gathering terms involving $\operatorname{dist}(x_k, \mathcal{X})$, we arrive at

$$\begin{aligned}\mathbb{E}(\operatorname{dist}(x_{k+1}, \mathcal{X})^2) \leq (1 - \mu\eta_k)\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})^2) + \eta_k^2 M_k^2 + 2\eta_k\gamma_k\alpha_k \\ - 2\eta_k(\tau\gamma_k - M)\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})).\end{aligned}$$

Since $\gamma_k \uparrow \infty$, there exists $k_1 \in \mathbb{N}$ such that $\tau_k\gamma_k - M \geq \tau\gamma_k/2$ for all natural numbers $k \geq k_1$. Thus,

$$\begin{aligned}\mathbb{E}(\operatorname{dist}(x_{k+1}, \mathcal{X})^2) \leq (1 - \mu\eta_k)\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})^2) + \eta_k^2 M_k^2 + 2\eta_k\gamma_k\alpha_k \\ - \tau\eta_k\gamma_k\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})),\end{aligned}$$

for all $k \geq k_1$. Multiplying both sides by $\gamma_k^{-1}\eta_k^{-2}$ yields

$$\begin{aligned}\gamma_k^{-1}\eta_k^{-2}\mathbb{E}(\operatorname{dist}(x_{k+1}, \mathcal{X})^2) \leq \gamma_k^{-1}\eta_k^{-2}(1 - \mu\eta_k)\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})^2) + \gamma_k^{-1}M_k^2 \\ + 2\eta_k^{-1}\alpha_k - \tau\eta_k^{-1}\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})),\end{aligned}$$

for all $k \geq k_1$. We have already shown that $\eta_k^{-2}(1 - \mu\eta_k) \leq \eta_{k-1}^2$ for all $k \in \mathbb{N}$. Also, since γ_k is nondecreasing, we have $\gamma_k^{-1} \geq \gamma_{k+1}^{-1}$ for all $k \in \mathbb{N}$. Combining these two facts, we get

$$\begin{aligned}\gamma_{k+1}^{-1}\eta_k^{-2}\mathbb{E}(\operatorname{dist}(x_{k+1}, \mathcal{X})^2) \leq \gamma_k^{-1}\eta_{k-1}^{-2}\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})^2) + \gamma_k^{-1}M_k^2 \\ + 2\eta_k^{-1}\alpha_k - \tau\eta_k^{-1}\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})),\end{aligned}$$

for all $k \geq k_1$. Setting $d_j := \gamma_j^{-1}\eta_{j-1}^{-2}\mathbb{E}(\operatorname{dist}(x_j, \mathcal{X})^2)\ \forall j \in \mathbb{N}$ and rearranging again, we obtain

$$\tau\eta_k^{-1}\mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})) \leq d_k - d_{k+1} + \gamma_k^{-1}M_k^2 + 2\eta_k^{-1}\alpha_k,$$

for all $k \geq k_1$. Summing over $k = k_1, \dots, K-1$ for $K \in \mathbb{N}$, we have

$$\begin{aligned}
\tau \sum_{k=k_1}^{K-1} \eta_k^{-1} \mathbb{E}(\operatorname{dist}(x_k, \mathcal{X})) &\leq \sum_{k=k_1}^{K-1} (d_k - d_{k+1}) + \sum_{k=k_1}^{K-1} \gamma_k^{-1} M_k^2 + 2 \sum_{k=k_1}^{K-1} \eta_k^{-1} \alpha_k \\
&= d_0 - d_K + \sum_{k=k_1}^{K-1} \gamma_k^{-1} M_k^2 + 2 \sum_{k=k_1}^{K-1} \eta_k^{-1} \alpha_k \\
&\leq d_0 + \sum_{k=k_1}^{K-1} \gamma_k^{-1} M_k^2 + 2 \sum_{k=k_1}^{K-1} \eta_k^{-1} \alpha_k,
\end{aligned}$$

where we used $d_K \geq 0$ in the last step. The distance functional $x \mapsto \operatorname{dist}(x, \mathcal{X})$ is convex (TODO: Show this in an example.), so if we divide both sides of the above inequality by $\tau \cdot S_{k_1,K}$, we obtain

$$\mathbb{E}(\operatorname{dist}(\bar{x}_{k_1,K})) \leq \frac{\tau^{-1} d_{k_1}}{S_{k_1,K}} + \frac{\tau^{-1} \sum_{k=k_1}^{K-1} \gamma_k^{-1} M_k^2}{S_{k_1,K}} + \frac{2\tau^{-1} \sum_{k=k_1}^{K-1} \eta_k^{-1} \alpha_k}{S_{k_1,K}}.$$

To derive a bound for $\mathbb{E}(\operatorname{dist}(\bar{x}_K))$, we proceed similarly as before. We have

$$\bar{x}_K = \frac{S_{0,k_1}}{S_K} \bar{x}_{0,k_1} + \frac{S_{k_1,K}}{S_K} \bar{x}_{k_1,K}$$

and

$$\frac{S_{0,k_1}}{S_K} + \frac{S_{k_1,K}}{S_K} = 1.$$

Combining with the latest bound on $\mathbb{E}(\operatorname{dist}(\bar{x}_{k_1,K}))$, we obtain the desired bound

$$\begin{aligned}
\mathbb{E}(\operatorname{dist}(\bar{x}_K)) &\leq \frac{S_{0,k_1}}{S_K} \mathbb{E}(\operatorname{dist}(\bar{x}_{0,k_1})) + \frac{S_{k_1,K}}{S_K} \mathbb{E}(\operatorname{dist}(\bar{x}_{k_1,K})) \\
&\leq \frac{S_{0,k_1}}{S_K} \mathbb{E}(\operatorname{dist}(\bar{x}_{0,k_1})) + \frac{\tau^{-1} d_{k_1}}{S_K} + \frac{\tau^{-1} \sum_{k=k_1}^{K-1} \gamma_k^{-1} M_k^2}{S_K} + \frac{2\tau^{-1} \sum_{k=k_1}^{K-1} \eta_k^{-1} \alpha_k}{S_K}
\end{aligned}$$

for all $K \geq k_1$. This concludes the proof. □

Theorem 3.12 (Convergence rates in expectation)**.** Let assumptions 2 and 3 hold and define $\eta_k := 2/(\mu k)$ for all $k \in \mathbb{N}$, $\eta_0 := 2/\mu$. Then, for any $e \in (0, \infty)$, there exist sequences $(\gamma_k)_{k \in \mathbb{N}_0}$, $(h_k)_{k \in \mathbb{N}_0}$, $(\beta_k)_{k \in \mathbb{N}}$ such that the iterates $(\bar{x}_k)_{k \in \mathbb{N}}$ generated from algorithm 2 with parameters $(\eta_k)_{k \in \mathbb{N}_0}$, $(\gamma_k)_{k \in \mathbb{N}_0}$, $(h_k)_{k \in \mathbb{N}_0}$, $(\beta_k)_{k \in \mathbb{N}}$ satisfy

$$\mathbb{E}(|f(\bar{x}_K) - f(x^\star)|) = \mathcal{O}(\log^{2e}(K)/K).$$

Further, it also holds that

$$\operatorname{dist}(\bar{x}_K, \mathcal{X}) = \mathcal{O}(\log^e(K)/K).$$

Proof. Lemma 3.11 gives us an upper bound for $\mathbb{E}(f(\bar{x}_k) - f(x^\star))$. By convexity, we have for all $k \in \mathbb{N}$

$$\begin{aligned} f(\bar{x}_k) - f(x^\star) &\geq \langle \tilde{\nabla} f(x^\star), \bar{x}_k - x^\star \rangle \\ &= \langle \tilde{\nabla} f(x^\star), \bar{x}_k - \Pi_{\mathcal{X}}(\bar{x}_k) \rangle + \langle \tilde{\nabla} f(x^\star), \Pi_{\mathcal{X}}(\bar{x}_k) - x^\star \rangle, \end{aligned}$$

for any subgradient $\tilde{\nabla} f(x^\star) \in \partial f(x^\star)$. Optimimality of $x^\star$ for f on $\mathcal{X}$ implies that there exists a subgradient $\tilde{\nabla} f(x^\star) \in \partial f(x^\star)$ such that $\langle \tilde{\nabla} f(x^\star), \Pi_{\mathcal{X}}(\bar{x}_k) - x^\star \rangle \geq 0$. Furthermore, lemma 3.1 implies that we can bound the subgradients in $\partial f(x^\star)$ by a constant $M \in (0, \infty)$. Combining these facts with the above inequality, and applying Cauchy-Schwarz, we obtain

$$f(\bar{x}_k) - f(x^\star) \geq -M \operatorname{dist}(\bar{x}_k),$$

for all $k \in \mathbb{N}$. From lemma 3.11 we know that there exists $k_1 \in \mathbb{N}$ such that for all $K \in \mathbb{N}$ with $K \geq k_1$, it holds that

$$\mathbb{E}(\operatorname{dist}(\bar{x}_K, \mathcal{X})) \leq \frac{S_{0,k_1}}{S_K} \mathbb{E}(\operatorname{dist}(\bar{x}_{0,k_1})) + \frac{\tau^{-1} d_{k_1}}{S_K} + \frac{\tau^{-1} \sum_{k=k_1}^{K-1} \gamma_k^{-1} M_k^2}{S_K} + \frac{2\tau^{-1} \sum_{k=k_1}^{K-1} \eta_k^{-1} \alpha_k}{S_K},$$

where $S_K := \sum_{k=0}^{K-1} \eta_k^{-1}$ and $M_k^2 = \mathcal{O}(\beta_k^{-1}(1+\gamma_k^2) + \gamma_k^2)$. Since $\eta_k^{-1} = \mathcal{O}(k)$, it follows that $S_K = \mathcal{O}(K^2)$. Also, since $\alpha_k = \mathcal{O}(k)$, we have $\eta_k^{-1} \alpha_k \leq c$ for some $c \in (0, \infty)$. Further, $\gamma_k^{-1} M_k^2 = \mathcal{O}(\gamma_k)$. Therefore

$$\begin{aligned} \mathbb{E}(\operatorname{dist}(\bar{x}_K, \mathcal{X})) &\leq \mathcal{O}(K^{-2}) + \mathcal{O}\Big(K^{-2} \sum_{k=1}^{K-1} \log^e(k)\Big) + \mathcal{O}(K^{-1}) \\ &= \mathcal{O}(K^{-2}) + \mathcal{O}(\log^e(K)/K) + \mathcal{O}(K^{-1}) \\ &= \mathcal{O}(\log^e(K)/K). \end{aligned}$$

Therefore,

$$f(\bar{x}_k) - f(x^\star) \geq -M \cdot \mathcal{O}(\log^e(K)/K).$$

The upper bound from lemma 3.11 reads

$$\mathbb{E}(f(\bar{x}_K) - f(x^\star)) \leq \frac{S_{0,k_0}}{S_K} \mathbb{E}(f(\bar{x}_{0,k_0}) - f(x^\star)) + \frac{e_{k_0}}{2S_K} + \frac{\sum_{k=k_0}^{K-1} M_k^2}{2S_K} + \frac{\sum_{k=k_0}^{K-1} \eta_k^{-1} \gamma_k \alpha_k}{S_K},$$

and since $M_k^2 = O(\log^{2e} k)$ and $\eta_k^{-1}\gamma_k\alpha_k = O(\log^e k)$, it follows that

$$\begin{aligned}\mathbb{E}(f(\bar{x}_K) - f(x^\star)) &\leq O(K^{-2}) + O\Big(K^{-2}\sum_{k=1}^{K-1}\log^{2e}(k)\Big) + O\Big(K^{-2}\sum_{k=1}^{K-1}\log^{e}(k)\Big) \\ &= O(\log^{2e}(K)/K).\end{aligned}$$

Putting the two bounds together, we obtain

$$\mathbb{E}(|f(\bar{x}_K) - f(x^\star)|) = O(\log^{2e}(K)/K),$$

as desired. □

3.3. High-probability guarantees

In the previous sections we proved $O(\log k/k)$ convergence to zero of the expected suboptimality, $\mathbb{E}|f(\bar{x}_k) - f(x^\star)|$, and the expected subfeasibility, $\operatorname{dist}(\bar{x}_k, \mathcal{X})$, of the sequence $(\bar{x}_k)_{k\in\mathbb{N}}$ generated by algorithm 2, under suitable choice of parameters $(\eta_k)_{k\in\mathbb{N}_0}$, $(\gamma_k)_{k\in\mathbb{N}_0}$, $(\beta_k)_{k\in\mathbb{N}_0}$, and $(h_k)_{k\in\mathbb{N}_0}$. We will now turn our attention to establishing *high-probability guarantees*, in contrast to guarantees that hold only in expectation. In particular, we will investigate under which conditions we can guarantee that an iterate $\bar{x}_k$ is both a) close to the feasible set and b) has function value close to the optimal value $f(x^\star)$ with high probability. This notion of "closeness with high probability" is formalized in the following definition.

Definition 3.13 ((ϵ,δ)-solution)**.** Let $\epsilon \in (0,\infty)$ and $\delta \in (0,1)$. We call a random variable $x\colon \Omega \to \mathbb{R}^n$ an (ϵ,δ)**-solution** of (P), if

$$\mathbb{P}\Big(\max\big(|f(x) - f(x^\star)|, \operatorname{dist}(x,\mathcal{X})\big) \geq \epsilon\Big) \leq \delta.$$

In other words, an iterate x_k of algorithm 2 is an (ϵ,δ)-solution, if we can guarantee that $|f(x_k) - f(x^\star)| < \epsilon$ and $\operatorname{dist}(x_k,\mathcal{X}) < \epsilon$ with probability greater than $1-\delta$. We can now state the central questions we aim to answer in this section as follows:

1. For what choice of parameters can we guarantee that the sequence $(\bar{x}_k)_{k\in\mathbb{N}}$ generated by algorithm 2 reaches an (ϵ,δ)-solution of (P) as quickly as possible, for any given choice of ϵ and δ?

2. What is the relatioship between ϵ, δ, and the number of iterations of algorithm 2 needed to reach an (ϵ,δ)-solution of (P)?

To resolve these questions, we will first need a more general version of lemma 3.5, which is established in the next lemma. We will then proceed similarly to how we did in the previous section, though we will need to carry around an extra term throughout the calculations.

Lemma 3.14 (One-step improvement II)**.** Let $\rho \in (0, 1)$ and $\eta_k \in (0, \rho L^{-1}]$ for all $k \in \mathbb{N}_0$. Then the iterates $(x_k)_{k\in\mathbb{N}}$ generated by algorithm 2 with step size schedule $(\eta_k)_{k\in\mathbb{N}_0}$ satisfy

$$\begin{aligned} 2\eta_k(f_k(x_{k+1}) - f_k(x)) \leq (1 - \mu\eta_k)\|x - x_k\|^2 + 2\eta_k\langle z_k, x_k - x\rangle - \|x - x_{k+1}\|^2 \\ + \frac{8G^2\eta_k^2\gamma_k^2}{1-\rho} + \frac{2\eta_k^2}{1-\rho}\|z_k\|^2 \end{aligned}$$

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

Proof. This is just inequality (3.4). Since we make the same assumptions in this lemma as in lemma 3.5, we can simply follow the exact same steps until we arrive at the desired result. □

Lemma 3.15. Let assumption 4 hold. For all $k \in \mathbb{N}$, it holds that

$$\|z_k\|^2 \leq 8\hat{L}^2 + 8\gamma_k^2 N^2 G^2,$$

with probability one, where $\hat{L}, G, N \in (0, \infty)$.

Proof. 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 have

$$\begin{aligned} \|z_k\|^2 &= \left\|g_k - \nabla\psi_k(x_k)\right\|^2 \\ &= \left\|\nabla F_{\xi_k}(x_k) + \gamma_k A(\xi_k)^\top \nabla h_k(A(\xi_k)x_k - b(\xi_k)) + \nabla\mathbb{E}(F_{\xi_k}(x_k)) + \gamma_k\nabla\pi_k(x_k)\right\|^2 \\ &\leq 4\Big(\left\|\nabla F_{\xi_k}(x_k)\right\|^2 + \gamma_k^2\|A(\xi_k)\|_F^2\|\nabla h_k(A(\xi_k)x_k - b(\xi_k))\|^2 \\ &\qquad + \left\|\nabla\mathbb{E}(F_{\xi_k}(x_k))\right\|^2 + \gamma_k^2\|\nabla\pi_k(x_k)\|^2\Big) \end{aligned}$$

By almost sure L-smoothness of $x \mapsto F_\xi(x)$ and $x \mapsto \mathbb{E}(F_\xi(x))$ (assumption 1), as well as the fact that $(x_k)_{k\in\mathbb{N}}$ is almost surely bounded (theorem 3.8), there exists a constant $\hat{L} \in (0, \infty)$ such that

$$\left\|\nabla F_{\xi_k}(x_k)\right\| \leq \hat{L} \text{ and } \left\|\nabla\mathbb{E}(F_{\xi_k}(x_k))\right\| \leq \hat{L}.$$

Therefore,

$$\left\|\nabla F_{\xi_k}(x_k) + \gamma_k A(\xi_k)^\top \nabla h_k(A(\xi_k)x_k - b(\xi_k))\right\| \leq \hat{L} + \gamma_k N G.$$

Further $\|A(\xi)\|_F \leq N$ a. s. (assumption 1), and

$$\sup_{k\in\mathbb{N}} \sup_{x\in\mathbb{R}^m} \|\nabla h_k(x)\| \leq G \quad \text{and} \quad \sup_{k\in\mathbb{N}} \sup_{x\in\mathbb{R}^n} \|\nabla \pi_k(x)\| \leq \mathbb{E}\|A(\xi)\|_F\, G$$

for a constant $G \in (0, \infty)$ (lemma 3.1). We thus obtain

$$\|z_k\|^2 \leq 4(2\hat{L}^2 + 2\gamma_k^2 N^2 G^2) = 8\hat{L}^2 + 8\gamma_k^2 N^2 G^2,$$

as desired. □

3.4. Infeasible problems

As we have seen in the SVM example (TODO), some problems of interest may not be feasible. Yet, our methods can still be applied in those cases. The question is then: What do the iterates converge to, if anything?

Definition 3.16. Let $\delta \in [0, 1]$. A point $x \in \mathbb{R}^n$ is called δ**-feasible**, if

$$\mathbb{P}(A(\xi)x - b(\xi) > 0) \leq \delta.$$

The δ**-set**, denoted $\mathcal{X}_\delta$, is the set of all δ-feasible points. A point $x \in \mathbb{R}^n$ is called **maximally feasible**, if there exists $\delta \in [0, 1]$ such that (x, δ) solves

$$\min_{(x,\delta)\in\mathbb{R}^n\times[0,1]} \delta \quad \text{s. t.} \quad x \in \mathcal{X}_\delta.$$

Conjecture: Consider the following two statements:

1. For any $\delta \in (0, 1]$, the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ is eventually contained in $\mathcal{X}_\delta$ in probability.

2. The sequence of iterates $(x_k)_{k\in\mathbb{N}}$ converges to a maximally feasible point in probability.

At least one of these two statements must hold, regardless of whether (P) is feasible or not. Both statements hold iff. problem (P) is feasible.

4

Numerical Examples

5
Summary and Outlook

Bibliography

[1] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[2] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[3] Angelia Nedić and Tatiana Tatarenko. Huber loss-based penalty approach to problems with linear constraints. *arXiv preprint arXiv:2311.00874*, 2023.

[4] Meng Li, Paul Grigas, and Alper Atamtürk. New penalized stochastic gradient methods for linearly constrained strongly convex optimization. *Journal of Optimization Theory and Applications*, 205(2):29, 2025.

[5] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

[6] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *Journal of machine learning research*, 24(147):1–56, 2023.

[7] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

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

[9] Hui Hu and Qing Wang. On approximate solutions of infinite systems of linear inequalities. *Linear Algebra and its applications*, 114:429–438, 1989.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Mathematics selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel — insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen — benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe.

Hamburg, am . . . .

Amir Miri Lavasani