A proximal subgradient method for nonconvex stochastic optimization under the Kurdyka–Łojasiewicz condition

Felipe Atenas · Alejandro Jofré · Pedro Pérez-Aros · David Torregrosa-Belén

August 5, 2026

Abstract This work introduces a proximal stochastic subgradient method for minimizing the sum of an expected cost, whose integrand is potentially nonsmooth and nonconvex, and a lower semicontinuous, prox-bounded function. We target a broad class of integrands obeying a nonsmooth, localized variant of the descent lemma in the decision variable, a structural assumption that simultaneously covers smooth losses with Lipschitz gradient and differences of such losses with convex functions. At each iteration the expected cost is replaced by a sample average that is progressively refined, and the proximal-subgradient stepsize is selected by an Armijo-type line search enforcing a sufficient-decrease property up to stochastic errors induced by the sample-based approximation. This framework accommodates substantially more general problem formulations than existing methods, in particular, it requires neither (weak) convexity of the regularizer nor a uniform bound on the variance of the stochastic oracle, and our analysis yields convergence guarantees that are new even in the smooth setting. Specifically, we establish almost sure convergence of the sequence of function values and stationarity of every accumulation point of the trajectories under the relaxed requirement that the sample-size sequence be merely nondecreasing and unbounded, with no prescribed growth rate. Leveraging the Kurdyka–Łojasiewicz (KL) property, we further upgrade this subsequential guarantee to convergence of the whole trajectory to a single stationary point. Finally, for exponential-type KL desingularizing functions and polynomially growing sample sizes, we derive explicit polynomial convergence rates, up to a logarithmic factor, for both the function values and the iterates.

Keywords Stochastic programming · proximal subgradient method · stochastic approximation · nonconvex optimization · Kurdyka–Łojasiewicz condition · convergence rates

Mathematics Subject Classification (2020) 90C15, 49J53, 49J52, 90C26, 65K05

1 Introduction

A large class of optimization problems under uncertainty are modeled as minimizing an expected value written as an integral with respect to the underlying probability distribution. In many applications arising in machine learning, statistics, and operations research, this expectation cannot be evaluated exactly, either because the distribution is not available in closed form or because the resulting integral

This paper has not been published, already submitted or will be submitted anywhere else.
Research of the first three authors was supported by Centro de Modelamiento Matemático (CMM), ACE210010 and FB210005, BASAL funds for center of excellence from ANID-Chile. The third author was supported by ANID-Chile grant: Fondecyt Regular 1240335, Fondecyt Regular 1240120, Fondecyt Regular 1261728. The fourth author was partially supported by grant PID2022-136399NB-C21 funded by ERDF/EU and by MICIU/AEI/10.13039/501100011033.

Felipe Atenas
Centro de Modelamiento Matemático (CNRS IRL2807), Universidad de Chile, Santiago, Chile
E-mail: fatenas@cmm.uchile.cl

Alejandro Jofré and Pedro Pérez-Aros
Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático (CNRS IRL2807), Universidad de Chile, Santiago, Chile
E-mail: ajofre@uchile.cl, pperez@dim.uchile.cl

David Torregrosa-Belén
Department of Mathematics, University of Alicante, Alicante, Spain
E-mail: david.torregrosa@ua.es

is computationally intractable in high dimensions. To overcome this issue, stochastic algorithms replace exact first-order information by sample-based approximations, leading to methods that must balance computational effort, statistical accuracy, and convergence guarantees; see, for instance, [23, 43–45, 74].

In this paper, we consider the following structured nonconvex stochastic programming problem

$$\min_{x\in\mathbb{R}^d} \Phi(x) := \mathbb{E}_\mu\big[\varphi(\xi,x)\big] + \psi(x), \tag{$\mathcal{P}$}$$

where the expected cost is defined on a probability space $(\Xi, \mathcal{A}, \mu)$ by

$$\mathbb{E}_\mu\big[\varphi(\xi,x)\big] := \int_\Xi \varphi(\xi,x)\, d\mu(\xi).$$

Here, $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ is assumed to be an upper-C^2 normal integrand with respect to the decision variable in the sense of Definition 2.5. Roughly speaking, for all $\xi \in \Xi$, the mapping $x \mapsto \varphi(\xi, x)$ may be nonsmooth and nonconvex, satisfying a local generalized version of the so-called *descent lemma*, suitable for variational analysis. The term $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is assumed to be proper, lower semicontinuous, and prox-bounded. This allows us to encode constraints, nonsmooth regularization terms, and convex/nonconvex penalties. Models of the form ($\mathcal{P}$) emerge naturally in stochastic problems where nonsmooth losses and regularizers appear, such as optimal quantization, sparse and partial-label learning, matrix factorization and statistical estimation, among others; see, e.g., [25, 30, 39, 69, 72, 73, 92].

To tackle problem ($\mathcal{P}$), we propose a method of proximal subgradient type, in which stochastic first-order information is estimated by average approximations. A sufficient decrease condition of the sample-based estimations tracks the progress of the method, though it does not yield a descent method in the usual sense, but rather a method of approximate descent up to vanishing stochastic errors. The method is presented in Algorithm 1.

Related work Stochastic optimization methods have been extensively studied in recent years (see, e.g., [13, 23, 86, 90]), motivated by their relevance in modern applied mathematics, and especially by their prominent role in machine learning and related scientific fields. A substantial part of this literature has focused on convex and strongly convex optimization, where significant advances have been achieved through several complementary approaches, including refined stepsize rules (see, e.g., [74]), variance reduction techniques (see, e.g., [54]), and central limit theorems describing the asymptotic behavior of stochastic algorithms (see, e.g., [65] and references therein).

Despite the tremendous progress made in convex stochastic programming, the nonconvex and nonsmooth setting has received comparatively less attention. This is due to the intrinsic difficulties arising from the interaction of nonconvexity, nonsmoothness, and stochasticity. Indeed, deterministic techniques do not transfer directly to stochastic frameworks, where first-order information is typically available through oracles only. Moreover, in nonconvex optimization, convergence of function values does not necessarily characterize convergence to the optimal value, and stationary points need not be global, or even local, minimizers. Nevertheless, several important contributions have been made in this direction. For smooth and composite nonconvex stochastic programming, complexity guarantees and accelerated variants have been obtained in [43–45]. More recent developments include variance-reduced, recursive-gradient, mini-batch, and hybrid stochastic methods for finite-sum and expected value composite problems; see, e.g., [68, 77, 89]. These methods typically assume that the expected value term is smooth, or that the nonsmooth component has a convex structure that can be handled by a proximal operator. For nonsmooth expected value objectives, proximal stochastic subgradient and stochastic prox-linear methods have been proposed for weakly convex functions; see, e.g., [31–33]. Stochastic difference-of-convex algorithms have also been investigated in [63].

Sample-based stochastic optimization methods A classical approach to tackle the minimization of expected costs is the sample average approximation (SAA) method, which replaces the expected value function by an empirical average over a finite sample. This leads to a deterministic optimization problem solved by deterministic optimization techniques. SAA methods have been extensively studied from the viewpoints of consistency, asymptotic convergence, stability of optimal solutions, and finite-sample error estimates; see, among others, [48, 58, 59, 87]. This approach is especially useful when the sampled problem preserves exploitable deterministic structure. However, solving a sequence of large deterministic problems can become expensive when high accuracy requires large sample sizes or when each sampled problem is itself nonsmooth and nonconvex. A different paradigm is provided by stochastic approximation (SA) methods. This line of research goes back to the seminal works [57, 83], and has become one of the central

algorithmic frameworks for stochastic optimization. Unlike SAA, these methods are able to address the original stochastic problem by calling stochastic oracles for approximating first-order information of the expected cost; see, e.g., [62, 74, 80].

Regarding stepsize selection, we can mention two approaches. In the absence of dynamically increasing sample sizes, stochastic approximation schemes require rapidly diminishing stepsizes to ensure convergence; see, e.g., [42, 45, 74]. More precisely, in these works, the sequence of stepsizes $(\gamma_k)_{k\in\mathbb{N}}$ satisfies $\sum_{k=0}^{\infty}\gamma_k = +\infty$ and $\sum_{k=0}^{\infty}\gamma_k^2 < +\infty$. Such conditions are standard in the convergence analysis of (proximal) stochastic gradient methods, both in convex and nonconvex settings. As an alternative, the stepsizes can be defined to adapt to the progress of the iteration process via backtracking line search procedures using function and (sub)gradient estimates; see, e.g., [75, 76].

Kurdyka–Łojasiewicz condition in nonconvex optimization The analysis of optimization algorithms dealing with nonconvex problems such as ($\mathcal{P}$) is significantly different from the convex case. In the nonconvex scenario, the existence of multiple local minima and saddle points complicates establishing global convergence guarantees for the sequence of iterates. A fundamental tool to overcome this challenge of deterministic nonconvex methods is the Kurdyka–Łojasiewicz (KL) inequality; see, e.g., [1, 10–12, 22]. The KL framework enables improving subsequential convergence results to convergence of the entire sequence of iterates to a single stationary point, as well as characterizing local convergence rates. Recall that the satisfaction of a sufficient decrease condition is pivotal for the convergence principle of the seminal work [12]. This condition might be compromised in the stochastic setting, due to the presence of errors induced by stochastic approximations, preventing the straightforward extrapolation of KL-based techniques for analyzing stochastic methods.

Our contribution The purpose of this work is to analyze a new proximal stochastic subgradient (PSS) algorithm for solving ($\mathcal{P}$). At every iteration, PSS constructs a surrogate model of ($\mathcal{P}$) where the expected cost is replaced by a sample average approximation refined along iterations. Specifically, given an increasing sequence of sample sizes $(n_k)_{k\in\mathbb{N}}$, where $k \in \mathbb{N}$ represents the algorithm's current iteration, PSS draws a collection $\xi_{n_{k-1}+1}, \xi_{n_{k-1}+2}, \ldots, \xi_{n_k}$ of independent and identically distributed (i.i.d.) random variables to construct an *empirical average* approximation of the integral in ($\mathcal{P}$), that is, for all $x \in \mathbb{R}^d$,

$$\mathbb{E}_\mu\big[\varphi(\xi, x)\big] \approx \frac{1}{n_k}\sum_{j=1}^{n_k}\varphi(\xi_j, x).$$

In this manner, PSS generates an iterative sequence of the form

$$\begin{cases} v^k \in \dfrac{1}{n_k}\displaystyle\sum_{j=1}^{n_k}\partial\varphi(\xi_j, x), \\ x^{k+1} \in \operatorname{prox}_{\gamma\psi}(x^k - \gamma v^k), \end{cases}$$

thus combining stochastic (Clarke) subgradient information of φ taken from the empirical approximation with a proximal evaluation of ψ. The stepsize $\gamma > 0$ for this proximal subgradient update is computed by an Armijo-type line search [5] to ensure a sufficient decrease of the sample model of the objective function Φ in ($\mathcal{P}$). For more details, we refer to Algorithm 1. The sample errors lead to a nonmonotone stochastic scheme with respect to the original objective Φ, hindering the direct application of standard techniques for the convergence analysis of descent methods, and particularly the KL framework. Nevertheless, we show that our method can be viewed as a descent scheme perturbed by stochastic errors that vanish along the iteration process, as long as the sequence of iterates remains bounded. In this regard, and in contrast to standard approaches, we do not impose any bounded variance condition, but rather assume local Lipschitzianity of $x \mapsto \varphi(\xi, \cdot)$; see Remark 3.1. This allows us to combine a stopping time based analysis together with new results on uniform error bounds for average approximations on compact sets; see Section 2.4.

From the perspective of the proximal operation, our analysis only relies on the ability to compute at least one element of the proximal mapping of ψ, without requiring an additional geometric structure such as (weak) convexity, different from most existing works on stochastic approximation schemes with proximal steps; see, e.g., [32, 40, 53, 81, 93]. Thus, our setting naturally covers the class of proper, lower semicontinuous, prox-bounded functions for which such a proximal point is available. From a numerical standpoint, this allows us to cover many relevant classes of constraints and nonconvex regularizers; see, e.g., [8, 18, 39, 69, 73, 92].

Our main results can be summarized as follows:

- In Theorem 3.5, we prove that almost surely every accumulation point of bounded sequences generated by PSS is a stationary point of $(\mathcal{P})$. We do not impose any assumption on the sequence of sample sizes $(n_k)_{k\in\mathbb{N}}$ besides being nondecreasing and unbounded. To the best of our knowledge, this feature remains novel even within the smooth setting; see Remark 3.10 for details.
- In Theorem 4.2, we upgrade the previous subsequential guarantee to global convergence of the whole sequence of iterates generated by PSS to a single stationary point of $(\mathcal{P})$. This trajectory-level result is obtained for non-differentiable objectives by leveraging the KL condition (see (32)), provided the approximation error is controlled at a summable rate.
- In Theorem 4.9, we establish explicit local convergence rates when the desingularizing function of the KL condition belongs to the exponential family (see Remark 4.1) and the sample sizes grow at least polynomially. We show that both the function values and the iterates converge at a rate of the order $\mathcal{O}\big(k^{-p}\ln(k)\big)$, where the exponent $p > 0$ is given in closed form as a function of the KL exponent β and of the polynomial decay rate γ of the inverse sample-size sequence $(n_k^{-1})_{k\in\mathbb{N}}$; see Figure 1 for a visualization of the attainable exponents. The underlying recursion estimate (Lemma 2.12), which tracks a contraction perturbed by polynomially decaying errors, is of independent interest and, to our knowledge, has not previously appeared in this form.

In particular, our results show that deterministic KL-type ideas can still be used in a stochastic nonmonotone setting, provided that the sample errors are controlled with sufficient accuracy alongside iterations.

Organization of the paper In Section 2, we introduce the notation and concepts used throughout this work and show some preliminary, technical results. In particular, we introduce the notion of upper-C^2 normal integrands (Definition 2.5), pivotal for our analysis. Section 3 presents our proposed PSS method (Algorithm 1) to solve problem $(\mathcal{P})$ and its analysis of almost surely subsequential convergence. In Section 4, under the assumptions of the KL condition, we deduce global convergence results and estimates of the rate of convergence of the PSS method. In Section 5 we illustrate the proposed method in numerical experiments on two tasks: optimal quantization and partial-label linear regression. Section 6 concludes with some final remarks and future research directions.

2 Preliminaries of variational and stochastic analysis

2.1 Elements from nonsmooth analysis

For a function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, Lipschitz continuous around $x \in \mathbb{R}^d$, the *Clarke subdifferential* is defined as

$$\partial f(x) = \{v \in \mathbb{R}^d : \langle v, d\rangle \le f^\circ(x; d) \text{ for all } d \in \mathbb{R}^d\},$$

where $f^\circ(x; d)$ stands for the Clarke subderivative at x in the direction $d \in \mathbb{R}^d$, which is defined as

$$f^\circ(x; d) = \limsup_{\substack{y\to x\\ t\to 0^+}} \left(\frac{f(y + td) - f(y)}{t}\right).$$

A function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is said to be *prox-bounded* if there exists some $\gamma > 0$ such that $x \mapsto f(x) + \frac{1}{2\gamma}\|x\|^2$ is bounded from below. The supremum of the set of all such $\gamma > 0$ is called the *threshold of prox-boundedness* for f, and is denoted by γ^f. Given a proper, lower semicontinuous (lsc) function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ and a constant $\gamma \in (0, +\infty]$, the *proximal mapping* is the multifunction $\mathrm{prox}_{\gamma f} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ defined as the solution set of the optimization problem

$$\mathrm{prox}_{\gamma f}(x) := \operatorname*{argmin}_{u\in\mathbb{R}^d}\left\{f(u) + \frac{1}{2\gamma}\|u - x\|^2\right\},$$

with the convention $\frac{1}{2\gamma} = 0$ for $\gamma = +\infty$. If f is prox-bounded, then the set $\mathrm{prox}_{\gamma f}(x)$ is nonempty for every $\gamma \in (0, \gamma^f)$. We say that $\bar{x}$ is a stationary point of $(\mathcal{P})$ if there exists $\bar{v} \in \partial_x \mathbb{E}_\mu\big[\varphi(\xi, \bar{x})\big]$ and $\bar{\gamma} > 0$ such that

$$\bar{x} \in \mathrm{prox}_{\bar{\gamma}\psi}(\bar{x} - \bar{\gamma}\bar{v}). \tag{1}$$

The following lemma shows a type of upper-semicontinuity property of the proximal mapping for prox-bounded functions. Its proof follows the same arguments in [82, Lemma 2.3], and it is included for completeness.

Lemma 2.1 *Let $\psi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ be a proper, lsc and prox-bounded function with threshold γ^ψ. Consider a point $\bar{x} \in \operatorname{dom}\psi$, a bounded sequence $(v^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$, and, for all $k \in \mathbb{N}$, $\widehat{x}^k \in \operatorname{prox}_{\gamma_k\psi}\left(x^k - \gamma_k v^k\right)$ for some sequence $x^k \to \bar{x}$ and $\gamma_k \to 0^+$ with $\gamma_k \in (0, \gamma^\psi)$. Then $\widehat{x}^k \to \bar{x}$.*

Proof. Let $\gamma \in (0, \gamma^\psi)$. Since ψ is prox-bounded and $(x^k)_{k\in\mathbb{N}}$ is bounded, using Young's inequality we have that there exists $\alpha \in \mathbb{R}$ such that

$$\psi(x) \geq -\frac{1}{2\gamma}\|x - x^k\|^2 + \alpha, \quad \text{for all } x \in \mathbb{R}^d.$$

Now, using the definition of the proximal mapping we get that

$$\psi(\widehat{x}^k)+\frac{1}{2\gamma_k}\|\widehat{x}^k - x^k\|^2 + \langle v^k, \widehat{x}^k - x^k\rangle + \frac{\gamma_k}{2}\|v^k\|^2 \leq \psi(\bar{x}) + \frac{1}{2\gamma_k}\|\bar{x} - x^k\|^2 + \langle v^k, \bar{x} - x^k\rangle + \frac{\gamma_k}{2}\|v^k\|^2.$$

Observe that $\langle v^k, \widehat{x}^k - x^k\rangle \geq -\frac{1}{2}\|\widehat{x}^k - x^k\|^2 - \frac{1}{2}\|v^k\|^2$. Hence, by using the two above inequalities one gets

$$\left(\frac{1}{2\gamma_k}-\frac{1}{2\gamma} - \frac{1}{2}\right)\|\widehat{x}^k - x^k\|^2 \leq \psi(\bar{x}) + \frac{1}{2\gamma_k}\|\bar{x} - x^k\|^2 + \langle v^k, \bar{x} - x^k\rangle - \alpha + \frac{1}{2}\|v^k\|^2.$$

It suffices to multiply the above by γ_k and pass to the limit as $k \to \infty$ to conclude that $\|\widehat{x}^k - x^k\| \to 0$, and consequently $\widehat{x}^k \to \bar{x}$. □

2.2 Elements from measure theory

In what follows, $(\Xi, \mathcal{A}, \mu)$ will be a measure space, and we use the notation L^p_μ for denoting the space of all p-integrable functions. Moreover, we consider a (sample) probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Consider a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ and an integrable function $f : \Omega \to \mathbb{R}$. We denote by $\mathbb{E}[f \mid \mathcal{G}]$ the conditional expectation of f with respect to $\mathcal{G}$. Given a measurable set $F \in \mathcal{F}$, the indicator of the set F is given by:

$$\mathbb{1}_F(\omega) = \begin{cases} 1, & \text{if } \omega \in F, \\ 0, & \text{if } \omega \notin F. \end{cases}$$

A set-valued mapping $M\colon \Xi \rightrightarrows \mathbb{R}^d$ is said to be measurable if $M^{-1}(U) \in \mathcal{A}$ for every open set $U \subset \mathbb{R}^d$, where $M^{-1}(U) := \{t \in \Xi \mid M(t) \cap U \neq \emptyset\}$. Additionally, a function $\varphi\colon \Xi \times \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is called a *normal integrand* if the multifunction $\xi \in \Xi \mapsto \operatorname{epi}\varphi_\xi$ is measurable with closed values, where the function $\varphi_\xi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is given by

$$\varphi_\xi(x) := \varphi(\xi, x), \qquad \text{for all } x \in \mathbb{R}^d. \tag{2}$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration, namely, a family of sub-σ-algebras such that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$. A random variable $\tau : \Omega \to \mathbb{N}\cup\{+\infty\}$ is called a *stopping time* with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if for every $n \in \mathbb{N}$, $\{\tau \leq n\} \in \mathcal{F}_n$. Equivalently, τ is a stopping time if and only if $\{\tau = n\} \in \mathcal{F}_n$, for all $n \geq 0$.

Following standard notation in probability theory, we omit the dependence on $\omega \in \Omega$ when writing random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, in order to avoid overloading the notation. Whenever needed, this dependence will be made explicit to prevent ambiguity.

The following result, known as the *freezing lemma*, provides a formula for computing conditional expectations of compositions involving integrands and independent random variables, extending the classical conditional-expectation formula under independence; see, e.g., [17, Lemma 10.10.2]. The proof follows from standard approximation techniques based on simple functions and can be found in [14, Lemma 4.11].

Lemma 2.2 *Let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ be a measurable mapping. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω, and let $\xi : \Omega \to \Xi$ be independent of $\mathcal{G}$ and $y : \Omega \to \mathbb{R}^d$ a $\mathcal{G}$-measurable function. Then, for every $\omega \in \Omega$,*

$$\mathbb{E}\left[\varphi(\xi(\cdot), y(\cdot)) \,|\, \mathcal{G}\right](\omega) = \int_{\Omega} \varphi(\xi(\omega_1), y(\omega))\, d\mathbb{P}(\omega_1), \quad a.s.,$$

provided that $\omega \mapsto \varphi(\xi(\omega), y(\omega))$ is integrable.

Next, let us establish a notion of local Lipschitz continuity for normal integrands.

Definition 2.3 (locally Lipschitz normal integrand) A normal integrand $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}^m$ is said to be locally Lipschitz around $\bar{x}$ with square integrable modulus if $\varphi(\cdot, \bar{x}) \in L^2_\mu$ and there exist $\kappa \in L^2_\mu$ and a neighborhood V of $\bar{x}$ such that

$$\|\varphi_\xi(y) - \varphi_\xi(z)\| \leq \kappa(\xi)\|y - z\|, \text{ for all } y, z \in V,\ \xi \in \Xi.$$

The following proposition shows that the local Lipschitz continuity stated above is equivalent to what appears to be a stronger condition: a uniform Lipschitz condition over bounded sets. More precisely, we have the following result, whose proof follows from a classical compactness argument.

Proposition 2.4 *Let us consider an integrand $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ and let $U \subseteq \mathbb{R}^d$ be a convex compact set. Then the following are equivalent:*

(a) for every $\bar{x} \in U$, the integrand φ is locally Lipschitz around $\bar{x}$ with square integrable modulus.
(b) there exist $\kappa \in L^2_\mu$ and a neighborhood V of U such that

$$\varphi(\cdot, y) \in L^2_\mu \text{ and } |\varphi_\xi(y) - \varphi_\xi(z)| \leq \kappa(\xi)\|y - z\|, \text{ for all } y, z \in V,\ \xi \in \Xi. \tag{3}$$

2.3 Upper-C^2 normal integrands

We now introduce the definition of upper-C^2 integrands anticipated in ($\mathcal{P}$), and that will constitute a fundamental structural assumption for this problem throughout the manuscript.

Definition 2.5 (upper-C^2 normal integrand) Let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a normal integrand. We say that φ is upper-C^2 around $\bar{x}$ provided that φ is locally Lipschitz around $\bar{x}$ with square integrable modulus and there exist a neighborhood V of $\bar{x}$ and (a nonnegative) $\ell \in L^1_\mu$ such that for all $\xi \in \Xi$, $x \in V$ and $v \in \partial\varphi_\xi(x)$, it holds

$$\varphi_\xi(y) \leq \varphi_\xi(x) + \langle v, y - x\rangle + \ell(\xi)\|y - x\|^2, \quad \text{for all } y \in V. \tag{4}$$

Furthermore, we simply say that φ is an upper-C^2 normal integrand provided that φ is upper-C^2 around every point $\bar{x} \in \mathbb{R}^d$.

Remark 2.6 Definition 2.5 extends the notion of upper-C^2 functions for integrands. The classical definition of upper-C^2 functions, namely, the minimum over a family of twice-continuously differentiable functions (see, e.g., [85]) is equivalent, as established in [3, Proposition 2.2] (see also [28, Theorem 5.1]), to (4) when Ξ is a singleton. This equivalence (and further characterizations) can be extended to general sets Ξ, this is the object of the working paper "A note on upper and lower-C^2 normal integrands". Nonetheless, we adopt this alternative definition because it highlights that this class of functions satisfies a nonsmooth version of the descent lemma, a property well-recognized in smooth optimization (see, e.g., [52, Lemma A.11]). Notably, it is well-known that smooth functions with Lipschitz continuous gradients satisfy an inequality of the kind of (4). Furthermore, it has been shown that the class of upper-C^2 functions is broader, encompassing structured functions formed as the difference between a smooth function with a Lipschitz continuous gradient and a (potentially nonsmooth) convex function (see, e.g., [3, 4, 55] for further details).

Similarly to Proposition 2.4, using compactness arguments, we can demonstrate that the integrable function ℓ in (4) can be chosen uniformly on bounded sets.

Proposition 2.7 *Let φ be an upper-C^2 normal integrand. Then for every convex compact set B there exists $\ell \in L^1_\mu$ such that for all $x \in B$ there exists a neighborhood V of x such that (4) holds.*

2.4 Average approximation errors

We now introduce the notation for the expected cost in problem ($\mathcal{P}$) and its sample-based approximations that we will use throughout this paper. For ease of notation, we let $\mathbf{E}_\varphi : \mathbb{R}^d \to \mathbb{R}$ be the expected value function given by

$$\mathbf{E}_\varphi(x) := \mathbb{E}_\mu\left[\varphi(\xi, x)\right] = \int_\Xi \varphi(\xi, x)\, d\mu(\xi), \text{ for all } x \in \mathbb{R}^d.$$

Since the exact evaluation of $\mathbf{E}_\varphi$ may be unavailable or computationally expensive, we approximate it by empirical averages constructed from samples. More precisely, for $k \in \mathbb{N}$ representing an iteration index, the *empirical average* of $\mathbf{E}_\varphi$ at a point $x \in \mathbb{R}^d$ is given by

$$\mathbf{E}^k_\varphi(x) := \frac{1}{n_k}\sum_{j=1}^{n_k} \varphi(\xi_j, x), \text{ for all } x \in \mathbb{R}^d, \tag{5}$$

and where $n_k \in \mathbb{N}$ is the sample size, and for $j = 1, \ldots, n_k$, the random variables ξ_j are i.i.d. In this manner, the sample *average approximation* of Φ in ($\mathcal{P}$) at iteration k is given by

$$\Phi_k(x) := \mathbf{E}^k_\varphi(x) + \psi(x), \text{ for all } x \in \mathbb{R}^d. \tag{6}$$

In the context of the approximation in (5)-(6), we make the following standing assumption on the sample $(\xi_i)_{i\geq 1}$ used throughout the paper.

Assumption 1 (Sampling framework) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\xi_i : \Omega \to \Xi$, $i \geq 1$, be a sequence of i.i.d. random variables with common law μ, and a divergent and nondecreasing sequence of sample sizes $(n_k)_{k\in\mathbb{N}} \subseteq \mathbb{N}\backslash\{0\}$.*

For notational convenience, we omit the symbol $\omega \in \Omega$ for the functions ξ_i. In particular, using the notation already introduced in (2) for a normal integrand $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$, in what follows we shall write

$$\varphi_{\xi_i}(x) := \varphi(\xi_i(\omega), x).$$

To establish convergence of the PSS method, we need to control the error associated with replacing the expected value with the empirical average in (6) during the iteration process. The following lemma furnishes a nonasymptotic bound for the uniform error of the average approximation generated by the sampling.

Lemma 2.8 *Let $(\xi_i)_{i\geq 1}$ be a sampling with sample size $(n_k)_{k\in\mathbb{N}}$ satisfying Assumption 1, and let $F : \Xi \times V \to \mathbb{R}^m$ be a measurable mapping with $V \subset \mathbb{R}^d$ a closed and bounded set. Suppose that there exist $\kappa \in L^2_\mu$ and $\bar{x} \in V$ such that*

$$F(\cdot, \bar{x}) \in L^2_\mu \text{ and } \|F(\xi, y) - F(\xi, z)\| \leq \kappa(\xi)\|y - z\|, \text{ for all } y, z \in V, \xi \in \Xi. \tag{7}$$

Then, the following hold.

(i) There exists $a > 0$ such that

$$\mathbb{E}\left(\sup_{x\in V}\left\|\int_\Xi F(\xi, x)d\mu(\xi) - \frac{1}{k}\sum_{i=1}^k F(\xi_i, x)\right\|\right) \leq \frac{a}{\sqrt{k}}, \quad \text{for all } k \geq 1. \tag{8}$$

(ii) In particular, let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ be a normal integrand which is locally Lipschitz around any point $x \in \mathbb{R}^d$ with square integrable modulus. Then, for all $r > 0$, there exists a constant a_r such that

$$\mathbb{E}\left(\sup_{\|x\|\leq r}\left|\mathbf{E}_\varphi(x) - \mathbf{E}^k_\varphi(x)\right|\right) \leq \frac{a_r}{\sqrt{n_k}}, \quad \text{for all } k \in \mathbb{N}. \tag{9}$$

(iii) There exists a measurable set $\widehat{\Omega}$ with $\mathbb{P}(\Omega\backslash\widehat{\Omega}) = 0$ such that for all $\omega \in \widehat{\Omega}$ and every sequence $(y^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$ with $y^k \to y$ we have that $\mathbf{E}^k_\varphi(y^k) \to \mathbf{E}_\varphi(y)$.

Proof. The first assertion follows directly from [60, Theorem 2.1 and Proposition 2.6], applied to the family of real-valued functions

$$G(\xi, x^*, x) := \langle x^*, F(\xi, x)\rangle, \qquad \xi \in \Xi, \quad (x^*, x) \in \mathbb{B}_1[0] \times V. \tag{10}$$

Indeed, this application yields precisely estimate (8). Moreover, since φ is Lipschitz continuous with respect to x on the compact set $\mathbb{B}_r[0]$, with a square-integrable Lipschitz modulus, the same argument applies to φ. Therefore, (9) follows from (i). Finally, the pointwise convergence follows from [88, Theorem 9.60].

□

The following lemma establishes an almost sure convergence rate for the uniform gap between the empirical expectation and its exact counterpart. Its proof relies on the technique of *normalized convergence*; see [35, 36] and the references therein for further details.

Lemma 2.9 *Under the assumptions of Lemma 2.8, we have a.s. that*

$$\sup_{x\in V} \left\| \int_{\Xi} F(\xi, x) d\mu(\xi) - \frac{1}{k}\sum_{i=1}^{k} F(\xi_i, x) \right\| = \mathcal{O}\left(\frac{\ln(k)}{\sqrt{k}}\right). \tag{11}$$

Proof. A closer inspection of the proof of [60, Theorem 2.2] (see the first inequality in [60, p. 1283], equation (5.8) therein, and the first equation in [60, p. 1284]) applied to the function G defined in (10), shows that there exist constants $a, b > 0$ and $k_0 \in \mathbb{N}$ such that, for every $k \geq k_0$ and every $\varepsilon > b/\sqrt{k}$,

$$\mathbb{P}\left(\left\{ \sup_{x\in V} \left\| \int_{\Xi} F(\xi, x)\, d\mu(\xi) - \frac{1}{k}\sum_{i=1}^{k} F(\xi_i, x) \right\| > \varepsilon \right\} \cap B_k \right) \leq \exp\big(-a\sqrt{k}\,\varepsilon\big), \tag{12}$$

where $B_k := \left\{ \frac{1}{k}\sum_{j=1}^{k} \kappa(\xi_j)^2 \leq 2\mathbb{E}\big[\kappa(\xi)^2\big] \right\}$, and κ denotes the Lipschitz modulus appearing in (7). Now, without loss of generality we may assume that $\mathbb{E}\big[\kappa(\xi)^2\big] > 0$. By the strong law of large numbers (see, e.g., [34, Theorem 8.3.5]), $\frac{1}{k}\sum_{j=1}^{k} \kappa(\xi_j)^2 \longrightarrow \mathbb{E}\big[\kappa(\xi)^2\big]$, a.s. Hence, by Egorov's theorem [17, Theorem 2.2.1], for every $\varsigma > 0$ there exists a measurable set $\Omega_\varsigma \subseteq \Omega$ such that $\mathbb{P}(\Omega \setminus \Omega_\varsigma) \leq \varsigma$, and the convergence above is uniform on Ω_ς. Consequently, there exists an integer $k_1 \geq k_0$ such that $\Omega_\varsigma \subseteq B_k$ for all $k \geq k_1$. Now fix $\delta > a^{-1}$, and choose $k_2 \geq k_1$ such that $\delta \ln(k) > b$ for all $k \geq k_2$. For such k, we may apply (12) with $\varepsilon_k := \delta\frac{\ln(k)}{\sqrt{k}}$. Define

$$A_k^\varsigma := \left\{ \omega \in \Omega : \frac{\sqrt{k}}{\ln(k)} \sup_{x\in V} \left\| \int_{\Xi} F(\xi, x)\, d\mu(\xi) - \frac{1}{k}\sum_{i=1}^{k} F(\xi_i, x) \right\| > \delta \right\} \cap \Omega_\varsigma.$$

Since $\Omega_\varsigma \subseteq B_k$ for every $k \geq k_2$, we have

$$A_k^\varsigma \subseteq \left\{ \sup_{x\in V} \left\| \int_{\Xi} F(\xi, x)\, d\mu(\xi) - \frac{1}{k}\sum_{i=1}^{k} F(\xi_i, x) \right\| > \varepsilon_k \right\} \cap B_k.$$

Therefore, by (12),

$$\mathbb{P}(A_k^\varsigma) \leq \exp\big(-a\sqrt{k}\varepsilon_k\big) = \exp\big(-a\delta\ln(k)\big) = \frac{1}{k^{a\delta}}, \qquad \text{for all } k \geq k_2.$$

Since $a\delta > 1$, the Borel–Cantelli lemma [17, Exercise 1.12.89 (i)] yields

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k^\varsigma) < +\infty \implies \mathbb{P}\left(\limsup_{k\to\infty} A_k^\varsigma\right) = 0.$$

Consequently, for almost every $\omega \in \Omega_\varsigma$, there exists $\widehat{k}(\omega) \in \mathbb{N}$ such that, for every $k \geq \widehat{k}(\omega)$,

$$\sup_{x\in V} \left\| \int_{\Xi} F(\xi, x)\, d\mu(\xi) - \frac{1}{k}\sum_{i=1}^{k} F(\xi_i, x) \right\| \leq \delta\frac{\ln(k)}{\sqrt{k}}.$$

Since $\varsigma > 0$ was arbitrary, we conclude that

$$\sup_{x\in V} \left\| \int_{\Xi} F(\xi, x)\, d\mu(\xi) - \frac{1}{k}\sum_{i=1}^{k} F(\xi_i, x) \right\| = \mathcal{O}\left(\frac{\ln(k)}{\sqrt{k}}\right), \qquad \text{a.s.}$$

This proves (11). □

We establish in the next lemma the convergence of the Clarke sample-based subgradients. Formally, we have the following result.

Lemma 2.10 *Let $(\xi_i)_{i\geq 1}$ be a sampling with sample size $(n_k)_{k\in\mathbb{N}}$ satisfying Assumption 1, and let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be an upper-C^2 normal integrand. Then there exists a measurable set $\widehat{\Omega} \subset \Omega$ with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$ such that, for all $\omega \in \widehat{\Omega}$, every sequence $x^k \to x$ and every $v^k \in \partial \mathbf{E}^k_\varphi(x^k)$, all the accumulation points of $(v^k)_{k\in\mathbb{N}}$ belong to $\partial \mathbf{E}_\varphi(x)$.*

Proof. For all $x, h \in \mathbb{R}^d$, $k \in \mathbb{N}$, define $f(x,h) := \big(\mathbf{E}_\varphi\big)^\circ(x;h)$ and $f^k(x,h) := \big(\mathbf{E}^k_\varphi\big)^\circ(x;h)$. By the Clarke interchange formula for subdifferentiation under the integral [27, Theorem 2.7.2], together with the Clarke regularity of $-\varphi_\xi$ for every $x \in \mathbb{R}^d$, we have

$$\partial \mathbf{E}_\varphi(x) = \int_\Xi \partial \varphi_\xi(x)\, d\mu(\xi), \qquad f(x,h) = \int_\Xi \varphi^\circ_\xi(x;h)\, d\mu(\xi),$$

for all $h \in \mathbb{R}^d$. Similarly, by the sum rule for the Clarke subdifferential and the Clarke regularity of each function φ_ξ [27, Proposition 2.3.3 and Corollary 3], we obtain

$$\partial \mathbf{E}^k_\varphi(x) = \frac{1}{n_k}\sum_{j=1}^{n_k} \partial \varphi_{\xi_j}(x), \qquad f^k(x,h) = \frac{1}{n_k}\sum_{j=1}^{n_k} \varphi^\circ_{\xi_j}(x;h), \tag{13}$$

for all $x, h \in \mathbb{R}^d$ and all $k \in \mathbb{N}$. Therefore, by [6, Theorem 2.3], there exists a measurable set $\widehat{\Omega} \subset \Omega$, with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$, such that, for every $\omega \in \widehat{\Omega}$, the sequence $(f^k)_{k\in\mathbb{N}}$ hypographically converges to f on $\mathbb{R}^d \times \mathbb{R}^d$.

Next, fix $\omega \in \widehat{\Omega}$. Let $x^{\nu_k} \to x$ and $v^{\nu_k} \to v$, with $v^{\nu_k} \in \partial \mathbf{E}^{\nu_k}_\varphi(x^{\nu_k})$ for all $k \in \mathbb{N}$. We prove that $v \in \partial \mathbf{E}_\varphi(x)$. Indeed, for every $h \in \mathbb{R}^d$, the definition of the Clarke subdifferential gives

$$\langle v^{\nu_k}, h\rangle \leq \big(\mathbf{E}^{\nu_k}_\varphi\big)^\circ(x^{\nu_k};h) = f^{\nu_k}(x^{\nu_k},h), \qquad \text{for all } k \in \mathbb{N}.$$

Since $(x^{\nu_k}, h) \to (x,h)$ and f^k hypographically converges to f, [85, Eq. 7(9)] implies

$$\langle v, h\rangle = \lim_{k\to\infty} \langle v^{\nu_k}, h\rangle \leq \limsup_{k\to\infty} f^{\nu_k}(x^{\nu_k},h) \leq f(x,h) = \big(\mathbf{E}_\varphi\big)^\circ(x;h).$$

Since $h \in \mathbb{R}^d$ was arbitrary, it follows from the characterization of the Clarke subdifferential that $v \in \partial \mathbf{E}_\varphi(x)$, which proves the desired inclusion. □

2.5 Technical lemmas on real sequences

To conclude this section, we state two technical lemmas involving sequences that appear in the subsequent convergence analysis. We defer the proof of these results to Appendix A.

The first technical lemma shows the convergence of a series involving the sequence of sample sizes used to approximate the expected value in $(\mathcal{P})$, which is instrumental for proving subsequential convergence in Theorem 3.5. The proof of the following result can be found in Appendix A.1.

Lemma 2.11 *Let $(n_k)_{k\in\mathbb{N}}$ be a divergent, nondecreasing sequence of natural numbers. Then*

$$\sum_{k=1}^{\infty} \frac{n_{k+1} - n_k}{\sqrt{n_k} n_{k+1}} < +\infty.$$

The second lemma of this section characterizes the speed of convergence of a sequence satisfying a recursion up to errors with polynomial decay. Therein, the error term w_k naturally determines the rate of convergence. Its proof can be found in Appendix A.3. To the best of our knowledge, this result has not been shown in the literature. Typically, recursions of the type (14) below appear with no error terms, see e.g. [2, Lemma 1] and [16, Lemma 2]. We refer to [26] and [79, Lemmas 4 & 5, Chap.2, p.45] for recursion formulas with decaying errors.

Lemma 2.12 *Let $(\alpha_k)_{k\in\mathbb{N}}$ be a nonnegative sequence converging to zero, and $(w_k)_{k\in\mathbb{N}}$ be a nonnegative sequence such that for some $p_1, p_2 > 0$,*

$$w_k = \mathcal{O}\big(k^{-p_2}\ln^{p_1}(k)\big).$$

Suppose that there exist $c > 0$ and $q \in (0,2)$, such that

$$\alpha_{k+1}^q \leq c(\alpha_k - \alpha_{k+1}) + w_k, \quad \text{for all } k \in \mathbb{N}. \tag{14}$$

Then the following hold.

(i) If $q \in (0,1)$, then $\alpha_k = o\big(k^{-p_2}\ln^{p_1}(k)\big)$;
(ii) If $q = 1$, then $\alpha_k = \mathcal{O}\big(k^{-p_2}\ln^{p_1}(k)\big)$;
(iii) If $q \in (1,2)$, $p_1 = q$ and $p_2 > 1$, then $\alpha_k = \mathcal{O}\left(k^{-\min\{p_2, \frac{q}{q-1}\}+1}\ln(k)\right)$.

3 Proximal stochastic subgradient algorithm

In this section, we formulate and analyze a PSS algorithm for solving the nonconvex problem ($\mathcal{P}$), corresponding to Algorithm 1. Throughout this section and thereafter, we have the following blanket assumptions.

Assumption 2 (On problem ($\mathcal{P}$)) *Let $\Phi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ be the objective function of problem ($\mathcal{P}$). We assume the following:*

(i) $\psi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a proper, lsc, prox-bounded function;
(ii) $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ is an upper-C^2 normal integrand;
(iii) The function $\xi \mapsto \inf_{x\in\mathbb{R}^d}(\varphi(\xi,x)+\psi(x))$ is integrable.

Remark 3.1 A few comments on Assumption 2 are in order.

(i) It is common in the literature to impose smoothness assumptions on the integrand φ and require the stochastic gradient to be unbiased, that is, $\mathbb{E}\big[\nabla_x\varphi(\xi_i,x)\big] = \nabla\mathbf{E}_\varphi(x)$ for a suitable class of points $x \in \mathbb{R}^d$. However, this property is inapplicable in our setting as the integrand is nonsmooth. Instead, one can show that any measurable selection $v_i^k \in \partial_x\varphi(\xi_i,x)$ satisfies $\mathbb{E}[v_i^k] \in \partial\mathbf{E}_\varphi(x)$. This condition is closely related to the one used in [31], where the authors consider a measurable selection of subgradients $g : \Xi\times\mathbb{R}^d \to \mathbb{R}^d$ such that $\mathbb{E}_\mu[g(\xi,x)] \in \partial\mathbf{E}_\varphi(x)$ for all $x \in U$, where U is an open set containing the constraint set. Furthermore, the latter work imposes a global second-moment bound on the stochastic subgradient oracle $g : \Xi\times\mathbb{R}^d \to \mathbb{R}^d$, namely, $\sup_{x\in U}\mathbb{E}_\mu\big[\|g(\xi,x)\|^2\big] < +\infty$. In contrast, since we assume only local Lipschitz continuity of φ; see Definition 2.3; such a global moment bound may fail in our framework.

(ii) Let us emphasize that, in the definition of an upper-C^2 normal integrand, we implicitly assume that the integrand is locally Lipschitz in the sense of Definition 2.3. This assumption is closely related to the standard uniform bounded-variance condition

$$\sup_{x\in\mathbb{R}^d}\mathbb{E}_\mu\left[\|\nabla_x\varphi_\xi(x) - \nabla\mathbf{E}_\varphi(x)\|^2\right] < +\infty,$$

or to its corresponding version on a prescribed subset of $\mathbb{R}^d$; see, e.g., [44,45] and the references therein. More recently, this condition has been weakened by allowing multiplicative noise [54]. Particularly, the stochastic oracle is allowed to have a variance that grows with the distance between points, rather than being uniformly bounded on the whole set. Moreover, when φ is convex with respect to the decision variable $x \in \mathbb{R}^d$, the multiplicative-noise assumption implies local Lipschitz continuity of the integrand in the sense of Definition 2.3 provided that C is an open convex set and that there exists $\widehat{x} \in \mathbb{R}^d$ such that $\varphi(\cdot,\widehat{x}) \in L_\mu^2$. Indeed, for convex normal integrands, suitable boundedness of the values is equivalent to Lipschitz continuity of the integrand; see, e.g., [29, Theorem 2].

The method proposed in Algorithm 1 approximates the expected value in ($\mathcal{P}$) by means of an empirical average that is calculated in each iteration by adding new terms to the sum coming from new samples of the underlying random variable of the model. Then, a proximal subgradient step is performed on the approximation, where the stepsize is defined via line search. Observe that, in principle, the line search can restart the initial candidate stepsize in each (outer) iteration, thus allowing a nonmonotone sequence of stepsizes. In addition, note that only one stochastic subgradient is computed every time the line search

is conducted, in contrast to other proposed line searches for stochastic methods that require new gradient computations at every trial; see, e.g., [75, 76].

Algorithm 1 Proximal Stochastic Subgradient (PSS) algorithm for problem ($\mathcal{P}$)

Input: $x^0 \in \operatorname{dom}\psi$, $\rho \in (0,1)$, $\sigma > 0$, $\gamma_{\min} > 0$ and sample size $(n_k)_{k\in\mathbb{N}}$ with $0 < n_k \leq n_{k+1}$ and $n_k \to \infty$ as $k \to \infty$. Set $n_{-1} := 0$ and $k := 0$.

1: If $n_k > n_{k-1}$, draw $\xi_j \sim \mu$ i.i.d. for $j = n_{k-1}+1, \ldots, n_k$.

2: Choose $v_i^k \in \partial\varphi_{\xi_i}(x^k)$ for $i = 1, \ldots, n_k$.

3: Define $v^k := \frac{1}{n_k}\sum_{i=1}^{n_k} v_i^k$ and set Φ_k using (6). ▷ Subgradient estimate

4: Take $\overline{\gamma}_k \in [\gamma_{\min}, \gamma^\psi[$. Set $\gamma_k := \overline{\gamma}_k$.

5: Compute ▷ Proximal line search

$$\widehat{x}^k \in \operatorname{prox}_{\gamma_k\psi}\left(x^k - \gamma_k v^k\right). \tag{15}$$

6: **while** $\Phi_k(\widehat{x}^k) > \Phi_k(x^k) - \sigma\|\widehat{x}^k - x^k\|^2$ ▷ Sample-based descent test
do $\gamma_k \leftarrow \rho\gamma_k$ and recompute $\widehat{x}_k$ according to (15).

7: Set $x^{k+1} := \widehat{x}^k$, update $k \leftarrow k+1$, and go back to Step 1.

3.1 Well-posedness of the iterative scheme in Algorithm 1

We start the analysis of Algorithm 1 by showing it is well-defined, which accounts for proving that the line search terminates after finitely many trials in each iteration. We first show a technical proposition from which the desired result will follow.

Proposition 3.2 *Let φ and ψ be functions satisfying Assumption 2(i)–(ii), and suppose that Assumption 1 holds for the sampling ξ_i, with $i \geq 1$. Let $\bar{x} \in \operatorname{dom}\psi$ and $\sigma > 0$. Then, there exists a measurable set $\widehat{\Omega}$ with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$ such that, for all $\omega \in \widehat{\Omega}$, there exist a neighborhood W of $\bar{x}$ and a constant $\vartheta > 0$ such that for all $x \in W$, $k \in \mathbb{N}$, and $v \in \partial\mathbf{E}_\varphi^k(x)$, one has*

$$\Phi_k(w) \leq \Phi_k(x) - \sigma\|w - x\|^2,$$

provided that

$$w \in \operatorname{prox}_{\gamma\psi}(x - \gamma v), \quad \text{with } \gamma \in (0, \vartheta). \tag{16}$$

Proof. First, by Proposition 2.4 and Proposition 2.7, for every $r \in \mathbb{N}$, we can consider $\kappa_r \in L^2_\mu$ and $\ell_r \in L^1_\mu$ such that (3) and (4) hold on $\mathbb{B}_r[0]$. Now, by the strong law of large numbers, we can assume that there exists a measurable set Ω_r with $\mathbb{P}(\Omega_r) = 1$ such that for all $\omega \in \Omega_r$

$$\varsigma_r := \sup_{k\in\mathbb{N}} \frac{1}{n_k}\sum_{i=1}^{n_k} \ell_r(\xi_i) < +\infty \quad \text{and} \quad \widetilde{\varsigma}_r := \sup_{k\in\mathbb{N}} \frac{1}{n_k}\sum_{i=1}^{n_k} \kappa_r(\xi_i) < +\infty.$$

Now, we set $\widehat{\Omega} := \bigcap_{r\in\mathbb{N}} \Omega_r$, a set of full measure. Fix $\omega \in \widehat{\Omega}$, $\bar{x} \in \operatorname{dom}\psi$ and $\sigma > 0$, and take $r_0 \in \mathbb{N}$ such that $\|\bar{x}\| \leq r_0 - 1$. Let V be a neighborhood of $\bar{x}$ such that $V \subset \mathbb{B}_{r_0}[0]$. In addition, for any $x \in V$ and $v \in \partial\mathbf{E}_\varphi^k(x)$, in view of (13) and [27, Proposition 2.1.2], there exist $v_i \in \partial\varphi_{\xi_i}(x)$, $j = 1, \ldots, n_k$, such that

$$\|v\| \leq \frac{1}{n_k}\sum_{i=1}^{n_k} \|v_i\| \leq \frac{1}{n_k}\sum_{i=1}^{n_k} \kappa_{r_0}(\xi_i) \leq \widetilde{\varsigma}_{r_0}.$$

Hence, the set $\{v \in \partial\mathbf{E}_\varphi^k(x) : x \in V,\ k \in \mathbb{N}\}$ is bounded. Then, by Lemma 2.1, we can consider a neighborhood $W \subseteq V$ of $\bar{x}$, a positive constant ϑ_0 so that any vector w given by (16) with $x \in W$ and $\gamma \in (0, \vartheta_0)$ belongs to V. By resorting again to (4), we have that, for all $k \in \mathbb{N}$ and all ξ_i, $i = 1, \ldots, n_k$, the following inequality holds

$$\varphi_{\xi_i}(w) - \varphi_{\xi_i}(x) - \ell_{r_0}(\xi_i)\|w - x\|^2 \leq \langle v_i, w - x\rangle.$$

By summing the above inequalities for $i = 1, \ldots, n_k$ and dividing by n_k, we get

$$\mathbf{E}_\varphi^k(w) - \mathbf{E}_\varphi^k(x) - \frac{1}{n_k}\sum_{i=1}^{n_k} \ell_{r_0}(\xi_i)\|w - x\|^2 \leq \langle v, w - x\rangle, \quad \text{for all } k \in \mathbb{N}. \tag{17}$$

By (16), the definition of the proximity operator and rearranging terms we get

$$\psi(w) + \langle v, w - x\rangle + \frac{1}{2\gamma}\|w - x\|^2 \leq \psi(x). \tag{18}$$

Summing (17) and (18) yields

$$\Phi_k(w) + \left(\frac{1}{2\gamma} - \frac{1}{n_k}\sum_{i=1}^{n_k} \ell_{r_0}(\xi_i)\right)\|w - x\|^2 \leq \Phi_k(x), \quad \text{for all } k \in \mathbb{N}.$$

Therefore, it suffices to take $\vartheta := \min\{\vartheta_0, \big(2(\sigma + \varsigma_{r_0})\big)^{-1}\}$ to obtain the desired result for all $k \in \mathbb{N}$. □

A direct consequence of Proposition 3.2 is that the line search procedure in Algorithm 1 is well-defined, as shown in the next result.

Corollary 3.3 (Finite termination of line search) *In Algorithm 1, under Assumption 2(i)–(ii), for each iteration $k \in \mathbb{N}$, the line search procedure described in Step 6 terminates after a finite number of (inner) steps.*

Proof. Fix $k \in \mathbb{N}$. Apply Proposition 3.2 with $\bar{x} = x = x^k$ and $v = v^k \in \partial \mathbf{E}^k_\varphi(x^k)$, where the inclusion holds from (13). Then, after finitely many trials of the line search, $\gamma_k \in (0, \vartheta)$, thus the line search terminates with $\Phi_k(\widehat{x}^k) \leq \Phi_k(x^k) - \sigma\|\widehat{x}^k - x^k\|^2$. □

3.2 Subsequential convergence analysis of PSS

We now proceed to show subsequential convergence of the sequence generated by Algorithm 1, for which we need the following assumption.

Assumption 3 (On Algorithm 1) *Let $(n_k)_{k\in\mathbb{N}} \subseteq \mathbb{N}\backslash\{0\}$ be a nondecreasing sequence of sample sizes such that $n_k \to +\infty$, and let $(x^k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. We consider a filtration $\mathbb{F} := (\mathcal{F}_k)_{k\in\mathbb{N}}$, representing all the information used by Algorithm 1 up to iteration k, such that ξ_j is $\mathcal{F}_k$-measurable for every $j \leq n_k$, and ξ_j is independent of $\mathcal{F}_k$ whenever $j > n_k$.*

Remark 3.4 (Measurability of the sequence generated by Algorithm 1) It is important to note that there is no guarantee that the sequence produced by Algorithm 1 will be a sequence of measurable functions, even more adapted to the filtration $\mathbb{F}$ in Assumption 1. This issue arises primarily from the fact that the first order information considered in the algorithm comes from set-valued mappings (Step 2), or from the proximal subproblem (15) possibly admitting multiple solutions. A common approach to overcome this issue is to assume the existence of a selector for the subdifferential and for the proximal mapping in (15). Under these assumptions, the sequence $(x^k)_{k\in\mathbb{N}}$ becomes an adapted process with respect to the filtration, which is sufficient for analyzing the asymptotic behavior of the sequence. However, we prefer to state this condition as an assumption rather than impose restrictions on the potential applications of Algorithm 1.

We first state the main theorem of this section. Then, in Section 3.2.1, we present some technical lemmas instrumental for the proof of the theorem, which will be developed in Section 3.2.2.

Theorem 3.5 (Subsequential convergence) *Let $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ denote the objective function of problem ($\mathcal{P}$), and assume that Assumption 2 holds. Given an initial point $x^0 \in \operatorname{dom}\Phi$, consider the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 under Assumptions 1 and 3. Then, there exists a measurable set $\widehat{\Omega}$ with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$, such that for all $\omega \in \widehat{\Omega}$, if the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded, then both $\big(\Phi_k(x^k)\big)_{k\in\mathbb{N}}$ and $\big(\Phi(x^k)\big)_{k\in\mathbb{N}}$ converge to the same limit,*

$$\sum_{k\in\mathbb{N}} \|x^{k+1} - x^k\|^2 < +\infty, \tag{19}$$

and the following assertions hold:

(i) $\inf_{k\in\mathbb{N}} \gamma_k > 0$;

(ii) if $x^{k_j} \to \bar{x}$ as $j \to \infty$, then $\bar{x}$ is a stationary point of problem ($\mathcal{P}$) in the sense of (1), and $\Phi(\bar{x}) = \lim_{k\to\infty} \Phi_k(x^k) = \lim_{k\to\infty} \Phi(x^k)$;

(iii) the set of accumulation points of $(x^k)_{k\in\mathbb{N}}$ is nonempty, closed, and connected;

(iv) if $(x^k)_{k\in\mathbb{N}}$ has an isolated accumulation point $\bar{x}$, then the entire sequence $(x^k)_{k\in\mathbb{N}}$ converges to the stationary point $\bar{x}$ as $k \to \infty$.

3.2.1 Technical results for Theorem 3.5

We start the analysis by defining a stopping time related to the boundedness of the sequence $(x^k)_{k\in\mathbb{N}}$. Its proof is direct by definition.

Lemma 3.6 *Suppose that Assumption 3 holds. Given $r > 0$ such that $\|x^0\| \leq r$, define*

$$\tau_r := \inf\left\{k \in \mathbb{N} : \|x^{k+1}\| > r\right\},$$

with the convention $\inf \emptyset = +\infty$, and consider the stopped process

$$x_r^k := x^{k\wedge\tau_r}, \qquad k \wedge \tau_r := \min\{k, \tau_r\}, \qquad \text{for every } k \in \mathbb{N}.$$

Then τ_r is a stopping time with respect to the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ in Assumption 3. Furthermore, x_r^k is $\mathcal{F}_{k-1}$-measurable for every $k \in \mathbb{N}\backslash\{0\}$, and the stopped process $(x_r^k)_{k\in\mathbb{N}}$ satisfies

$$\tau_r \geq p \quad \Longrightarrow \quad \|x^p\| \leq r, \qquad \text{for every } p \in \mathbb{N}.$$

In particular,

$$\|x_r^k\| \leq r, \qquad \text{for every } k \in \mathbb{N}.$$

In the next lemma, we show that the stopped sequence defined in Lemma 3.6 satisfies a sufficient decrease condition up to errors.

Lemma 3.7 *Suppose that the assumptions of Theorem 3.5 hold. In the setting of Lemma 3.6 define the sequence of random variables*

$$\Phi_k^r := \Phi_{k\wedge\tau_r}(x^{k\wedge\tau_r}), \quad \text{for all } k \in \mathbb{N}.$$

Then, for all $k \in \mathbb{N}$, it holds that

$$\sigma\|x_r^{k+1} - x_r^k\|^2 \leq \Phi_k^r - \Phi_{k+1}^r + W_k^r, \tag{20}$$

where the random variable W_k^r is defined by

$$W_k^r := \left(\mathbf{E}_\varphi^{k+1}(x^{k+1}) - \mathbf{E}_\varphi^k(x^{k+1})\right)\mathbb{1}_{\{\tau_r \geq k+1\}}. \tag{21}$$

Proof. From the definition of x^{k+1} in Step 7 of Algorithm 1 it follows that

$$\Phi_k(x^{k+1}) \leq \Phi_k(x^k) - \sigma\|x^{k+1} - x^k\|^2, \quad \text{for all } k \in \mathbb{N}. \tag{22}$$

By (5)-(6), then

$$\sigma\|x^{k+1} - x^k\|^2 \leq \Phi_k(x^k) - \Phi_k(x^{k+1}) = \Delta\,\Phi_k + \mathbf{E}_\varphi^{k+1}(x^{k+1}) - \mathbf{E}_\varphi^k(x^{k+1}), \tag{23}$$

where

$$\Delta\,\Phi_k := \Phi_k(x^k) - \Phi_{k+1}(x^{k+1}), \quad \text{for all } k \in \mathbb{N}.$$

Given a fixed $k \in \mathbb{N}$, note that

$$\|x_r^{k+1} - x_r^k\| = \begin{cases} 0, & \text{if } \tau_r \leq k, \\ \|x^{k+1} - x^k\|, & \text{if } \tau_r \geq k+1, \end{cases} \tag{24}$$

and

$$\Phi_k^r = \begin{cases} \Phi_{\tau_r}(x^{\tau_r}), & \text{if } \tau_r \leq k, \\ \Phi_k(x^k), & \text{if } \tau_r \geq k+1. \end{cases} \tag{25}$$

Hence, on the one hand, using (23) and (24) on $\{\tau_r \geq k+1\}$, we have that

$$\sigma\|x_r^{k+1} - x_r^k\|^2 \leq \Delta\,\Phi_k + \mathbf{E}_\varphi^{k+1}(x^{k+1}) - \mathbf{E}_\varphi^k(x^{k+1}) = \Phi_k^r - \Phi_{k+1}^r + W_k^r.$$

On the other hand, on $\{\tau_r \leq k\}$ we simply have

$$\sigma\|x_r^{k+1} - x_r^k\|^2 = 0 \quad \text{and} \quad \Phi_k^r - \Phi_{k+1}^r + W_k^r = 0,$$

so we conclude that (20) holds. □

The estimate shown in Lemma 3.7 is not a descent condition in the traditional sense due to the random variables $(W_k^r)_{k\in\mathbb{N}}$, defined in (21), appearing on the right-hand side of (20). Nevertheless, as shown in the next result, (the conditional expectation of) this sequence is vanishing.

Lemma 3.8 *Let us assume the setting of Lemma 3.7. Then, for all $k \in \mathbb{N}$ and $r > 0$ such that $\|x^0\| \leq r$, the random variable W_k^r given in (21) is integrable. Furthermore, let $Z_k^r := \mathbb{E}\big[W_k^r | \mathcal{F}_k\big]$. Then $\sum_{k\in\mathbb{N}} \mathbb{E}\big[|Z_k^r|\big] < +\infty$.*

Proof. Let us consider an arbitrary $r > 0$ such that $\|x^0\| \leq r$. From Lemma 3.6, $\|x_r^k\| \leq r$ for all $k \in \mathbb{N}$. Furthermore, by Proposition 2.4, there exist square integrable functions κ and δ such that for all $\xi \in \Xi$,

$$|\varphi_\xi(u)| \leq \kappa(\xi)\|u\| + \delta(\xi), \quad \text{for all } u \in \mathbb{B}_r[0]. \tag{26}$$

Therefore, for all $k \in \mathbb{N}$,

$$|W_k^r| \leq \frac{1}{n_{k+1}} \sum_{i=1}^{n_{k+1}} (\kappa(\xi_i)r + \delta(\xi_i)) + \frac{1}{n_k} \sum_{i=1}^{n_k} (\kappa(\xi_i)r + \delta(\xi_i)) .$$

Hence, $\mathbb{E}\big[|W_k^r|\big] \leq 2\mathbb{E}[\kappa]r + 2\mathbb{E}[\delta]$. Moreover, observe that

$$W_k^r = \left(\frac{1}{n_{k+1}} \sum_{j=n_k+1}^{n_{k+1}} \varphi_{\xi_j}(x_r^{k+1}) + \frac{n_k - n_{k+1}}{n_{k+1}} \mathbf{E}_\varphi^k(x_r^{k+1}) \right) \mathbb{1}_{\{\tau_r \geq k+1\}} .$$

Now, using that $\{\tau_r \geq k+1\} \in \mathcal{F}_k$, because τ_r is a stopping time (Lemma 3.6), we get that

$$Z_k^r = \mathbb{E}\left[\frac{1}{n_{k+1}} \sum_{j=n_k+1}^{n_{k+1}} \varphi_{\xi_j}(x_r^{k+1}) + \frac{n_k - n_{k+1}}{n_{k+1}} \mathbf{E}_\varphi^k(x_r^{k+1}) \,\middle|\, \mathcal{F}_k \right] \mathbb{1}_{\{\tau_r \geq k+1\}} .$$

Recall from Lemma 3.6 that the function x_r^{k+1} is $\mathcal{F}_k$-measurable, and by (26) the function $\omega \mapsto \varphi_{\xi_j(\omega)}(x_r^{k+1}(\omega))$ is integrable, so using Lemma 2.2 we get that

$$\mathbb{E}\left[\varphi_{\xi_j}(x_r^{k+1}) \mid \mathcal{F}_k\right](\omega) = \mathbf{E}_\varphi(x_r^{k+1}(\omega)) \text{ a.s.},$$

for all $j = n_k + 1, \ldots, n_{k+1}$. This shows that

$$Z_k^r = \frac{n_{k+1} - n_k}{n_{k+1}} \left(\mathbf{E}_\varphi(x_r^{k+1}) - \mathbf{E}_\varphi^k(x_r^{k+1}) \right) \mathbb{1}_{\{\tau_r \geq k+1\}} \text{ a.s.}$$

Since $\|x_r^k\| \leq r$ for all $k \in \mathbb{N}$, it follows from Lemma 2.8(ii) that there exists a constant $a_r > 0$, independent of k, such that

$$\mathbb{E}\big[|Z_k^r|\big] \leq a_r \frac{n_{k+1} - n_k}{n_{k+1}} \frac{1}{\sqrt{n_k}}.$$

By Lemma 2.11, this yields $\sum_{k\in\mathbb{N}} \mathbb{E}\big[|Z_k^r|\big] < +\infty$, concluding the proof. □

We now prove that the sequence generated by Algorithm 1 has square-summable increments. In addition, we show that the sequence of value approximations along the iterates is convergent. This is stated in the following lemma.

Lemma 3.9 *Suppose the assumptions of Lemma 3.8 hold. Then, $(\Phi_k^r)_{k\in\mathbb{N}}$ converges a.s. and*

$$\sum_{k=0}^{\infty} \|x_r^{k+1} - x_r^k\|^2 < +\infty \text{ a.s.} \tag{27}$$

Proof. Let κ and δ be the functions in (26) (which exist in view of Proposition 2.4). Let us define

$$s_k := \frac{1}{n_k} \sum_{i=1}^{n_k} \big(r\kappa(\xi_i) + \delta(\xi_i)\big) - \mathfrak{i}_\psi^r \quad \text{and} \quad s_k^r := s_{k\wedge\tau_r}, \qquad \text{for all } k \in \mathbb{N},$$

where $\mathfrak{i}_\psi^r := \inf_{x\in\mathbb{B}_r[0]} \psi(x)$. It is clear that s_k^r is (square) integrable. Moreover, by (20) and (26),

$$-s_{k+1}^r \leq \Phi_{k+1}^r \leq \Phi_0^r + \sum_{i=0}^{k} W_i^r,$$

which particularly implies that Φ_{k+1}^r is integrable in view of Lemma 3.8. Next, we introduce the process

$$Y_k^r := \mathbb{E}\big[\, s_{k+1}^r - s_k^r \,\big|\, \mathcal{F}_k\big], \qquad \text{for all } k \in \mathbb{N}.$$

Notice that (20) yields, for all $k \in \mathbb{N}$,

$$\mathbb{E}\big[\Phi_{k+1}^r + s_{k+1}^r \mid \mathcal{F}_k\big] \leq \Phi_k^r + s_k^r - \sigma\|x_r^{k+1} - x_r^k\|^2 + |Y_k^r| + |Z_k^r|. \tag{28}$$

A direct computation yields

$$Y_k^r = \frac{n_{k+1} - n_k}{n_{k+1}} \left(\mathbb{E}\big[r\kappa(\xi) + \delta(\xi)\big] - \frac{1}{n_k} \sum_{i=1}^{n_k} r\kappa(\xi_i) + \delta(\xi_i) \right).$$

Hence, by Lemma 2.8, we conclude the existence of a constant $b_r > 0$ such that

$$\mathbb{E}\big[|Y_k^r|\big] \leq b_r \frac{n_{k+1} - n_k}{n_{k+1}} \frac{1}{\sqrt{n_k}},$$

where b_r is the variance of the square integrable function $\xi \mapsto r\kappa(\xi) + \delta(\xi)$. In this manner, Lemma 2.11 implies $\sum_{k\in\mathbb{N}} \mathbb{E}\big[|Y_k^r|\big] < +\infty$. Therefore, by Lemma 3.8, the nonnegativity of the sequence $\big(\Phi_k^r + s_k^r\big)_{k\in\mathbb{N}}$, and (28), it follows from the well-known Robbins–Siegmund supermartingale convergence theorem (see [84, Theorem 1]) that $\big(\Phi_k^r + s_k^r\big)_{k\in\mathbb{N}}$ converges and (27) holds. Moreover, by the strong law of large numbers we have, a.s., that

$$\lim_{k\to\infty} s_k^r = \begin{cases} \mathbb{E}(r\kappa + \delta) - \mathfrak{i}_\psi^r, & \text{if } \tau_r = +\infty, \\ s_{\tau_r} - \mathfrak{i}_\psi^r, & \text{if } \tau_r < +\infty. \end{cases}$$

Consequently, $\big(\Phi_k^r\big)_{k\in\mathbb{N}}$ converges a.s.

□

3.2.2 Proof of Theorem 3.5

We are now in a position to prove each assertion of Theorem 3.5.

Proof of Theorem 3.5. We consider a measurable set $\widehat{\Omega}$ of full measure such that for all $\omega \in \widehat{\Omega}$, the conclusions of Proposition 3.2 and the following conditions hold: for all $r \in \mathbb{N}$ with $\|x^0\| \leq r$,

a_ω) from Lemma 3.9,

$$\sum_{k\in\mathbb{N}} \|x_r^{k+1}(\omega) - x_r^k(\omega)\|^2 < +\infty \tag{29}$$

and Φ_k^r converges to some random variable;

b_ω) by Lemma 2.9 we get that

$$\big|\mathbf{E}_\varphi(x_r^k) - \mathbf{E}_\varphi^k(x_r^k)\big| \to 0;$$

c_ω) $\frac{1}{n_k}\sum_{j=1}^{n_k} \kappa_r(\xi_j(\omega)) \to \mathbb{E}\,[\kappa_r]$, where κ_r is chosen according to (3) in Proposition 2.4 for the bounded set $\mathbb{B}_r[0]$ with $r \in \mathbb{N}$.

Let $\omega \in \widehat{\Omega}$ such that $(x^k(\omega))_{k\in\mathbb{N}}$ is bounded. Let us consider $r \geq \sup_{k\in\mathbb{N}} \|x^k(\omega)\| + 1$. So, due to the equations (24) and (25), we get that $\Phi_k(x^k) = \Phi_k^r$ and $\|x^{k+1} - x^k\| = \|x_r^{k+1} - x_r^k\|$ for all $k \in \mathbb{N}$. Therefore, $\big(\Phi_k(x^k)\big)_{k\in\mathbb{N}}$ converges and (19) holds due to (29).

Moreover, by b_ω), (24) and the fact that $\tau_r = +\infty$, we get that

$$\big|\mathbf{E}_\varphi(x^k) - \mathbf{E}_\varphi^k(x^k)\big| \to 0,$$

and since $\Phi(x^k) - \Phi_k(x^k) = \mathbf{E}_\varphi(x^k) - \mathbf{E}_\varphi^k(x^k)$, we deduce that $\big(\Phi(x^k)\big)_{k\in\mathbb{N}}$ converges a.s. to the same limit as $\big(\Phi_k(x^k)\big)_{k\in\mathbb{N}}$.

Now, let us continue with item (i), showing that $\inf_{k\in\mathbb{N}} \gamma_k(\omega) > 0$. We proceed by way of contradiction, assuming that there exists a subsequence $(\gamma_{\nu_i})_{i\in\mathbb{N}}$ of $(\gamma_k)_{k\in\mathbb{N}}$ such that $\gamma_{\nu_i} \to 0^+$ as $i \to \infty$. Since $(x^k)_{k\in\mathbb{N}}$

is bounded, we may assume that $x^{\nu_i} \to \bar{x}$. Take $\lambda_{\nu_i} := \rho^{-1}\gamma_{\nu_i}$ and $w^{\nu_i} \in \operatorname{prox}_{\lambda_{\nu_i}\psi}(x^{\nu_i} - \lambda_{\nu_i} v^{\nu_i})$, where $v^{\nu_i} = \frac{1}{n_{\nu_i}} \sum_{j=1}^{n_{\nu_i}} v_j^{\nu_i}$ and $v_j^{\nu_i} \in \partial\varphi_{\xi_i}(x^{\nu_i})$ for all $j = 1, \dots, n_{\nu_i}$. In particular, λ_{ν_i} is the stepsize of one trial prior to passing the test in Step 6 of Algorithm 1 in iteration ν_i. Hence, for all $i \in \mathbb{N}$,

$$\Phi_{\nu_i}(w^{\nu_i}) > \Phi_{\nu_i}(x^{\nu_i}) - \sigma\|w^{\nu_i} - x^{\nu_i}\|^2.$$

Furthermore, since $\gamma_{\nu_i} \to 0^+$, then for all $i_0 \in \mathbb{N}$ sufficiently large, in view of Proposition 3.2, applied to $\bar{x}$, $x = x^{\nu_{i_0}}$ and $w = w^{\nu_{i_0}}$, we get that

$$\Phi_k(w^{\nu_{i_0}}) \leq \Phi_k(x^{\nu_{i_0}}) - \sigma\|w^{\nu_{i_0}} - x^{\nu_{i_0}}\|^2, \quad \text{for all } k \in \mathbb{N},$$

yielding a contradiction. Hence item (i) holds true.

To prove item (ii), let $x^{k_j}(\omega) \to \bar{x}$ as $j \to \infty$. We first prove that $\bar{x}$ is a stationary point of ($\mathcal{P}$). First, from item (i), we can assume that $\gamma_{\nu_i} \to \bar{\gamma} \in (0, +\infty]$. By (29), it follows that $x^{k_j+1} \to \bar{x}$. Moreover, by the definition of r, we conclude that $\bar{x} \in \mathbb{B}_{r'}(0)$, where $r' := r - \frac{1}{2}$. Since φ_{ξ_i} is Lipschitz on $\mathbb{B}_r(0)$ with modulus $\kappa_r(\xi_i)$, we have

$$\|v^{k_j}\| \leq \frac{1}{n_{k_j}} \sum_{i=1}^{n_{k_j}} \kappa_r(\xi_i), \text{ whenever } x^{k_j} \in \mathbb{B}_r(0),$$

which shows that v^{k_j} is bounded in view of c_ω). Therefore, up to a subsequence and without relabeling, we may assume that $v^{k_j} \to \bar{v}$. Now, by Lemma 2.10, we have that $\bar{v} \in \partial\mathbf{E}_\varphi(\bar{x})$. In addition, by (15) and the definition of the proximal operator, for all $u \in \mathbb{R}^d$, we have

$$\psi(x^{k_j+1}) + \langle v^{k_j}, x^{k_j+1}\rangle + \frac{1}{2\gamma_{k_j}}\|x^{k_j+1} - x^{k_j}\|^2 \leq \psi(u) + \langle v^{k_j}, u\rangle + \frac{1}{2\gamma_{k_j}}\|u - x^{k_j}\|^2. \tag{30}$$

Now, using the above inequality with $u = \bar{x}$, we get that $\limsup_{j\to\infty} \psi(x^{k_j+1}) \leq \psi(\bar{x})$, and from the lower semicontinuity of ψ we obtain $\psi(\bar{x}) \leq \liminf_{j\to\infty} \psi(x^{k_j+1})$, which allows to conclude that $\lim_{j\to\infty} \psi(x^{k_j+1}) = \psi(\bar{x})$. Note that Lemma 2.8(iii) particularly implies

$$\Phi_{k_j+1}(x^{k_j+1}) \to \Phi(\bar{x}). \tag{31}$$

In this manner, fixing $u \in \mathbb{R}^d$ in the right-hand side of (30), letting $j \to \infty$ and recalling that $\gamma_{k_j} \to \bar{\gamma}$, we deduce that

$$\psi(\bar{x}) \leq \psi(u) + \langle \bar{v}, u - \bar{x}\rangle + \frac{1}{2\bar{\gamma}}\|u - \bar{x}\|^2,$$

with the convention $\frac{1}{2\bar{\gamma}} = 0$ for $\bar{\gamma} = \infty$. Therefore, $\bar{x}$ is a stationary point of problem ($\mathcal{P}$) in the sense of (1). This proves the first claim in item (ii). For the second claim, let us show that $\Phi(\bar{x}) = \lim_{k\to\infty} \Phi_k(x^k) = \lim_{j\to\infty} \Phi_{k_j}(x^{k_j})$. We already know from the beginning of this proof that $\lim_{k\to\infty} \Phi_k(x^k) = \lim_{k\to\infty} \Phi(x^k)$ exists, and thus combined with (31) yields

$$\Phi(\bar{x}) = \lim_{j\to\infty} \Phi_{k_j+1}(x^{k_j+1}) = \lim_{k\to\infty} \Phi_{k+1}(x^{k+1}) = \lim_{k\to\infty} \Phi_k(x^k).$$

To conclude the proof, we now proceed to show that items (iii) and (iv) hold. First, by (29), it follows that $\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$ a.s. Hence, the sequence $(x^k)_{k\in\mathbb{N}}$ satisfies the *Ostrowski condition*, which implies directly our last claim (see, e.g., [38, Theorem 8.3.9 and Proposition 8.3.10]).

□

Remark 3.10 A few comments about Theorem 3.5 are in order.

(i) It is worth mentioning that if one requires that (a.s.) $\|x^k\| \leq r$ for all $k \in \mathbb{N}$ for some $r > 0$, then it is not difficult to show that $\tau_r = +\infty$ a.s., for every $r > \sup_{k\in\mathbb{N}} \|x^k\|$. Therefore, taking expectation in (28) we get that

$$\sum_{k\in\mathbb{N}} \mathbb{E}\|x^{k+1} - x^k\|^2 < +\infty.$$

and that $\big(\Phi_k(x^k)\big)_{k\in\mathbb{N}}$ converges to an integrable random variable.

(ii) Due to Theorem 3.5(i), if we construct the sequence of stepsizes $(\gamma_k)_{k\in\mathbb{N}}$ to be nonincreasing (i.e. setting $\bar{\gamma}_k = \gamma_{k-1}$ for all $k \in \mathbb{N}$ in Step 4 of Algorithm 1), then this sequence needs to stabilize. Indeed, by way of contradiction, suppose that the sequence of stepsizes does not stabilize. Then there are infinitely many iterations where the inner loop of the line search shrinks the stepsize at least a factor $\rho \in (0,1)$, contradicting the fact that $\inf_{k\in\mathbb{N}} \gamma_k > 0$. In other words, there exists an index $\bar{k}$ such that at iteration $k \geq \bar{k}$ the line search procedure terminates upon the first trial.

(iii) A distinctive feature of Theorem 3.5 is that it does not impose any prescribed growth rate on the sample sizes n_k; it is sufficient to assume that $(n_k)_{k\in\mathbb{N}}$ is a nondecreasing sequence satisfying $n_k \to \infty$. To the best of our knowledge, this level of flexibility is novel. Many classical and recent frameworks in the literature mandate pre-scheduled growth conditions on the sample sizes to establish subsequential convergence; see, e.g., [42, 54, 63–65]. Specifically, within the smooth, convex setting, the reduced-variable stochastic algorithm in [54] requires a sample-size sequence growing at a rate of at least $\mathcal{O}(k^3 \ln k)$. In nonconvex scenarios, the mini-batch sizes in [45] are assumed to scale at least linearly, and no trajectory-level convergence guarantees are obtained. Central limit theorems and asymptotic properties studied in [65] similarly rely on polynomial sample-size growth schedules of the form $n_k^{-1} = \mathcal{O}(k^\nu)$ with $\nu \geq 1$ to enforce noise stabilization. For nonconvex settings, even more demanding polynomial configurations are typically assumed; for instance, the stochastic difference-of-convex framework in [63] requires $n_k^{-1} = \mathcal{O}(k^\nu)$ with a power of at least $\nu \geq 2$, while related nonconvex schemes [64] obtain convergence guarantees with strictly increasing polynomial or geometric sample paths. In the infinite-dimensional Hilbert space setting, the convergence result in [42] needs the summability condition $\sum_{k\in\mathbb{N}} 1/n_k < +\infty$.

4 Convergence under the Kurdyka–Łojasiewicz condition

From the previous section, we see that the PSS is not a descent method, although the descent-type condition (22) is satisfied. In light of this, we now turn our attention to the convergence analysis of the (random) sequence produced by Algorithm 1 under the well-known KL property [61, 70, 78]. Firstly, let us recall this property.

A function $\Phi\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is said to satisfy the KL property at $\bar{x}$ if there exist a constant $\eta > 0$ and a continuous, concave function $\theta : [0,\eta] \to [0,+\infty[$, called the *desingularizing function*, such that $\theta(0) = 0$, θ is continuously differentiable on $]0,\eta[$ with $\theta' > 0$, and the inequality

$$\theta'\big(\Phi(x) - \Phi(\bar{x})\big)\operatorname{dist}\big(0;\partial\Phi(x)\big) \geq 1 \tag{32}$$

holds for all $x \in \mathbb{B}_\eta(\bar{x})$ satisfying $\Phi(\bar{x}) < \Phi(x) < \Phi(\bar{x}) + \eta$. Here, $\operatorname{dist}(\,\cdot\,;C)$ denotes the distance to the set C. It is known that the KL property holds at critical points of semi-algebraic and functions definable in o-minimal structures [19–21]. For functions defined in expected value form as the objective function in $(\mathcal{P})$, it is clear that the KL property holds when φ_ξ is, for instance, an analytic function for all ξ, and ψ is semi-algebraic.

Next, we impose an additional structural condition on the desingularizing function θ that has proven useful for establishing convergence rates; see, e.g., [56, 67]. Its validity for a broad class of desingularizing functions is discussed in detail in [56, Remark G.1].

Assumption 4 (Quasi-additivity type property) *Let θ be the desingularizing function in the KL condition (32) for which there exists a constant $C \in (0,1]$ such that for all $t_1, t_2 \in (0,\eta)$ with $t_1 + t_2 < \eta$, we have*

$$C\,[\theta'(t_1+t_2)]^{-1} \;\leq\; [\theta'(t_1)]^{-1} + [\theta'(t_2)]^{-1}. \tag{33}$$

Remark 4.1 (Exponential KL condition) Consider the exponential version of the KL condition, where the desingularizing function is given by $\theta(t) = Mt^{1-\beta}$ with M a positive constant and $\beta \in (0,1)$. Since the map $t \mapsto t^\beta$ is concave on $[0,+\infty[$, one obtains the inequality

$$(t_1+t_2)^\beta \leq t_1^\beta + t_2^\beta \qquad \forall\, t_1, t_2 \geq 0.$$

Therefore, the corresponding desingularizing function $\theta(t) = Mt^{1-\beta}$ satisfies Assumption 4 with constant $C = 1$ (see the discussion after [67, eq. (2.3)]). We refer to [66] for a treatment of the exponential version of the KL property.

In what follows, we say that a normal integrand φ is $C^{1,1}$ around a point $\bar{x}$ provided that φ_ξ is differentiable for all $\xi \in \Xi$, $\nabla_x \varphi(\cdot, \bar{x}) \in L^2_\mu$, and there exist $\kappa \in L^2_\mu$ and $\varepsilon > 0$ such that

$$\|\nabla_x \varphi(\xi, y) - \nabla_x \varphi(\xi, z)\| \leq \kappa(\xi)\|y - z\|, \quad \forall\, y, z \in \mathbb{B}_\varepsilon[\bar{x}], \forall\, \xi \in \Xi.$$

Moreover, we say that φ is $C^{1,1}$ around a nonempty set S, provided that it is $C^{1,1}$ around every $\bar{x} \in S$. When φ is differentiable at $x \in \mathbb{R}^d$, it is clear that the gradient of its empirical average (5) is given by

$$\nabla_x \mathbf{E}^k_\varphi(x) = \frac{1}{n_k} \sum_{j=1}^{n_k} \nabla_x \varphi_{\xi_j}(x).$$

4.1 Global convergence of PSS

The following analysis of the global convergence of the sequence generated by Algorithm 1 extends the arguments introduced in the seminal paper [12] to the stochastic framework considered here. We establish a global convergence result for Algorithm 1 under suitable assumptions on the sample sizes, which allow us to control the approximation error induced by the use of empirical averages. Recall that, due to this error, our scheme does not belong to the class of descent methods, thereby preventing a direct application of the classical KL convergence framework; see, e.g., [12, 16] and the references therein. We prove that, provided the approximation error vanishes at an appropriate rate, the use of sample averages does not alter the anticipated convergence properties of the generated sequence. More precisely, we obtain the following global convergence result.

Theorem 4.2 (Global convergence) *Suppose that the assumptions of Theorem 3.5 hold. Let $\mathcal{S}$ be the set of stationary points of Φ. Suppose that φ is $C^{1,1}$ around $\mathcal{S}$. Furthermore, let us assume that the sequence of sample sizes $(n_k)_{k\in\mathbb{N}}$ satisfies*

$$\sum_{k=1}^{\infty} \frac{\ln(n_k)}{\sqrt{n_k}} < +\infty. \tag{34}$$

Let $(x^k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Then, there exists a measurable set $\widehat{\Omega}$ with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$ such that, for every $\omega \in \widehat{\Omega}$, whenever the sequence $\big(x^k(\omega)\big)_{k\in\mathbb{N}}$ is bounded and has an accumulation point $\bar{x}$ where the KL condition holds with a desingularizing function θ satisfying Assumption 4 such that

$$\sum_{k=p_0}^{\infty} \left(\theta' \left(\sum_{t=k}^{\infty} \frac{\ln(n_t)}{\sqrt{n_t}} \right) \right)^{-1} < +\infty, \quad \text{for some } p_0 \in \mathbb{N}, \tag{35}$$

and that $\Phi(x^k) > \Phi(\bar{x})$ for all $k \geq p_0$, then the whole sequence $\big(x^k(\omega)\big)_{k\in\mathbb{N}}$ converges to $\bar{x}$ as $k \to \infty$.

We shall present the proof of Theorem 4.2 in Section 4.1.2 below. For this purpose, we first establish a set of preliminary results in the following section.

4.1.1 Technical results for Theorem 4.2

Let us start formally stating the following result, which is a simple consequence of the compactness of the set $\mathcal{S}_r := \{x \in \mathcal{S} : \|x\| \leq r\}$.

Lemma 4.3 *Assume that φ is $C^{1,1}$ around $\mathcal{S}$. Then for each $r \in \mathbb{N}$, there exist $\varepsilon_r > 0$, a function $\kappa_r \in L^2_\mu$, and finitely many points $\bar{x}^r_1, \ldots, \bar{x}^r_{m_r} \in \mathcal{S}$ such that*

$$\mathcal{S}_r \subseteq \bigcup_{i=1}^{m_r} \mathbb{B}_{\varepsilon_r/2}(\bar{x}^r_i),$$

and, for every $\xi \in \Xi$, every $i = 1, \ldots, m_r$, and all $y, z \in \mathbb{B}_{\varepsilon_r}[\bar{x}^r_i]$, one has

$$\big\|\nabla_x \varphi(\xi, y) - \nabla_x \varphi(\xi, z)\big\| \leq \kappa_r(\xi)\|y - z\|. \tag{36}$$

Let us recall that we already showed sufficient decrease conditions, up to errors, in (20) and (22), that play the role of [12, (H1)]. We next establish a relative error condition, akin to [12, (H2)], which, in combination with the sufficient decrease property, will lead to the convergence guarantees in Theorem 4.2.

Lemma 4.4 *Under the assumptions of Theorem 3.5, let $(x^k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Let $\mathcal{S}$ be the set of stationary points of Φ. Assume that φ is $C^{1,1}$ around $\mathcal{S}$. Then, there exists a measurable set $\widehat{\Omega}$ with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$, such that for every $\omega \in \widehat{\Omega}$, if $\bar{x}$ is an accumulation point of $\big(x^k(\omega)\big)_{k\in\mathbb{N}}$, there exist constants $\widehat{r}, \zeta > 0$ and an index $\widehat{k} \in \mathbb{N}$ such that*

$$\operatorname{dist}\big(0;\, \partial\, \Phi_k(x^{k+1})\big) \;\leq\; \zeta\, \|x^{k+1} - x^k\|, \quad \text{for all } k \geq \widehat{k} \text{ whenever } x^k \in \mathbb{B}_{\widehat{r}}[\bar{x}]. \tag{37}$$

Proof. Let $\widehat{\Omega}$ be a measurable set of full probability such that the conclusions of Theorem 3.5 hold. Furthermore, by Assumption 1 and the strong law of large numbers we can assume (shrinking the set if necessary) that for all $\omega \in \widehat{\Omega}$ and $r \in \mathbb{N}$,

$$\frac{1}{n_k} \sum_{j=1}^{n_k} \kappa_r(\xi_j) \to \mathbb{E}\left[\kappa_r\right], \tag{38}$$

as $k \to +\infty$, where κ_r are the (square-integrable) Lipschitz functions given in Lemma 4.3. Now, let $\omega \in \widehat{\Omega}$ and $\bar{x}$ be an accumulation point of $\big(x^k(\omega)\big)_{k\in\mathbb{N}}$. It follows by Theorem 3.5 that $\bar{x} \in \mathcal{S}$. Now, let us consider $r_0 \in \mathbb{N}$ such that $\|\bar{x}\| \leq r_0$. Due to (36) and (38), we can assume that there exist $L_1 > 0$ and $\widehat{r} \in (0, \varepsilon_r/2)$ such that $\mathbf{E}_\varphi^k(\,\cdot\,)$ is continuously differentiable with L_1-Lipschitz gradient on $\mathbb{B}_{2\widehat{r}}(\bar{x})$ for every $k \in \mathbb{N}$. In view of (19), let $\widehat{k} \in \mathbb{N}$ be such that

$$\|x^{k+1} - x^k\| \leq \widehat{r}, \quad \text{for all } k \geq \widehat{k}.$$

Hence, for every $x^k \in \mathbb{B}_{\widehat{r}}(\bar{x})$ with $k \geq \widehat{k}$, its next iterate x^{k+1} belongs to $\mathbb{B}_{2\widehat{r}}(\bar{x})$. Using the optimality condition of the problem in (15), we obtain

$$\frac{x^k - x^{k+1}}{\gamma_k} - \nabla_x \mathbf{E}_\varphi^k(x^k) \;\in\; \partial\psi(x^{k+1}).$$

This in turn implies that

$$d^k := \frac{x^k - x^{k+1}}{\gamma_k} \;+\; \nabla_x \mathbf{E}_\varphi^k(x^{k+1}) - \nabla_x \mathbf{E}_\varphi^k(x^k) \in \partial\, \Phi_k(x^{k+1}).$$

The L_1-Lipschitz continuity of $\nabla \mathbf{E}_\varphi^k$ on $\mathbb{B}_{2\widehat{r}}(\bar{x})$ yields

$$\|d^k\| \;\leq\; \left(\frac{1}{\gamma_k} + L_1\right) \|x^k - x^{k+1}\|.$$

Finally, by Theorem 3.5(i), there exists a sufficiently large constant $\zeta > 0$ such that

$$\operatorname{dist}\big(0;\, \partial\, \Phi_k(x^{k+1})\big) \;\leq\; \|d^k\| \;\leq\; \zeta\, \|x^{k+1} - x^k\|,$$

which concludes the proof. □

We continue with a technical result that establishes the setting to apply the KL property.

Lemma 4.5 *Suppose that the assumptions of Theorem 4.2 hold. Then, relative to a measurable set $\widehat{\Omega}$ with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$, the following assertions hold. Let $\bar{x}$ be an accumulation point of the sequence $\big(x^k(\omega)\big)_{k\in\mathbb{N}}$, for $\omega \in \widehat{\Omega}$, where the KL condition (32) is satisfied for a given $\eta > 0$. There exist $r_1, r_2 > 0$ such that $\mathbf{E}_\varphi^k$, $\mathbf{E}_\varphi$ are $C^{1,1}$ on $\mathbb{B}_{r_1}[\bar{x}]$, and $\mathbb{B}_\eta[\bar{x}] \subset \mathbb{B}_{r_2}[0]$. Moreover, for all $k \in \mathbb{N}$, define the (random) sequences*

$$a_k := 2 \sup\left\{\left|\mathbf{E}_\varphi(x) - \mathbf{E}_\varphi^k(x)\right| : \|x\| \leq r_2\right\}, \tag{39}$$

$$u_k := \sum_{t=k}^{\infty} a_t, \tag{40}$$

and

$$b_k := \sup\left\{\left\|\nabla_x \mathbf{E}_\varphi(x) - \nabla_x \mathbf{E}_\varphi^k(x)\right\| : x \in \mathbb{B}_{r_1}[\bar{x}]\right\}. \tag{41}$$

Then, for any $\varepsilon \in (0, \min\{\widehat{r}, \eta/2, 5r_1/6\})$*, there exists* $k_0 \in \mathbb{N}$ *such that for all* $k \geq k_0$*, the following conditions hold:*

$$\|x^{k+1} - x^k\| \leq \varepsilon/5; \tag{P1}$$

$$\Phi(x^k) - \Phi(\bar{x}) + u_k < \eta; \tag{P2}$$

$$\Phi(\bar{x}) < \Phi(x^k) < \Phi(\bar{x}) + \eta; \tag{P3}$$

$$\max\left\{\|x^{k_0} - \bar{x}\|, \frac{s_{k_0+1}}{\mathfrak{C}}, \frac{1}{\zeta}\sum_{k=k_0}^{\infty}\left(b_k + [\theta'(u_{k+1})]^{-1}\right)\right\} \leq \varepsilon/5; \tag{P4}$$

where $\mathfrak{C} := \zeta^{-1}C\sigma$*,* C *is the constant in the quasi-additive property* (33)*, and the constants* $\widehat{r}$ *and* ζ *are such that they satisfy* (37)*, and*

$$s_k := \theta\big(\Phi(x^k) - \Phi(\bar{x}) + u_k\big) \geq 0.$$

Proof. Let $\widehat{\Omega}$ be a measurable set with $\mathbb{P}(\Omega \setminus \widehat{\Omega}) = 0$, on which the conclusions of Theorem 3.5 and Lemma 4.4 hold. Moreover, applying Lemma 2.9 with $F = \varphi$, we obtain that, for every $r > 0$,

$$a_k^r := \sup_{\|x\| \leq r} \left|\mathbf{E}_\varphi(x) - \mathbf{E}_\varphi^k(x)\right| = \mathcal{O}\left(\frac{\ln(n_k)}{\sqrt{n_k}}\right) \quad \text{a.s.} \tag{42}$$

Furthermore, let $r \in \mathbb{N}$, and consider the points $\bar{x}_i^r$, $i = 1, \ldots, m_r$, together with the function κ_r, provided by Lemma 4.3. Applying Lemma 2.9 on each ball $\mathbb{B}_{\varepsilon_r}[\bar{x}_i^r]$, we obtain that, for every $r \in \mathbb{N}$ and every $i = 1, \ldots, m_r$,

$$\sup_{x \in \mathbb{B}_{\varepsilon_r}[\bar{x}_i^r]} \left\|\nabla_x \mathbf{E}_\varphi(x) - \nabla_x \mathbf{E}_\varphi^k(x)\right\| = \mathcal{O}\left(\frac{\ln(n_k)}{\sqrt{n_k}}\right) \quad \text{a.s.} \tag{43}$$

Thus, after possibly replacing $\widehat{\Omega}$ by another measurable subset of full probability, we may assume that, in addition, both estimates (42) and (43) hold for every $\omega \in \widehat{\Omega}$.

Fix now $\omega \in \widehat{\Omega}$, and let $\bar{x}$ be an accumulation point of the sequence $(x^k(\omega))_{k\in\mathbb{N}}$. Choose $r_2 > 0$ such that $\mathbb{B}_\eta[\bar{x}] \subseteq \mathbb{B}_{r_2}[0]$. By (42), it directly follows that,

$$a_k = a_k^{r_2} = \mathcal{O}\left(\frac{\ln(n_k)}{\sqrt{n_k}}\right). \tag{44}$$

Next, take $r \in \mathbb{N}$ such that $\|\bar{x}\| \leq r$. Since $\mathcal{S}_r \subseteq \bigcup_{i=1}^{m_r} \mathbb{B}_{\varepsilon_r/2}(\bar{x}_i^r)$, there exists $i \in \{1, \ldots, m_r\}$ such that $\bar{x} \in \mathbb{B}_{\varepsilon_r/2}(\bar{x}_i^r)$. Consequently, for every $r_1 \in (0, \varepsilon_r/2)$ sufficiently small, we have $\mathbb{B}_{r_1}[\bar{x}] \subseteq \mathbb{B}_{\varepsilon_r}[\bar{x}_i^r]$, with $\mathbf{E}_\varphi^k$ and $\mathbf{E}_\varphi$ being $C^{1,1}$ on $\mathbb{B}_{r_1}[\bar{x}]$. Therefore, by (43),

$$b_k = \sup_{x \in \mathbb{B}_{r_1}[\bar{x}]} \left\|\nabla_x \mathbf{E}_\varphi(x) - \nabla_x \mathbf{E}_\varphi^k(x)\right\| \leq \sup_{x \in \mathbb{B}_{\varepsilon_r}[\bar{x}_i^r]} \left\|\nabla_x \mathbf{E}_\varphi(x) - \nabla_x \mathbf{E}_\varphi^k(x)\right\| = \mathcal{O}\left(\frac{\ln(n_k)}{\sqrt{n_k}}\right). \tag{45}$$

Moreover, by (45) and our sampling-size assumption (34), we have $\sum_{k\in\mathbb{N}} b_k < +\infty$. Using (44) and (34), there are $m, \widehat{p} \in \mathbb{N}\setminus\{0\}$ such that

$$a_k \leq m \cdot \frac{\ln(n_k)}{\sqrt{n_k}} \quad \text{and} \quad \sum_{t=k}^{\infty} \frac{\ln(n_t)}{\sqrt{n_t}} < \frac{\eta}{2m}, \quad \text{for all } k \geq \widehat{p},$$

so that $(u_k)_{k\in\mathbb{N}}$ is well-defined, and

$$u_k \leq m \cdot \sum_{t=k}^{\infty} \frac{\ln(n_t)}{\sqrt{n_t}} < \frac{\eta}{2}, \quad \text{for all } k \geq \widehat{p}. \tag{46}$$

By using the quasi-additivity property of the desingularizing function θ followed by the fact that θ' is nonincreasing (as θ is concave), a simple induction argument shows that for every $m \in \mathbb{N}\setminus\{0\}$ and every $t > 0$ satisfying $mt \in (0, \eta)$, one has

$$[\theta'(mt)]^{-1} \leq \left(\frac{2}{C}\right)^{m-1} [\theta'(t)]^{-1},$$

from where we obtain for all $k \geq \widehat{p}$,

$$\left[\theta'(u_k)\right]^{-1} \leq \left[\theta'\left(m\sum_{t=k}^{\infty}\frac{\ln(n_t)}{\sqrt{n_t}}\right)\right]^{-1} \leq \left(\frac{2}{C}\right)^{m-1}\left[\theta'\left(\sum_{t=k}^{\infty}\frac{\ln(n_t)}{\sqrt{n_t}}\right)\right]^{-1}.$$

Consequently, by (35), we have

$$\sum_{k=\widehat{p}}^{\infty}\left[\theta'(u_k)\right]^{-1} < +\infty. \tag{47}$$

Altogether, the above implies that we can take an $\varepsilon \in (0, \min\{\widehat{r}, \eta/2, 5r_1/6\})$ and choose $k_0 \geq \max\{\widehat{k}, \widehat{p}\}$, where $\widehat{k}$ is from Lemma 4.4, so that for all $k \geq k_0$, (P1) follows from (19), while (P2) and (P3) follow from Theorem 3.5(ii) and (46). In particular, in view of (P2), the sequence $(s_k)_{k\in\mathbb{N}}$ is well-defined. Regarding (P4), if necessary, increase k_0 so that $\|x^{k_0} - \bar{x}\| \leq \varepsilon/5$ and $s_{k_0+1} \leq \varepsilon\mathfrak{C}/5$, where the latter is possible due to the continuity of θ'. The estimate of the remaining term follows from (34), (45) and (47), yielding (P4). □

The remainder of the analysis uses Lemma 4.5 to bound the distance from $(x^k)_{k\in\mathbb{N}}$ to the accumulation point $\bar{x}$. First, we bound the distance between two consecutive iterates x^{k+1} and x^{k+2}, whenever x^k is close enough to $\bar{x}$.

Lemma 4.6 *Under the assumptions of Lemma 4.5, for every $k \geq k_0$ with $x^k \in \mathbb{B}_\varepsilon[\bar{x}]$ it holds*

$$\|x^{k+2} - x^{k+1}\| \leq \frac{(s_{k+1} - s_{k+2})}{2\mathfrak{C}} + \frac{b_k}{2\zeta} + \frac{1}{2}\|x^{k+1} - x^k\| + \frac{1}{2\zeta}\left[\theta'(u_{k+1})\right]^{-1}.$$

Proof. Let us consider $x^k \in \mathbb{B}_\varepsilon[\bar{x}]$. As in Lemma 4.5, let $r_2 > 0$ such that $\mathbb{B}_\eta[\bar{x}] \subset \mathbb{B}_{r_2}[0]$. It follows from (P1) and the triangle inequality that $x^{k+1} \in \mathbb{B}_{r_1}[\bar{x}]$ and $x^{k+1}, x^{k+2} \in \mathbb{B}_\eta[\bar{x}] \subset \mathbb{B}_{r_2}[0]$. Then, by adding and subtracting both $\mathbf{E}_\varphi(x^{k+1})$ and $\mathbf{E}_\varphi(x^{k+2})$ in the right-hand side of (23), we get

$$\sigma\|x^{k+2} - x^{k+1}\|^2 \leq \Phi(x^{k+1}) + u_{k+1} - \left(\Phi(x^{k+2}) + u_{k+2}\right). \tag{48}$$

Now, in view of (P2), we can apply the concavity of θ and then (48) yielding

$$\begin{aligned} s_{k+1} - s_{k+2} &\geq \theta'(\Phi(x^{k+1}) - \Phi(\bar{x}) + u_{k+1})\left(\Phi(x^{k+1}) + u_{k+1} - \Phi(x^{k+2}) - u_{k+2}\right) \\ &\geq \sigma\theta'(\Phi(x^{k+1}) - \Phi(\bar{x}) + u_{k+1})\|x^{k+2} - x^{k+1}\|^2. \end{aligned}$$

Using the quasi-additive property (33), (P3) and the KL property (32), we get that

$$\begin{aligned} s_{k+1} - s_{k+2} &\geq \frac{C\sigma}{\left[\theta'(\Phi(x^{k+1}) - \Phi(\bar{x}))\right]^{-1} + \left[\theta'(u_{k+1})\right]^{-1}}\|x^{k+2} - x^{k+1}\|^2 \\ &\geq \frac{C\sigma}{\operatorname{dist}(0; \partial\,\Phi(x^{k+1})) + \left[\theta'(u_{k+1})\right]^{-1}}\|x^{k+2} - x^{k+1}\|^2. \end{aligned} \tag{49}$$

From the triangle inequality, (41) and the fact that we also have

$$\begin{aligned} \operatorname{dist}(0; \partial\,\Phi(x^{k+1})) &\leq \operatorname{dist}(0; \partial\,\Phi_k(x^{k+1})) + \operatorname{dist}(\partial\,\Phi(x^{k+1}); \partial\,\Phi_k(x^{k+1})) \\ &\leq \operatorname{dist}(0; \partial\,\Phi_k(x^{k+1})) + \left\|\nabla\mathbf{E}_\varphi(x^{k+1}) - \nabla\mathbf{E}^k_\varphi(x^{k+1})\right\| \\ &\leq \operatorname{dist}(0; \partial\,\Phi_k(x^{k+1})) + b_k. \end{aligned} \tag{50}$$

Since $\varepsilon < \eta < \widehat{r}$, then $x^k \in \mathbb{B}_\varepsilon[\bar{x}] \subseteq \mathbb{B}_{\widehat{r}}[\bar{x}]$, and thus the estimate in Lemma 4.4 holds. Therefore, combining (49) and (50) yields

$$(s_{k+1} - s_{k+2})\left(\|x^{k+1} - x^k\| + \zeta^{-1}b_k + \zeta^{-1}\left[\theta'(u_{k+1})\right]^{-1}\right) \geq \mathfrak{C}\|x^{k+2} - x^{k+1}\|^2.$$

Then it follows that

$$\begin{aligned} \mathfrak{C}\|x^{k+2} - x^{k+1}\| &\leq \sqrt{\mathfrak{C}}\sqrt{(s_{k+1} - s_{k+2})\left(\|x^{k+1} - x^k\| + \zeta^{-1}b_k + \zeta^{-1}\left[\theta'(u_{k+1})\right]^{-1}\right)} \\ &\leq \frac{(s_{k+1} - s_{k+2})}{2} + \frac{\mathfrak{C}}{2}\|x^{k+1} - x^k\| + \frac{\mathfrak{C}}{2}\zeta^{-1}b_k + \frac{\mathfrak{C}}{2}\zeta^{-1}\left[\theta'(u_{k+1})\right]^{-1}, \end{aligned}$$

which proves the claim. □

Finally, we bound the partial sums of the tail of the series defined by the step length of the sequence $(x_k)_{k\in\mathbb{N}}$, from where we will obtain global convergence in Section 4.1.2.

Lemma 4.7 *Under the assumptions of Lemma 4.5, let $k_1 > k_0$. If $x^k \in \mathbb{B}_\varepsilon[\bar{x}]$ for all $k = k_0, \dots, k_1$, then*

$$\sum_{k=k_0}^{k_1} \|x^{k+1} - x^k\| \leq 2\|x^{k_0+1} - x^{k_0}\| + \frac{s_{k_0+1}}{\mathfrak{C}} + \frac{1}{\zeta} \sum_{j=k_0+1}^{\infty} \left(b_{j-1} + [\theta'(u_j)]^{-1}\right). \tag{51}$$

Proof. Resorting to Lemma 4.6, by recursive substitution we get for all $k = k_0, \dots, k_1$,

$$\begin{aligned} \|x^{k+2} - x^{k+1}\| \leq \frac{1}{2^{k-k_0+1}}\|x^{k_0+1} - x^{k_0}\| + \frac{1}{\mathfrak{C}} \sum_{i=0}^{k-k_0} \left(\frac{1}{2}\right)^{i+1} (s_{k+1-i} - s_{k+2-i}) \\ + \zeta^{-1} \sum_{i=0}^{k-k_0} \left(\frac{1}{2}\right)^{i+1} \left(b_{k-i} + [\theta'(u_{k+1-i})]^{-1}\right). \end{aligned} \tag{52}$$

By adding (52) for $k = k_0, \dots, k_1 - 1$,

$$\begin{aligned} \sum_{k=k_0}^{k_1} \|x^{k+1} - x^k\| &= \|x^{k_0+1} - x^{k_0}\| + \sum_{k=k_0}^{k_1-1} \|x^{k+2} - x^{k+1}\| \\ &\leq 2\|x^{k_0+1} - x^{k_0}\| + \frac{1}{\mathfrak{C}} \sum_{k=k_0}^{k_1-1} \sum_{i=0}^{k-k_0} \left(\frac{1}{2}\right)^{i+1} (s_{k+1-i} - s_{k+2-i}) \\ &\quad + \zeta^{-1} \sum_{k=k_0}^{k_1-1} \sum_{i=0}^{k-k_0} \left(\frac{1}{2}\right)^{i+1} \left(b_{k-i} + [\theta'(u_{k+1-i})]^{-1}\right) \\ &= 2\|x^{k_0+1} - x^{k_0}\| + \frac{1}{\mathfrak{C}} \sum_{j=0}^{k_1-k_0-1} \frac{1}{2^{j+1}} \sum_{k=k_0+1}^{k_1-j} (s_k - s_{k+1}) \\ &\quad + \zeta^{-1} \sum_{j=0}^{k_1-k_0-1} \frac{1}{2^{j+1}} \sum_{k=k_0+1}^{k_1-j} \left(b_{k-1} + [\theta'(u_k)]^{-1}\right) \\ &\leq 2\|x^{k_0+1} - x^{k_0}\| + \frac{1}{\mathfrak{C}} s_{k_0+1} + \frac{1}{\zeta} \sum_{k=k_0+1}^{\infty} \left(b_{k-1} + [\theta'(u_k)]^{-1}\right), \end{aligned}$$

where to get the second inequality we apply the telescopic sum identity, the fact that $s_{k_1-j+1} \geq 0$ for all $j \leq k_1 - k_0 - 1$, and $\sum_{j\geq 1} 2^{-j} = 1$. This concludes the proof.

□

4.1.2 Proof of Theorem 4.2

We are now in a position to prove the main result of this section.

Proof of Theorem 4.2. It suffices to show that for all $\varepsilon > 0$ defined in Lemma 4.5, it holds that $x^k \in \mathbb{B}_\varepsilon[\bar{x}]$ for all $k \geq k_0$. We argue by induction. By assumption, the statement holds for $k = k_0$, implied by (P4), as well as for $k = k_0 + 1$, since due to (P1) and (P4),

$$\|x^{k_0+1} - \bar{x}\| \leq \|x^{k_0+1} - x^{k_0}\| + \|x^{k_0} - \bar{x}\| \leq \frac{2\varepsilon}{5}.$$

To proceed with our induction step, we assume that the claim holds for all $k \in \{k_0, \dots, k_1\}$, with $k_1 > k_0$. Let us prove that $x^{k_1+1} \in \mathbb{B}_\varepsilon[\bar{x}]$. Indeed, in view of Lemma 4.7, (P1) and (P4),

$$\sum_{k=k_0}^{k_1} \|x^{k+1} - x^k\| \leq \frac{2\varepsilon}{5} + \frac{\varepsilon}{5} + \frac{\varepsilon}{5} = \frac{4\varepsilon}{5}.$$

Calling (P4) once again and the triangle inequality, we deduce

$$\|x^{k_1+1} - \bar{x}\| \leq \|x^{k_0} - \bar{x}\| + \sum_{k=k_0}^{k_1} \|x^{k+1} - x^k\| \leq \varepsilon,$$

from where we deduce the desired result. □

Remark 4.8 We next comment on Theorem 4.2.

(i) The assumption $\Phi(x^k) > \Phi(\bar{x})$ for all large enough $k \in \mathbb{N}$ in Theorem 4.2 holds trivially whenever $\bar{x}$ is a local minimizer (unless x^k is a local minimizer itself). Alternatively, a slightly stronger variant of the KL condition [67, Definition 2.1] can be used, namely, requiring

$$\theta'\big(|\,\Phi(x) - \Phi(\bar{x})|\big)\mathrm{dist}\big(0; \partial\,\Phi(x)\big) \geq 1$$

whenever

$$x \in \mathbb{B}_\eta(\bar{x}) \text{ and } 0 < |\,\Phi(x) - \Phi(\bar{x})| < \eta.$$

Instead, we assume the condition $\Phi(x^k) > \Phi(\bar{x})$ for ease of presentation.

(ii) The proof of Theorem 4.2 uses the bound in Lemma 4.4, which is expected to hold for methods of *implicit* nature, such as proximal-type methods (see, e.g. [12,16,41]). Methods of *explicit* nature, such as the gradient method, are expected to satisfy Lemma 4.4 by evaluating the left-hand side of the estimate therein at x^k, namely,

$$\mathrm{dist}\big(0; \partial\,\Phi_k(x^k)\big) \;\leq\; \zeta\,\|x^{k+1} - x^k\|. \tag{53}$$

For instance, if $\psi \equiv 0$ and φ is smooth in ($\mathcal{P}$), then the update (15) in Algorithm 1, after passing the line search test, becomes

$$x^{k+1} = x^k - \gamma_k v^k, \text{ where } v^k = \frac{1}{n_k}\sum_{i=1}^{n_k} \nabla\varphi_{\xi_i}(x^k) = \nabla_x \mathbf{E}^k_\varphi(x^k).$$

Then (53) holds for $\zeta = (\inf_{k\in\mathbb{N}}\gamma_k)^{-1}$, which is strictly positive a.s. due to Theorem 3.5(i). Therefore, similar convergence results to Theorem 4.2, *mutatis mutandis*, follow for explicit methods satisfying a sufficient descent condition up to errors (Lemma 3.7), for nonconvex stochastic optimization problems. We refer to [9] for a unified treatment of implicit and explicit methods of descent in the nonconvex deterministic setting.

4.2 Local convergence rates of PSS

In this section we describe the rate of convergence of Algorithm 1 when the KL condition is satisfied in the exponential case. We first state the main theorem. Observe that in order to get convergence rates, on top of the KL condition, we assume that the sample sizes grow at least at a polynomial rate.

Theorem 4.9 (Convergence rates) *In the setting of Theorem 4.2, let $\bar{x}$ be the limit of the sequence $(x^k)_{k\in\mathbb{N}}$, and suppose that the KL condition holds at $\bar{x}$ with desingularizing function $\theta(t) = Mt^{1-\beta}$, for some $M > 0$ and $\beta \in (0,1)$. Furthermore, let the sequence of sample sizes $(n_k)_{k\in\mathbb{N}}$ satisfy*

$$n_k^{-1} = \mathcal{O}(k^{-\varrho}), \text{ where } \varrho \in \left(\frac{2+2\beta}{\beta}, +\infty\right). \tag{54}$$

Then, the following convergence rates hold.

(i) If $\beta \in \left(0, \frac{1}{2}\right]$, then

$$\Phi(x^k) - \Phi(\bar{x}) = \mathcal{O}\left(\frac{\ln(k)}{k^{\varrho/2-1}}\right), \text{ and } \|x^k - \bar{x}\| = \mathcal{O}\left(\frac{\ln(k)}{k^{1/2(\varrho/2-1)-1}}\right).$$

(ii) If $\beta \in \left(\frac{1}{2}, 1\right)$*, then*

$$\Phi(x^k) - \Phi(\bar{x}) = \begin{cases} \mathcal{O}\left(k^{-2\beta(\varrho/2-1)+1}\ln(k)\right), & \text{if } \frac{2+2\beta}{\beta} < \varrho \leq \frac{4\beta}{2\beta-1}, \\ \mathcal{O}\left(k^{-2\beta/(2\beta-1)+1}\ln(k)\right), & \text{if } \varrho > \frac{4\beta}{2\beta-1}, \end{cases}$$

and

$$\|x^k - \bar{x}\| = \begin{cases} \mathcal{O}\left(k^{-\beta(\varrho/2-1)+1}\ln(k)\right), & \text{if } \frac{2+2\beta}{\beta} < \varrho \leq \frac{4\beta}{2\beta-1}, \\ \mathcal{O}\left(k^{-(1-\beta)/(2\beta-1)}\ln^{1-\beta}(k)\right), & \text{if } \varrho > \frac{4\beta}{2\beta-1}. \end{cases}$$

Figure 1 shows a heat map that illustrates the convergence rates obtained in Theorem 4.9. The proof of Theorem 4.9 will be developed in Section 4.2.2 . We first present in the following section some results instrumental to it.

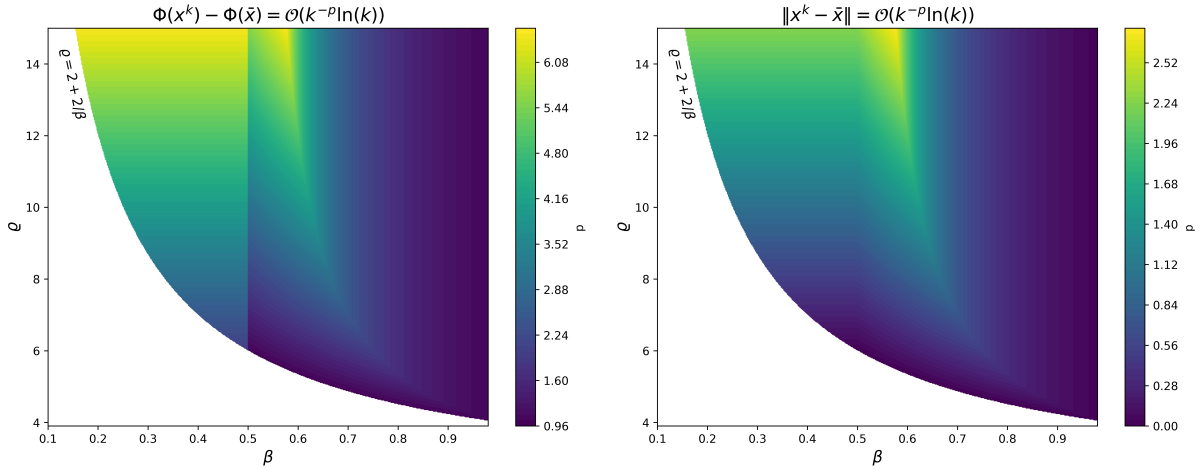


Fig. 1: Heat map of the exponent p in the convergence rate $\mathcal{O}\left(k^{-p}\ln(k)\right)$ obtained in Theorem 4.9 for the function values (left) and the iterates (right), displayed as a function of the KL exponent $\beta \in (0, 1)$ and the sample-size polynomial growth exponent $\varrho > 2 + 2/\beta$.

4.2.1 Preliminaries of Theorem 4.9

Before delving into the analysis, we note that a sequence of sample sizes $(n_k)_{k\in\mathbb{N}}$ verifying (54), satisfies conditions (34) and (35) required in Theorem 4.2. We defer the proof of this claim to Section A.2 in the Appendix, as it is merely a technicality.

Lemma 4.10 *Let* $\beta \in (0, 1)$*,* $\varrho > 0$ *and* $(n_k)_{k\in\mathbb{N}}$ *such that* (54) *holds. Then*

$$[\theta'(u_{k+1})]^{-1} = \mathcal{O}\left(\frac{\ln^{\beta}(k)}{k^{\beta(\varrho/2-1)}}\right), \quad b_k = \mathcal{O}\left(\frac{\ln(k)}{k^{\varrho/2}}\right) \text{ and } \sum_{j=k}^{\infty}(b_j + \theta'(u_{j+1})^{-1}) = \mathcal{O}\left(\frac{\ln(k)}{k^{\beta(\varrho/2-1)-1}}\right).$$

Furthermore, (34) *and* (35) *hold.*

Capitalizing on Lemma 2.12, we proceed to deduce convergence rates for Algorithm 1 for both the function values and the iterates. The arguments adapt those from the classical monotone setting to our nonmonotone case. For a different nonmonotone method, the authors in [67, Theorem 3.10] obtain similar convergence rates for the iterates alone. Observe that the rate at which the approximation error vanishes, which was given by Lemma 2.9, dominates over any possible faster linear rates, such as the classical linear [10, Theorem 2] or superlinear [16, Theorem 2].

In order to obtain the convergence rates in Theorem 4.9, we first show a recursion for the sequence of function values that complies with Lemma 2.12.

Lemma 4.11 *Suppose that the assumptions of Theorem 4.9 are satisfied. Set* $\alpha_k := \Phi(x^k)-\Phi(\bar{x})+u_k \geq 0$ *for all* $k \geq k_0$*, where* k_0 *is defined in Lemma 4.5. Then, there exist constants* $\widehat{C}_1, \widehat{C}_2 > 0$ *such that*

$$\alpha_{k+1}^{2\beta} \leq \widehat{C}_1(\alpha_k - \alpha_{k+1}) + \widehat{C}_2 \frac{\ln^{2\beta}(k)}{k^{2\beta(\varrho/2-1)}},$$

for all sufficiently large k.

Proof. From (P2) and (P3), we know that $\alpha_k, u_k, \Phi(x^k) - \Phi(\bar{x})$ belong to $(0, \eta)$ for all $k \geq k_0$, where $\eta > 0$ comes from the neighborhood of the KL condition. Furthermore, taking $\widehat{r} > 0$ for which Lemma 4.4 holds and $r_1 > 0$ such that φ is $C^{1,1}$ on $\mathbb{B}_{r_1}[\bar{x}]$ (as done in Lemma 4.5), Theorem 4.2 implies that for all $\widehat{\varepsilon} \in (0, \min\{\eta, r_1, \widehat{r}\})$, increasing k_0 if necessary, for all $k \geq k_0$, $x^k \in \mathbb{B}_{\widehat{\varepsilon}}[\bar{x}]$. Then

$$\begin{aligned}
\alpha_{k+1}^{2\beta} &= M^2(1-\beta)^2\theta'(\Phi(x^{k+1}) - \Phi(\bar{x}) + u_{k+1})^{-2} \\
&\leq M^2(1-\beta)^2\big(\theta'(\Phi(x^{k+1}) - \Phi(\bar{x}))^{-1} + [\theta'(u_{k+1})]^{-1}\big)^2 \\
&\leq M^2(1-\beta)^2\big(\operatorname{dist}\big(0; \partial\,\Phi(x^{k+1})\big) + [\theta'(u_{k+1})]^{-1}\big)^2 \\
&\leq M^2(1-\beta)^2\big(\zeta\|x^{k+1} - x^k\| + b_k + [\theta'(u_{k+1})]^{-1}\big)^2 \\
&\leq 2M^2(1-\beta)^2\big[\zeta^2\|x^{k+1} - x^k\|^2 + \big(b_k + [\theta'(u_{k+1})]^{-1}\big)^2\big] \\
&\leq 2M^2(1-\beta)^2\big[\frac{\zeta^2}{\sigma}(\Phi_k(x^k) - \Phi_k(x^{k+1})) + \big(b_k + [\theta'(u_{k+1})]^{-1}\big)^2\big] \\
&\leq 2M^2(1-\beta)^2\big[\frac{\zeta^2}{\sigma}(\alpha_k - \alpha_{k+1}) + \big(b_k + [\theta'(u_{k+1})]^{-1}\big)^2\big],
\end{aligned}$$

where in the first line we use $\theta'(t) = M(1-\beta)t^{-\beta}$, in the second line we apply the quasi-subadditivity property with $C = 1$ (see Remark 4.1), in the third line we use the KL property (32), in the fourth line we apply (50) and Lemma 4.4, in the fifth line we employ Young's inequality, in the sixth line we apply (22), and in the last line we combine (6), (39) and (40). Hence, for $\widehat{C}_1 = 2M^2(1-\beta)^2\frac{\zeta^2}{\sigma}$ and $C_2 = \widehat{C}_1\frac{\sigma}{\zeta^2}$,

$$\alpha_{k+1}^{2\beta} \leq \widehat{C}_1(\alpha_k - \alpha_{k+1}) + C_2\big(b_k + [\theta'(u_{k+1})]^{-1}\big)^2.$$

Finally, the second term on the right-hand side of the above equation is estimated by Lemma 4.10, which yields the existence of a constant $\widehat{C}_2 > 0$ such that

$$\alpha_{k+1}^{2\beta} \leq \widehat{C}_1(\alpha_k - \alpha_{k+1}) + \widehat{C}_2 \frac{\ln^{2\beta}(k)}{k^{2\beta(\varrho/2-1)}},$$

concluding the proof. □

We proceed to show a technical lemma that relates the rate of convergence of the function values and the error associated with the sample average approximation to the rate of convergence of the iterates (cf. [41, Theorem 3.3]).

Lemma 4.12 *Suppose that the assumptions of Theorem 4.9 are satisfied. Let* $\widetilde{\theta}(t) = \max\{\theta(t), \sqrt{t}\}$*, and* $(\alpha_k)_{k\in\mathbb{N}}$ *given as in Lemma 4.11. Then, for all* k *sufficiently large,*

$$\|x^k - \bar{x}\| = \mathcal{O}(\widetilde{\theta}(\alpha_k)) + \mathcal{O}\left(k^{-\beta(\varrho/2-1)+1}\ln(k)\right).$$

Proof. By rewriting (48) with $k-1$ instead of k, we have, for all sufficiently large k,

$$\sigma\|x^{k+1} - x^k\|^2 \leq \Phi(x^k) + u_k - (\Phi(x^{k+1}) + u_{k+1}) = \alpha_k - \alpha_{k+1} \leq \alpha_k. \tag{55}$$

Hence,

$$\begin{aligned}
\|\bar{x} - x^k\| &\leq \sum_{j=k}^{\infty}\|x^{j+1} - x^j\| \\
&\leq 2\|x^{k+1} - x^k\| + \frac{1}{\mathfrak{C}}\theta(\alpha_{k+1}) + \frac{1}{\zeta}\sum_{j=k+1}^{\infty}(b_{j-1} + [\theta'(u_j)]^{-1}) \\
&\leq \frac{2}{\sqrt{\sigma}}\sqrt{\alpha_k} + \frac{1}{\mathfrak{C}}\theta(\alpha_k) + \frac{1}{\zeta}\sum_{j=k}^{\infty}(b_j + [\theta'(u_{j+1})]^{-1}) \\
&\leq \mathcal{O}(\widetilde{\theta}(\alpha_k)) + \mathcal{O}\left(k^{-\beta(\varrho/2-1)+1}\ln(k)\right),
\end{aligned}$$

where in the second line we use (51), in the third line we apply (55) and the fact that θ is nondecreasing (since $\theta' > 0$), and the last line follows from the definition of $\widetilde{\theta}$ and Lemma 4.10. □

4.2.2 Proof of Theorem 4.9

We are now ready to deduce local convergence rates for both the sequence of function values and the iterates.

Proof of Theorem 4.9. We first address the convergence rates for the function values. Observe that if the KL condition (32) holds in the exponential case with $\beta \in (0, \frac{1}{2})$, then it also holds for $\beta = \frac{1}{2}$, so we concentrate on the latter case. For $\beta = \frac{1}{2}$, Lemma 4.11 yields a recursion of the type (14) with $q = p_1 = 1$ and $p_2 = (\frac{\varrho}{2} - 1)$. Then Lemma 2.12(ii) implies that $\alpha_k = \mathcal{O}\big(k^{-(\varrho/2-1)} \ln(k)\big)$, from where the first claim for the function values in item (i) follows as $u_k \geq 0$. Similarly, for $\beta \in (\frac{1}{2}, 1)$, combine Lemma 4.11 and Lemma 2.12(iii) to obtain $\alpha_k = \mathcal{O}\big(k^{-2\beta \min\{\frac{\varrho}{2}-1, \frac{1}{2\beta-1}\}+1} \ln(k)\big)$, from where the first estimate in (ii) follows as well.

As for the convergence rate of the iterates, if $\beta \in (0, \frac{1}{2}]$, for all k sufficiently large, $\widetilde{\theta}(\alpha_k) = \mathcal{O}(\sqrt{\alpha_k})$ and $k^{-1/2(\varrho/2-1)}\sqrt{\ln(k)} \leq k^{-1/2(\varrho/2-1)+1} \ln(k)$, then Lemma 4.12 implies $\|x^k - \bar{x}\| = \mathcal{O}\big(k^{-1/2(\varrho/2-1)+1} \ln(k)\big)$, showing the second rate in item (i). If $\beta \in (\frac{1}{2}, 1)$, then for all k large enough, $\widetilde{\theta}(\alpha_k) = \mathcal{O}(\alpha_k^{1-\beta})$, so that from Lemma 4.12,

$$\|x^k - \bar{x}\| = \mathcal{O}\big(k^{(1-\beta)(1-2\beta\nu)} \ln^{1-\beta}(k)\big) + \mathcal{O}\big(k^{-\beta(\varrho/2-1)+1} \ln(k)\big), \tag{56}$$

where $\nu = \min\{\frac{\varrho}{2} - 1, \frac{1}{2\beta-1}\}$. We separate the analysis in two cases.
Case 1: $\nu = \frac{\varrho}{2} - 1$ ($\frac{\varrho}{2} - 1 \leq \frac{1}{2\beta-1}$). In this case,

$$\widetilde{\nu} := (1-\beta)(-2\beta\nu + 1) + \beta\left(\frac{\varrho}{2} - 1\right) - 1 \leq 0 \implies \frac{k^{\widetilde{\nu}}}{\ln^{\beta}(k)} \to 0,$$

then for all k large enough,

$$k^{(1-\beta)(-2\beta\nu+1)} \ln^{1-\beta}(k) \leq k^{-\beta(\varrho/2-1)+1} \ln(k).$$

Hence the second term in (56) dominates for all k sufficiently large, therefore

$$\|x^k - \bar{x}\| = \mathcal{O}\left(k^{-\beta(\varrho/2-1)+1} \ln(k)\right).$$

Case 2: $\nu = \frac{1}{2\beta-1}$ ($\frac{\varrho}{2} - 1 > \frac{1}{2\beta-1}$). In this case,

$$\widetilde{\nu} := (1-\beta)(-2\beta\nu + 1) + \beta\left(\frac{\varrho}{2} - 1\right) - 1 > 0 \implies \frac{k^{\widetilde{\nu}}}{\ln^{\beta}(k)} \to +\infty,$$

then for all k large enough,

$$k^{(1-\beta)(-2\beta\nu+1)} \ln^{1-\beta}(k) \geq k^{-\beta\left(\frac{\varrho}{2}-1\right)+1} \ln(k).$$

Hence the first term in (56) dominates for all k sufficiently large, consequently

$$\|x^k - \bar{x}\| = \mathcal{O}\left(k^{(1-\beta)(-2\beta\nu+1)} \ln^{1-\beta}(k)\right),$$

from where we can conclude the proof.

□

Remark 4.13 A few comments about the convergence rates are in order.

(i) In Theorem 4.9, we skip the case $\beta = 0$ for the exponent, as the corresponding desingularizing function does not satisfy the condition in (35). For this exponent, one usually obtains termination of the algorithm in finitely many steps (see, e.g. [66, page 2]). Nonetheless, our conjecture is that such a result might not be possible to retrieve in general for our method due to the average approximation errors.

(ii) The rates in Theorem 4.9 refer to the *outer iterations* of Algorithm 1, without counting how many *inner iterations* the line search requires to produce a candidate point that passes the line search test. However, in the setting of Remark 3.10(ii), that is, when the stepsizes are nonincreasing, these rates become global.

(iii) The exponents in Theorem 4.9 reveal a sharp *transition* at $\beta = \frac{1}{2}$ and an interplay between the KL exponent β and the sample-size growth rate ϱ. For $\beta \in (0, \frac{1}{2}]$, the function-value exponent $\frac{\varrho}{2} - 1$ grows linearly and unboundedly with ϱ: a faster enrichment of the sample, which makes the approximation error vanish faster, directly translates into a faster decay of the objective gap. For $\beta \in (\frac{1}{2}, 1)$, however, the function-value exponent equals $2\beta\big(\frac{\varrho}{2} - 1\big) - 1$ only while $\frac{2+2\beta}{\beta} < \varrho \leq \frac{4\beta}{2\beta-1}$, and *saturates* at the value $\frac{1}{2\beta-1}$ once $\varrho > \frac{4\beta}{2\beta-1}$. In other words, beyond a threshold sample-size growth, the stochastic rate ceases to improve and recovers, up to the logarithmic factor, the classical deterministic rate $\mathcal{O}\big(k^{-1/(2\beta-1)}\big)$ associated with the exponential KL condition; see, e.g., [10, Theorem 2]. The same phenomenon holds for the iterates, whose exponent saturates at $\frac{1-\beta}{2\beta-1}$, matching the deterministic iterate rate $\mathcal{O}\big(k^{-(1-\beta)/(2\beta-1)}\big)$. Hence, the deterministic KL rate is the natural barrier of the method, and it is attained as soon as the sampling error decays sufficiently fast.

5 Numerical illustrations

This section provides numerical illustrations of the performance of PSS in two different applications: optimal quantization and partial-label linear regression.

5.1 Optimal quantization

The optimal quantization problem consists of approximating a (continuous) probability distribution by a discrete measure that minimizes the expected distortion. This problem has been extensively studied in multiple fields; see, for instance, [46] and the references therein. The problem can be mathematically formulated as follows. Let μ be a probability distribution in $\mathbb{R}^d$. The goal of the optimal quantization problem is to find $s \in \mathbb{N}$ points $x_1, x_2, \ldots, x_s \in \mathbb{R}^d$ solving the minimization problem

$$\min_{x_1,\ldots,x_s \in \mathbb{R}^d} \quad \int_{\mathbb{R}^d} \min_{t\in\{1,\ldots,s\}} \|x_t - \xi\|_2^2 \, d\mu(\xi).$$

The objective function of the problem above can be posed as the expectation of the upper-C^2 normal integrand given by, for $X := (x_1, x_2, \ldots, x_s) \in \mathbb{R}^{d\times s}$,

$$\varphi : \mathbb{R}^d \times \mathbb{R}^{d\times s} \to \mathbb{R}, \ (\xi, X) \mapsto \min_{t\in\{1,\ldots,s\}} \|x_t - \xi\|^2,$$

see [3, Section 4.2] for a deterministic, empirical-risk formulation of the problem.

In our experiments we introduce an additional constraint/penalization on the points X. Namely, we are interested in the problem

$$\min_{X\in\mathbb{R}^{d\times s}} \quad \mathbf{E}_\varphi(X) + \psi(X),$$

where ψ denotes a penalization term. In the following, we describe two setups that differ in the choice of ψ.

Setup 1: constrained optimal quantization. In this setting, we take ψ to be the indicator function ι_C of a nonempty closed set C, which is lsc and prox-bounded (with prox-boundedness threshold $+\infty$).

Setup 2: regularized optimal quantization. In this setting, we take

$$\psi(X) = \lambda_1 \sum_{j=1}^{s} \sum_{i<j} P(x_i - x_j),$$

where $\lambda_1 > 0$ is the level of penalization parameter, and P is a penalization term traditionally used to promote sparsity. This formulation is inspired by the models of fusion-based clustering [47], the discrete version of the optimal quantization problem. The expected effect of ψ on the solution is to induce the merge (fusion) of quantization points that are close enough to minimize the overall cost, reducing the number of effective centroids that approximate the original probability distribution.

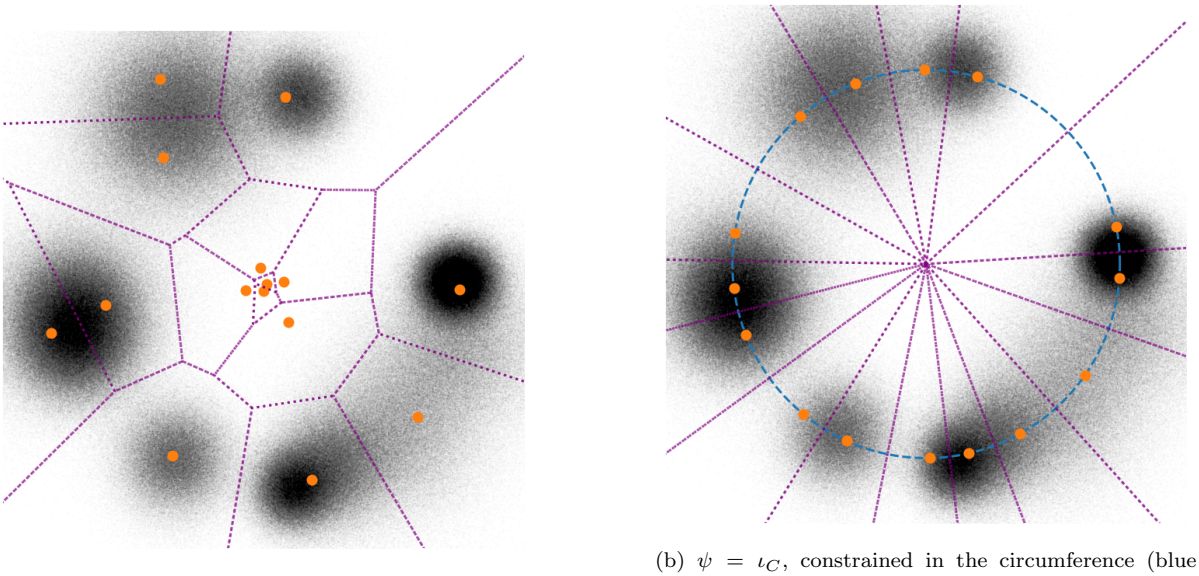
For both of the setups above, we construct the original distribution μ as a uniform mixture of ten bivariate Gaussian distributions whose parameters are generated randomly. The centers of these

distributions are expressed in polar coordinates as $\big(r\cos(\theta), r\sin(\theta)\big)$, where the angles θ and the radii r are drawn uniformly in the intervals $[0, 2\pi]$ and $[10, 11]$, respectively. For each distribution, the covariance matrix is given by $\sigma^2 I$, where σ^2 is sampled independently and uniformly from $[1, 3]$. To generate samples $\xi \sim \mu$, one of the ten Gaussian components is selected with equal probability, followed by sampling from such a distribution. We aim to discretize the original distribution with $s = 15$ points. Regarding the numerical implementation of Algorithm 1, we take $n_0 = 1000$ and define $n_k = n_{k-1} + k$, for all $k \geq 1$. Further, it is well-known that $\varphi(\xi, \cdot)$ verifies the stronger property of being 1-upper-C^2; see, e.g., [3]. Hence, one can take $\gamma_k = \widetilde{\gamma} \in (0, 1/2)$. This avoids the computational burden of performing the *Armijo-type line search* at every iteration of the method. In particular, we fixed $\widetilde{\gamma} = 0.9/2$. In our experiments, Algorithm 1 stopped and returned the k-th iterate if

$$\frac{\|X^k - X^{k-1}\|}{\max\{1, \|X^{k-1}\|\}} < 10^{-3}. \tag{57}$$

For the initialization of the algorithm, we randomly picked the initial points uniformly from the interval $[-1, 1]$ for both setups.

Figure 2 illustrates the output of Algorithm 1 of experimental Setup 1. The original probability distribution μ is visualized as a gray-scale point cloud, where darker regions correspond to areas of higher probability mass. Figure 2a shows the optimal quantization results (orange dots) with no constraints (i.e., $C = \mathbb{R}^2$) for $s = 15$ target points. The purple dotted lines represent the Voronoi partition associated with the computed quantization points. Observe that a few points try to coalesce toward the origin, suggesting that these points might be redundant for the quantization. Figure 2b displays the outcome of adding the circumference constraint $C := \{x \in \mathbb{R}^2 : \|x\| = 10.5\}$ depicted in dashed blue lines, which induces no trivial quantization points.



(a) $\psi = 0$ (unconstrained)

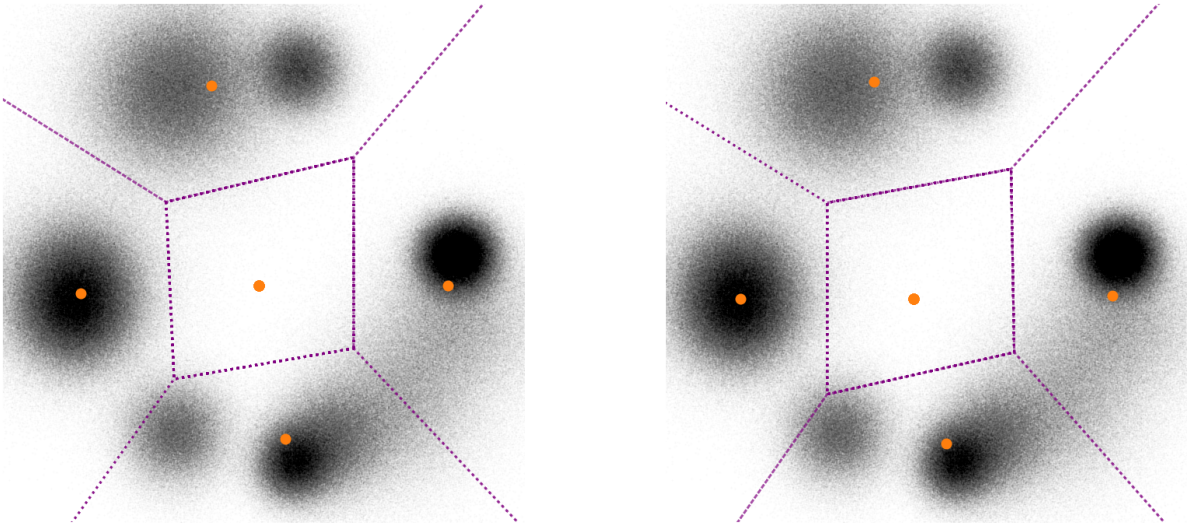
(b) $\psi = \iota_C$, constrained in the circumference (blue dashed line)

Fig. 2: Optimal quantization results (orange dots) of Setup 1 and corresponding Voronoi partition (purple dashed lines).

To promote the coalescence of quantization points observed in Figure 2a, we implement Setup 2 with the (lsc, prox-bounded) convex penalty ℓ_1 and with the nonconvex penalty $\ell_1 - \ell_2$ of [37]. For the latter, we evaluate the proximal operator only with respect to the ℓ_1-norm, since the ℓ_2-norm is twice continuously differentiable and is therefore absorbed into the upper-C^2 term φ. Figures 3 and 4 report the results obtained with the penalization parameters $\lambda_1 = 0.01$ and $\lambda_1 = 0.0025$, respectively.

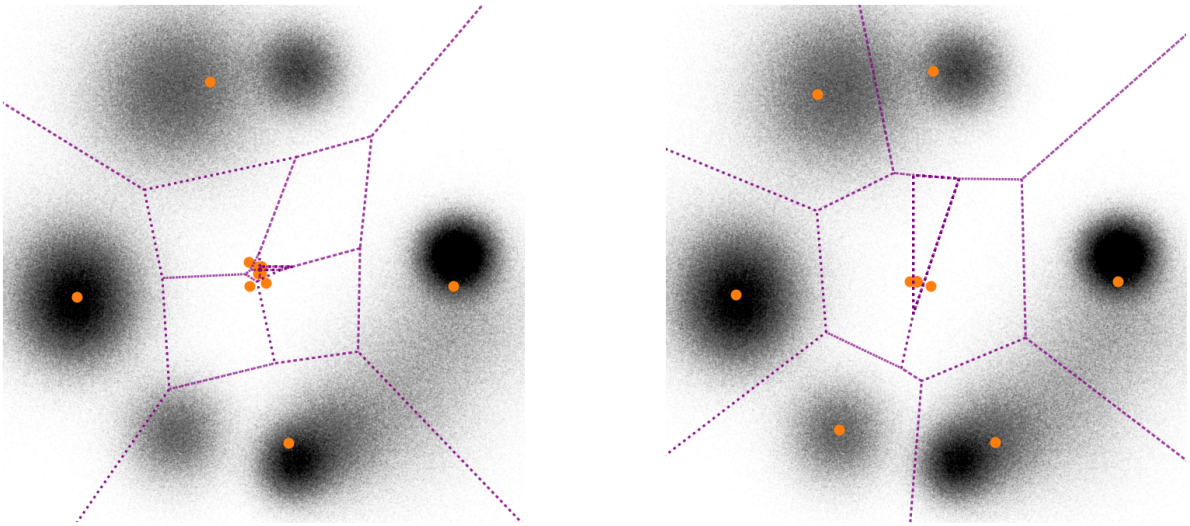
For the larger parameter $\lambda_1 = 0.01$, Figures 3a and 3b yield essentially the same quantization (though numerically distinct): the points clustered near the origin in Figure 2a merge, but the strong penalization also drives the remaining points to underrepresent the data clouds. Decreasing the penalization to $\lambda_1 = 0.0025$ produces a qualitatively different and more desirable behavior, displayed in Figures 4a and 4b. Among all displayed outcomes, the $\ell_1 - \ell_2$ penalty in Figure 4b gives the most faithful quantization,

assigning essentially one point to each data cloud, while only the redundant points near the origin are not fully fused. This illustrates the expected advantage of the nonconvex penalty in reducing the shrinking bias of the ℓ_1 regularization.



(a) ℓ_1 penalization (b) $\ell_1 - \ell_2$ penalization

Fig. 3: Optimal quantization results of Setup 2 with penalization parameter $\lambda_1 = 0.01$.



(a) ℓ_1 penalization (b) $\ell_1 - \ell_2$ penalization

Fig. 4: Optimal quantization results of Setup 2 with penalization parameter $\lambda_1 = 0.0025$.

5.2 Partial-label linear regression

Many real-world applications aim to learn information from regression models that present a set of multiple candidate labels. These labels may be produced from ambiguity and noise in data annotation tasks, and only a few of them might correspond to true labels. This scenario has driven the field of weakly supervised learning known as *partial-label learning* [30], which has found applications in face recognition systems and multimedia content analysis, among others; see, e.g., [24, 71, 91]. For regression tasks that

learn with real-valued labels, a new model has recently been proposed in [25] that minimizes the least loss incurred by candidate labels. In the context of linear regression, this model can be posed as a particular instance of $(\mathcal{P})$, which also allows including a regularizing penalty to promote sparse solutions.

The problem of interest can be formulated as follows. An agent seeks to find a sparse vector of weights $w \in \mathbb{R}^d$ from a variable environment that, given a feature vector $x \in \mathbb{R}^d$, provides multiple potential target values $y = (y_1, y_2, \ldots, y_s) \in \mathbb{R}^s$. Every y_t represents competing regimes dictated by the same underlying features. The problem can then be tackled by the stochastic program

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mu} \left[\min \left\{ \frac{1}{2}(y_t - x^T w)^2 \ : \ t = 1, \ldots, s \right\} \right] + \psi(w), \tag{58}$$

where μ is a joint probability distribution for (x, y) and ψ is a sparsity-inducing penalty. By using well-known characterizations of upper-C^2 functions (see, e.g., [4, Section 3]) it is not difficult to check that the above problem is in the form of $(\mathcal{P})$.

We present here a preliminary experiment to illustrate the performance of Algorithm 1 applied to problems in the form of (58). Specifically, we take $s = 2$, namely, only two candidate labels are given. To sample (x, y_1, y_2), we first randomly construct two sparse vectors $w_1^* \in \mathbb{R}^{50}$ and $w_2^* \in \mathbb{R}^{50}$ that represent the true weights of the regime. Then, for a feature vector x sampled from a standard normal distribution, we take $y_t = x^T w_t^* + e_t$, where the noise e_t is drawn from a normal distribution centered at the origin and with standard deviation 0.05, for $t = 1, 2$. For the regularization, in our experiments we take (1) $\psi = \lambda_1 \|\cdot\|_1$ with regularization parameter $\lambda_1 = 0.1$, and (2) $\psi = \text{MCP}$ [92]. Experiments on supervised learning tasks [7,8] indicate that the use of MCP and other nonconvex penalties can improve the quality of the solution compared to the results provided by the ℓ_1 penalization. This corresponds to reducing the shrinking bias mentioned in the introduction. The MCP is weakly convex [18], and thus (lsc) prox-bounded. Its proximity operator is usually called the *firm threshold*, its expression can be consulted, for instance, in [15]. Specifically, we consider the MCP regularizer and its proximity operator as given in [15] with $\tau = 0.1$ and $\rho = 2$.

In this setup, the implementation of our PSS method requires conducting a line search to estimate the stepsize parameter in Step 6. For it, we set the parameters $\sigma = 0.1$, $\rho = 0.5$, $\bar{\gamma}_0 = 2$ and $\bar{\gamma}_k = \max\{\gamma_{k-1}, 0.01\}$, for $k \geq 1$. Finally, we take $n_k = k + 1$, for all $k \in \mathbb{N}$. We stop the algorithm after obtaining a relative error (57) smaller than 10^{-4}.

Figure 5 presents a stem plot showcasing the comparison between the true weight vectors (w_1^* and w_2^*) and the weights obtained by PSS (red crosses) initialized at the origin for both sparse penalties. As observed, both cases approximate the weight w_1^*. Figure 6 shows a second representative instance where PSS initialized at a random vector converges to different weight vectors depending on the regularization strategy.

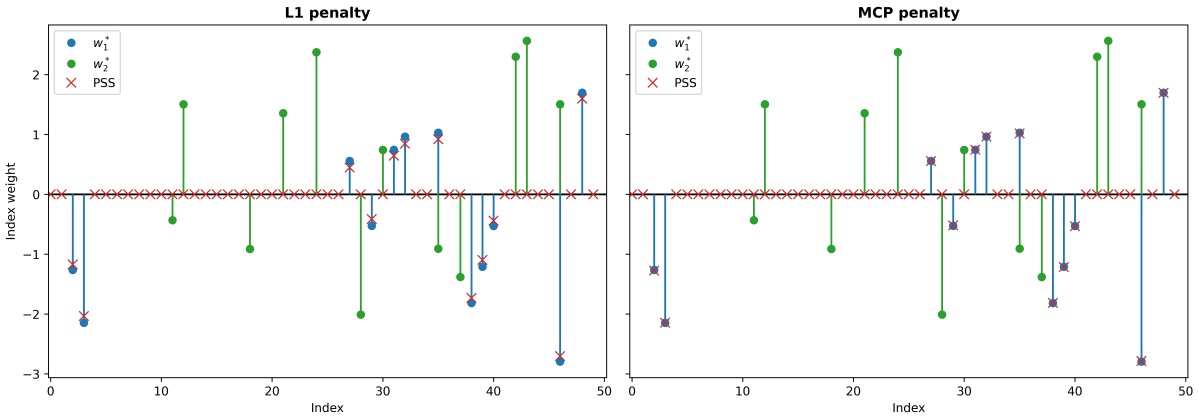


Fig. 5: Instance of problem (58) where PSS approximates the same true weight vector for the different regularization penalties.

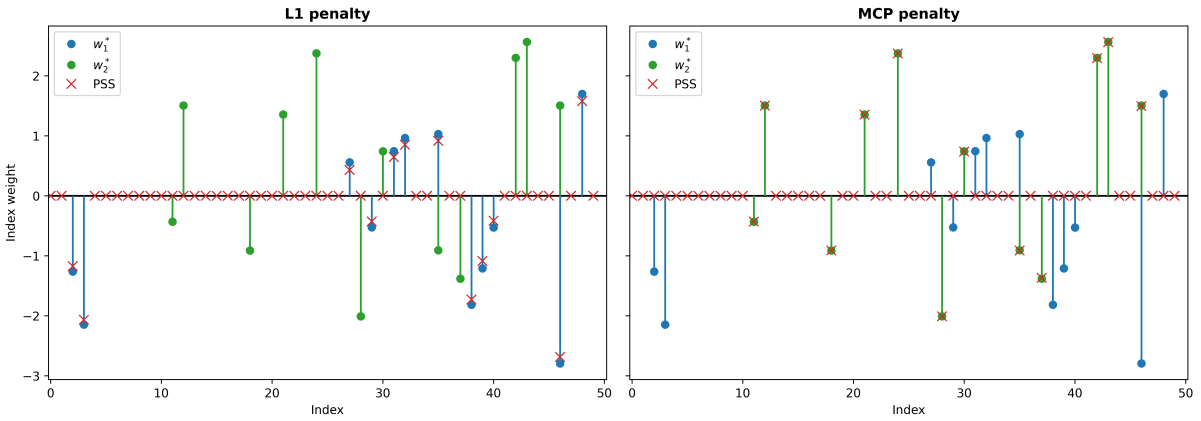


Fig. 6: Instance of problem (58) where PSS approximates different true weight vectors depending on the regularization penalty.

6 Conclusion

In this work, we developed a convergence analysis for a proximal sample-based subgradient method with an Armijo-type line search for nonsmooth and nonconvex stochastic optimization problems. The method applies to objectives given by the sum of an expected upper-C^2 normal integrand and a proper lower semicontinuous prox-bounded function, thereby covering nonsmooth losses, constraints, and possibly nonconvex regularization terms. The main idea of the analysis is to interpret the sample-based sufficient decrease condition as an inexact descent mechanism for the true objective, where the perturbation terms are induced by the empirical approximation and vanish asymptotically. This viewpoint is inspired by the deterministic convergence framework, but requires additional probabilistic estimates to handle the nonmonotonicity produced by stochastic errors.

We established almost sure convergence of both the sample-based and true objective values, square summability of the increments, and stationarity of every accumulation point of bounded trajectories. A distinctive feature of the result is that, for subsequential convergence, no prescribed growth rate is imposed on the sample-size sequence: it is enough to assume that the sequence is nondecreasing and unbounded. This flexibility is obtained through a localization argument based on stopping times, which allows the analysis to be carried out along bounded sample paths. In particular, the convergence result does not require an a priori almost sure boundedness assumption on the whole stochastic process.

Under the Kurdyka–Łojasiewicz condition, we further strengthened the subsequential convergence result to convergence of the whole trajectory to a single stationary point, provided that the sample errors satisfy suitable summability requirements. Moreover, for desingularizing functions of the form $\theta(t) = Mt^{1-\beta}$ and polynomially increasing sample sizes, we derived explicit local convergence rates for both the objective values and the iterates. The proof combines uniform convergence estimates for empirical averages, inexact descent inequalities, relative-error bounds adapted to sample-based models, and recursion estimates tailored to stochastic perturbations. These results show that the KL methodology, which is classical in deterministic nonconvex optimization, can also be effectively used in stochastic nonmonotone settings when the approximation errors are sufficiently controlled.

The numerical experiments on optimal quantization and partial-label learning illustrate the applicability of the proposed method to relevant nonsmooth stochastic models. They also suggest that the line search mechanism can provide a practical way to balance descent and statistical accuracy without enforcing a rigid sample-size schedule.

Several questions remain open for future research. On the theoretical side, it would be interesting to refine the stopping-time localization technique in order to weaken standard assumptions commonly imposed in stochastic optimization, such as uniform bounded variance or global uniform smoothness. Other natural directions to investigate are alternative sample-size regimes, adaptive sampling strategies, and variance-reduction mechanisms within the present nonsmooth nonconvex function setting and the extension of the variational inequality framework developed in previous works by Iusem, Jofre, Oliveira and Thompson [49–51]. In addition, Remark 4.8(ii) suggests the possibility of developing an abstract convergence theory for explicit and implicit stochastic methods satisfying descent-type conditions with controlled errors, in the spirit of deterministic KL-based frameworks. On the numerical side, further

experiments are needed to assess the practical performance of the method on large-scale learning problems and to compare different strategies for choosing the sample sizes and the initial trial stepsizes in the line search procedure.

Statements and Declarations

Competing interests The authors declare that they have no competing interests.

References

1. Absil, P.-A., Mahony, R., and Andrews, B. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization 16*, 2 (2005), 531–547.
2. Aragón Artacho, F. J., Fleming, R. M., and Vuong, P. T. Accelerating the DC algorithm for smooth functions. *Math. Program. 169*, 1 (2018), 95–118.
3. Aragón-Artacho, F. J., Pérez-Aros, P., and Torregrosa-Belén, D. The Boosted Double-proximal Subgradient Algorithm for nonconvex optimization. *Math. Program. 214*, 1-2, Ser. A (2025), 491–537.
4. Aragón-Artacho, F. J., Campoy, R., Pérez-Aros, P., and Torregrosa-Belén, D. Nonmonotone subgradient methods based on a local descent lemma. Preprint, available at: arXiv:2510.19341.
5. Armijo, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific J. Math. 16* (1966), 1–3.
6. Artstein, Z., and Wets, R. J.-B. Consistency of minimizers and the SLLN for stochastic programs. *J. Convex Anal. 2*, 1-2 (1995), 1–17.
7. Atenas, F. Shadow splitting methods for nonconvex optimisation: epi-approximation, convergence and saddle point avoidance. Preprint, available at: arXiv:2512.20433.
8. Atenas, F. Understanding the Douglas–Rachford splitting method through the lenses of Moreau-type envelopes. *Comput. Optim. Appl. 90*, 3 (2025), 881–910.
9. Atenas, F., Sagastizábal, C., Silva, P. J., and Solodov, M. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM J. Optim. 33*, 1 (2023), 89–115.
10. Attouch, H., and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program. 116*, 1–2 (2009), 5–16.
11. Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality. *Math. Oper. Res. 35*, 2 (2010), 438–457.
12. Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Math. Program. 137*, 1–2 (2013), 91–129.
13. Bach, F. *Learning theory from first principles*. MIT press, 2024.
14. Baldi, P. *Probability—an introduction through theory and exercises*. Universitext. Springer, Cham, [2023] ©2023.
15. Bayram, İ. On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Trans. Signal Process. 64*, 6 (2016), 1597–1608.
16. Bento, G., Mordukhovich, B., Mota, T., and Nesterov, Y. Convergence of descent optimization algorithms under Polyak-Łojasiewicz-Kurdyka conditions. *J. Optim. Theory Appl. 207*, 3 (2025), 41.
17. Bogachev, V. I. *Measure theory. Vol. I, II*. Springer-Verlag, Berlin, 2007.
18. Böhm, A., and Wright, S. J. Variable smoothing for weakly convex composite functions. *J. Optim. Theory Appl. 188*, 3 (2021), 628–649.
19. Bolte, J., Daniilidis, A., and Lewis, A. A nonsmooth Morse–Sard theorem for subanalytic functions. *J. Math. Anal. Appl. 321*, 2 (2006), 729–740.
20. Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim. 17*, 4 (2007), 1205–1223.
21. Bolte, J., Daniilidis, A., Lewis, A., and Shiota, M. Clarke subgradients of stratifiable functions. *SIAM J. Optim. 18*, 2 (2007), 556–572.
22. Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program. 146*, 1–2 (2014), 459–494.
23. Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review 60*, 2 (2018), 223–311.
24. Chen, C.-H., Patel, V. M., and Chellappa, R. Learning from ambiguously labeled face images. *IEEE Trans. Pattern Anal. Mach. Intell. 40*, 7 (2018), 1653–1667.
25. Cheng, X., Wang, D.-B., Feng, L., Zhang, M.-L., and An, B. Partial-label regression. *Proc. AAAI Conf. Artif. Intell. 37*, 6 (Jun. 2023), 7140–7147.
26. Chung, K. L. On a stochastic approximation method. *Ann. Math. Statistics 25* (1954), 463–483.
27. Clarke, F. H. *Optimization and nonsmooth analysis*, second ed., vol. 5 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1990.
28. Clarke, F. H., Stern, R. J., and Wolenski, P. R. Proximal smoothness and the lower-C^2 property. *J. Convex Anal. 2*, 1-2 (1995), 117–144.
29. Correa, R., Hantoute, A., and Pérez-Aros, P. Qualification conditions-free characterizations of the ε-subdifferential of convex integral functions. *Appl. Math. Optim. 83*, 3 (2021), 1709–1737.
30. Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *J. Mach. Learn. Res.h 12*, 42 (2011), 1501–1536.
31. Davis, D., and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim. 29*, 1 (2019), 207–239.

32. Davis, D., and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM J. Optim. 29*, 3 (2019), 1908–1930.
33. Duchi, J. C., and Ruan, F. Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim. 28*, 4 (2018), 3229–3259.
34. Dudley, R. M. *Real analysis and probability.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.
35. Ermoliev, Y. M., and Norkin, V. I. Normalized convergence of random variables and its applications. *Kibernetika (Kiev)*, 6 (1990), 85–89, 134.
36. Ermoliev, Y. M., and Norkin, V. I. Sample average approximation method for compound stochastic optimization problems. *SIAM J. Optim. 23*, 4 (2013), 2231–2263.
37. Esser, E., Lou, Y., and Xin, J. A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM J. Imaging Sci. 6*, 4 (2013), 2010–2046.
38. Facchinei, F., and Pang, J.-S. *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II.* Springer-Verlag, New York, 2003.
39. Fan, J., and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*, 456 (2001), 1348–1360.
40. Fatkhullin, I., He, N., and Hu, Y. Stochastic optimization under hidden convexity. *SIAM J. Optim. 35*, 4 (2025), 2544–2571.
41. Frankel, P., Garrigos, G., and Peypouquet, J. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *J. Optim. Theory Appl. 165*, 3 (2015), 874–900.
42. Geiersbach, C., and Scarinci, T. Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces. *Comput. Optim. Appl. 78*, 3 (2021), 705–740.
43. Ghadimi, S., and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim. 23*, 4 (2013), 2341–2368.
44. Ghadimi, S., and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program. 156*, 1-2 (2016), 59–99.
45. Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program. 155*, 1-2 (2016), 267–305.
46. Graf, S., and Luschgy, H. *Foundations of quantization for probability distributions*, vol. 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.
47. Hocking, T. D., Joulin, A., Bach, F., and Vert, J.-P. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th ICML* (2011), p. 1.
48. Homem-de Mello, T. On rates of convergence for stochastic optimization problems under non-IID sampling. *SIAM J. Optim. 19*, 2 (2008), 524–551.
49. Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization 27*, 2 (2017), 686–724.
50. Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM Journal on Optimization 29*, 1 (2019), 175–206.
51. Iusem, A. N., Jofré, A., and Thompson, P. Incremental constraint projection methods for monotone stochastic variational inequalities. *Mathematics of Operations Research 44*, 1 (2019), 236–263.
52. Izmailov, A. F., and Solodov, M. V. *Newton-Type Methods for Optimization and Variational Problems.* Springer, Cham, 2014.
53. Jia, Z., and Grimmer, B. First-order methods for nonsmooth nonconvex functional constrained optimization with or without slater points. *SIAM J. Optim. 35*, 2 (2025), 1300–1329.
54. Jofré, A., and Thompson, P. On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Math. Program. 174*, 1-2 (2019), 253–292.
55. Kanzow, C., and Lehmann, L. A nonmonotone descent method for optimization problems defined by upper-$\mathcal{C}^2$ functions over submanifolds. Preprint, available at: arXiv:2605.26909.
56. Khanh, P., Luong, H.-C., Mordukhovich, B., and Tran, D. Fundamental convergence analysis of sharpness-aware minimization. *NeurIPS 37* (2024), 13149–13182.
57. Kiefer, J., and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist. 23*, 3 (1952), 462–466.
58. King, A. J., and Rockafellar, R. T. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res. 18*, 1 (1993), 148–162.
59. Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim. 12*, 2 (2002), 479–502.
60. Krätschmer, V. Nonasymptotic upper estimates for errors of the sample average approximation method to solve risk-averse stochastic programs. *SIAM J. Optim. 34*, 2 (2024), 1264–1294.
61. Kurdyka, K. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble) 48*, 3 (1998), 769–783.
62. Lan, G. An optimal method for stochastic composite optimization. *Math. Program. 133*, 1-2, Ser. A (2012), 365–397.
63. Le Thi, H. A., Huynh, V. N., Pham Dinh, T., and Luu, H. P. H. Stochastic difference-of-convex-functions algorithms for nonconvex programming. *SIAM J. Optim. 32*, 3 (2022), 2263–2293.
64. Lei, J., and Shanbhag, U. V. Asynchronous variance-reduced block schemes for composite non-convex stochastic optimization: block-specific steplengths and adapted batch-sizes. *Optim. Methods Softw. 37*, 1 (2022), 264–294.
65. Lei, J., and Shanbhag, U. V. Variance-reduced accelerated first-order methods: central limit theorems and confidence statements. *Math. Oper. Res. 50*, 2 (2025), 1364–1397.
66. Li, G., and Pong, T. K. Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math. 18*, 5 (2018), 1199–1232.
67. Li, X., Milzarek, A., and Qiu, J. Convergence of random reshuffling under the Kurdyka–Łojasiewicz inequality. *SIAM J. Optim. 33*, 2 (2023), 1092–1120.
68. Li, Z. Simple and optimal stochastic gradient methods for nonsmooth nonconvex optimization. *J. Mach. Learn. Res 23*, 239 (2022), 1–61.

69. Loh, P.-L., and Wainwright, M. J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res 16* (2015), 559–616.
70. Łojasiewicz, S. *Ensembles semi-analytiques*. Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette (Seine-et-Oise), France, 1965.
71. Luo, J., and Orabona, F. Learning from candidate labeling sets. In *Adv. Neural Inf. Process. Syst.* (2010), J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23, Curran Associates, Inc.
72. Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res 11* (2010), 19–60.
73. Metel, M. R., and Takeda, A. Stochastic proximal methods for non-smooth non-convex constrained sparse optimization. *J. Mach. Learn. Res 22*, 115 (2021), 1–36.
74. Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim. 19*, 4 (2009), 1574–1609.
75. Nguyen, L. M., Scheinberg, K., and Tran, T. H. Stochastic ISTA/FISTA adaptive step search algorithms for convex composite optimization. *J. Optim. Theory Appl. 205*, 1 (2025), 10.
76. Paquette, C., and Scheinberg, K. A stochastic line search method with expected complexity analysis. *SIAM J. Optim. 30*, 1 (2020), 349–376.
77. Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *J. Mach. Learn. Res. 21*, 110 (2020), 1–48.
78. Polyak, B. Gradient methods for the minimisation of functionals. *USSR Comput. Math. & Math. Phys. 3*, 4 (1963), 864–878.
79. Polyak, B. T. *Introduction to optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.
80. Polyak, B. T., and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim. 30*, 4 (1992), 838–855.
81. Pougkakiotis, S., and Kalogerias, D. A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization. *SIAM J. Sci. Comput. 45*, 5 (2023), A2679–A2702.
82. Pérez-Aros, P., and Torregrosa-Belén, D. Randomized block proximal method with locally lipschitz continuous gradient. Preprint, available at: arXiv:2504.11410.
83. Robbins, H., and Monro, S. A stochastic approximation method. *Ann. Math. Statist. 22*, 3 (1951), 400–407.
84. Robbins, H., and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)* (1971), Academic Press, New York-London, pp. 233–257.
85. Rockafellar, R. T., and Wets, R. J.-B. *Variational analysis*, vol. 317 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.
86. Royset, J. O., and Wets, R. J.-B. *An optimization primer*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, [2021] ©2021.
87. Shapiro, A. Asymptotic behavior of optimal solutions in stochastic programming. *Math. Oper. Res. 18*, 4 (1993), 829–845.
88. Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on stochastic programming—modeling and theory*, third ed., vol. 28 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, [2021] ©2021.
89. Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. A hybrid stochastic optimization framework for composite nonconvex optimization. *Math. Program. 191*, 2 (2022), 1005–1071.
90. van Ackooij, W. S., and de Oliveira, W. L. Methods of nonsmooth optimization in stochastic programming. *International Series in Operations Research and Management Science* (2025).
91. Zeng, Z., Xiao, S., Jia, K., Chan, T.-H., Gao, S., Xu, D., and Ma, Y. Learning by associating ambiguously labeled images. In *2013 IEEE Conf. Comput. Vis. Pattern Recognit.* (2013), pp. 708–715.
92. Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*, 2 (2010), 894–942.
93. Zhang, J., and Xiao, L. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Math. Program. 195*, 1 (2022), 649–691.

A Appendix

A.1 Proof of Lemma 2.11

Proof of Lemma 2.11. Since $0 < n_k \leq n_{k+1}$, then $n_k\sqrt{n_{k+1}} \leq \sqrt{n_k}n_{k+1}$, and thus

$$\frac{n_{k+1}-n_k}{\sqrt{n_k}n_{k+1}} \geq \frac{n_{k+1}-n_k}{\sqrt{n_k}n_{k+1}+n_k\sqrt{n_{k+1}}} \geq \frac{n_{k+1}-n_k}{2\sqrt{n_k}n_{k+1}}.$$

For all $k \in \mathbb{N}$, let

$$p_k = \frac{n_{k+1}-n_k}{\sqrt{n_k}n_{k+1}} \text{ and } q_k = \frac{n_{k+1}-n_k}{\sqrt{n_k}n_{k+1}+n_k\sqrt{n_{k+1}}}.$$

The estimate above then yields $q_k \leq p_k \leq 2q_k$. Therefore, it suffices to verify that $\sum_{k\in\mathbb{N}} q_k$ converges for $\sum_{k\in\mathbb{N}} p_k$ to converge. Note that

$$q_k = \frac{n_{k+1}-n_k}{\sqrt{n_k}\sqrt{n_{k+1}}(\sqrt{n_{k+1}}+\sqrt{n_k})} = \frac{\sqrt{n_{k+1}}-\sqrt{n_k}}{\sqrt{n_k}\sqrt{n_{k+1}}} = \frac{1}{\sqrt{n_k}} - \frac{1}{\sqrt{n_{k+1}}}.$$

Since $n_{k+1} \to +\infty$, then

$$\sum_{k\in\mathbb{N}} q_k = \frac{1}{\sqrt{n_1}},$$

and thus $\sum_{k\in\mathbb{N}} p_k < +\infty$. □

A.2 Technical result for Theorem 4.9

Proof of Lemma 4.10. We first show that for all $k \geq 3$,

$$\sum_{t=k+1}^{\infty} \frac{\ln(n_t)}{\sqrt{n_t}} = \mathcal{O}\left(\frac{\ln(k)}{k^{\varrho/2-1}}\right).$$

By (54), there exists $C_1 > 0$ such that for all $k \in \mathbb{N}$, $n_k \geq C_1 k^{\varrho}$. Then, the result for b_k follows directly by recalling (45) and that $t \in (e^{2/\varrho}, +\infty) \mapsto \frac{\ln(t)}{t^{\varrho/2}}$ is nonincreasing. Using again that the latter function is nonincreasing and the integration by parts formula, it follows that for all $k > e$ ($> e^{2/\varrho}$),

$$\begin{aligned}\sum_{t=k+1}^{\infty} \frac{\ln(n_t)}{\sqrt{n_t}} &\leq \sum_{t=k+1}^{\infty} \frac{\ln(C_1) + \gamma \ln(t)}{\sqrt{C_1} t^{\varrho/2}} \leq \int_k^{\infty} \frac{\max\{\ln(C_1), 0\} + \varrho \ln(t)}{\sqrt{C_1} t^{\varrho/2}} dt \\ &= \frac{1}{\sqrt{C_1}} \left[\frac{\max\{\ln(C_1), 0\}}{(\frac{\varrho}{2}-1)k^{\varrho/2-1}} + \frac{\varrho}{\frac{\varrho}{2}-1}\left(\frac{\ln(k)}{k^{\varrho/2-1}} + \frac{1}{\frac{\varrho}{2}-1}\frac{1}{k^{\varrho/2-1}}\right)\right] \\ &\leq C_{\varrho} \frac{\ln(k)}{k^{\varrho/2-1}},\end{aligned}$$

with $C_{\varrho} = \frac{1}{\sqrt{C_1}}\left(\frac{\max\{\ln(C_1),0\}}{\frac{\varrho}{2}-1} + \frac{\varrho^2}{2(\frac{\varrho}{2}-1)^2}\right)$, from where the first claim follows, as well as (34). This same estimate and the fact that θ' is nonincreasing imply

$$\left[\theta'\left(\sum_{t=k+1}^{\infty} \frac{\ln(n_t)}{\sqrt{n_t}}\right)\right]^{-1} \leq \left[\theta'\left(C_{\varrho}\frac{\ln(k)}{k^{\varrho/2-1}}\right)\right]^{-1} = \frac{C_{\varrho}^{\beta}}{M(1-\beta)} \frac{\ln^{\beta}(k)}{k^{\beta(\varrho/2-1)}},$$

for sufficiently large k. Combining this last inequality with (40) and (44) yields the claim for $[\theta'(u_{k+1})]^{-1}$. Next, increasing p_0 if necessary, it holds

$$\sum_{k=p_0}^{\infty}\left[\theta'\left(\sum_{t=k}^{\infty} \frac{\ln(n_t)}{\sqrt{n_t}}\right)\right]^{-1} \leq \frac{C_{\varrho}^{\beta}}{M(1-\beta)} \sum_{k=p_0}^{\infty} \frac{\ln^{\beta}(k-1)}{(k-1)^{\beta(\varrho/2-1)}}.$$

Since $\beta(\frac{\varrho}{2}-1) > 1$, then the series converges, and (35) holds.

Finally, similarly to above, using that $t \in (e^{\beta^{-1}(\varrho/2-1)^{-1}}, +\infty) \mapsto \frac{\ln(t)}{t^{\beta(\varrho/2-1)}}$ is nonincreasing and integration by parts to bound the series,

$$\begin{aligned}\sum_{j=k}^{\infty}(b_j + \theta'(u_{j+1})^{-1}) &= \mathcal{O}\left(\sum_{j=k}^{\infty} \frac{\ln^{\beta}(j)}{j^{\beta(\varrho/2-1)}}\right) = \mathcal{O}\left(\sum_{j=k}^{\infty} \frac{\ln(j)}{j^{\beta(\varrho/2-1)}}\right) \\ &= \mathcal{O}\left(\int_{k-1}^{\infty} \frac{\ln(t)}{t^{\beta(\varrho/2-1)}} dt\right) = \mathcal{O}\left(\frac{\ln(k-1)}{(k-1)^{\beta(\varrho/2-1)-1}}\right),\end{aligned}$$

proving the last claim in (ii), since $\frac{\ln(k-1)}{(k-1)^{\beta(\varrho/2-1)-1}} \sim \frac{\ln(k)}{k^{\beta(\varrho/2-1)-1}}$ asymptotically. □

A.3 Proof of Lemma 2.12

In order to prove Lemma 2.12, we require two preliminary results, namely, Lemmas A.1 and A.2.

Lemma A.1 *Let $c > 0$, $p \in \mathbb{R}$, $\varrho_1 > 0$, and $\varrho_2 < 3\varrho_1 + 1$. For k large enough, define*

$$a_k := \left(1 + \frac{1}{k}\right)^p \frac{\ln^{\varrho_2}(k)}{\ln^{\varrho_1}(k+1)\left(c\ln^{\varrho_1}(k) - p\right)}.$$

Then

$$\lim_{k\to\infty} \frac{k}{c\ln^{\varrho_1}(k) - p}(a_k - a_{k+1}) = 0.$$

Proof. Define, for x large enough,

$$g(x) := \left(1 + \frac{1}{x}\right)^p \frac{\ln^{\varrho_2}(x)}{\ln^{\varrho_1}(x+1)\left(c\ln^{\varrho_1}(x) - p\right)}.$$

Then $a_k = g(k)$. Since $\varrho_1 > 0$ and $c > 0$, we have

$$|g(x)| = \mathcal{O}\left(\ln^{\varrho_2 - 2\varrho_1}(x)\right).$$

Moreover, differentiating logarithmically gives

$$\frac{g'(x)}{g(x)} = -\frac{p}{x^2\left(1+\frac{1}{x}\right)} + \frac{\varrho_2}{x\ln(x)} - \frac{\varrho_1}{(x+1)\ln(x+1)} - \frac{c\varrho_1 \ln^{\varrho_1 - 1}(x)}{x\left(c\ln^{\varrho_1}(x) - p\right)}.$$

Hence,

$$\frac{g'(x)}{g(x)} = \mathcal{O}\left(\frac{1}{x\ln(x)}\right),$$

and consequently

$$|g'(x)| = \mathcal{O}\left(\frac{\ln^{\varrho_2-2\varrho_1-1}(x)}{x}\right).$$

Therefore, there exist constants $c_0 > 0$ and $x_0 > 0$ such that

$$|g'(x)| \leq c_0\psi_0(x), \qquad x \geq x_0,$$

where

$$\psi_0(x) := \frac{\ln^{\varrho_2-2\varrho_1-1}(x)}{x}.$$

Since ψ_0 is nonincreasing for all sufficiently large x, we may enlarge x_0 if necessary so that ψ_0 is nonincreasing on $[x_0, \infty)$.

By the mean value theorem, for every k large enough, there exists $\xi_k \in (k, k+1)$ such that

$$a_k - a_{k+1} = g(k) - g(k+1) = -g'(\xi_k).$$

Thus, since $\xi_k \geq k$ and ψ_0 is nonincreasing,

$$|a_k - a_{k+1}| \leq c_0\psi_0(\xi_k) \leq c_0\psi_0(k) = c_0\frac{\ln^{\varrho_2-2\varrho_1-1}(k)}{k}.$$

It follows that

$$\left|\frac{k}{c\ln^{\varrho_1}(k) - p}(a_k - a_{k+1})\right| \leq \frac{c_0\ln^{\varrho_2-2\varrho_1-1}(k)}{c\ln^{\varrho_1}(k) - p}.$$

Since $c\ln^{\varrho_1}(k) - p \sim c\ln^{\varrho_1}(k)$, we obtain

$$\left|\frac{k}{c\ln^{\varrho_1}(k) - p}(a_k - a_{k+1})\right| = \mathcal{O}\left(\ln^{\varrho_2-3\varrho_1-1}(k)\right).$$

Finally, $\varrho_2 < 3\varrho_1 + 1$ implies $\varrho_2 - 3\varrho_1 - 1 < 0$. Hence

$$\ln^{\varrho_2-3\varrho_1-1}(k) \to 0,$$

and therefore

$$\lim_{k\to\infty}\frac{k}{c\ln^{\varrho_1}(k) - p}(a_k - a_{k+1}) = 0,$$

which concludes the proof. □

Lemma A.2 *Let $(u_k)_{k\geq 2}$ be a nonnegative sequence, let $\varrho_1 > 0$, and $\varrho_2 < 3\varrho_1 + 1$, and assume that*

$$u_{k+1} \leq \left(1 - c\frac{\ln^{\varrho_1}(k)}{k}\right)u_k + d\frac{\ln^{\varrho_2}(k)}{k^{p+1}}, \qquad d > 0, \quad p > 0, \quad c > 0. \tag{59}$$

Then

$$u_k = \mathcal{O}\left(\frac{\ln^{\varrho_2-\varrho_1}(k)}{k^p}\right) + o\left(\frac{\ln^{\varrho_1}(k)}{k^p}\right). \tag{60}$$

Proof. Assume that, (59) holds for every $k \geq 1$. Define

$$w_k := \frac{k^p}{\ln^{\varrho_1}(k)}u_k, \qquad k \geq 2.$$

Multiplying the recurrence by $(k+1)^p/\ln^{\varrho_1}(k+1)$, we obtain

$$w_{k+1} \leq \left(1 + \frac{1}{k}\right)^p \frac{\ln^{\varrho_1}(k)}{\ln^{\varrho_1}(k+1)}\left(1 - c\frac{\ln^{\varrho_1}(k)}{k}\right)w_k + d\left(1 + \frac{1}{k}\right)^p \frac{\ln^{\varrho_2}(k)}{k\ln^{\varrho_1}(k+1)}.$$

Since $\ln(k)/\ln(k+1) \leq 1$, for large enough $k \in \mathbb{N}$ we have

$$\left(1 + \frac{1}{k}\right)^p \frac{\ln^{\varrho_1}(k)}{\ln^{\varrho_1}(k+1)}\left(1 - c\frac{\ln^{\varrho_1}(k)}{k}\right) \leq \left(1 + \frac{1}{k}\right)^p\left(1 - c\frac{\ln^{\varrho_1}(k)}{k}\right).$$

Moreover,

$$\left(1 + \frac{1}{k}\right)^p = 1 + \frac{p}{k} + \mathcal{O}\left(\frac{1}{k^2}\right).$$

Thus,

$$w_{k+1} \leq \left(1 - \frac{c\ln^{\varrho_1}(k) - p}{k} + \mathcal{O}\left(\frac{\ln^{\varrho_1}(k)}{k^2}\right)\right)w_k + d\left(1 + \frac{1}{k}\right)^p \frac{\ln^{\varrho_2}(k)}{k\ln^{\varrho_1}(k+1)}.$$

Now define

$$v_k := w_k - a_k,$$

where

$$a_k := d\left(1+\frac{1}{k}\right)^p \frac{\ln^{\varrho_2}(k)}{\ln^{\varrho_1}(k+1)\left(c\ln^{\varrho_1}(k)-p\right)}.$$

For k large enough, $c\ln^{\varrho_1}(k)-p>0$, so a_k is well defined. This choice gives

$$\frac{c\ln^{\varrho_1}(k)-p}{k}a_k = d\left(1+\frac{1}{k}\right)^p \frac{\ln^{\varrho_2}(k)}{k\ln^{\varrho_1}(k+1)}.$$

Using $w_k = v_k + a_k$, we get

$$\begin{aligned} v_{k+1} &= w_{k+1} - a_{k+1} \\ &\le \left(1 - \frac{c\ln^{\varrho_1}(k)-p}{k} + \mathcal{O}\left(\frac{\ln^{\varrho_1}(k)}{k^2}\right)\right)(v_k + a_k) \\ &\quad + \frac{c\ln^{\varrho_1}(k)-p}{k}a_k - a_{k+1} \\ &= \left(1 - \frac{c\ln^{\varrho_1}(k)-p}{k} + \mathcal{O}\left(\frac{\ln^{\varrho_1}(k)}{k^2}\right)\right)v_k + a_k - a_{k+1} \\ &\quad + \mathcal{O}\left(\frac{\ln^{\varrho_1}(k)}{k^2}\right)a_k. \end{aligned}$$

Since

$$a_k = \mathcal{O}\left(\ln^{\varrho_2-2\varrho_1}(k)\right),$$

we have

$$\mathcal{O}\left(\frac{\ln^{\varrho_1}(k)}{k^2}\right)a_k = \mathcal{O}\left(\frac{\ln^{\varrho_2-\varrho_1}(k)}{k^2}\right).$$

Therefore,

$$v_{k+1} \le \left(1 - \frac{c\ln^{\varrho_1}(k)-p}{k} + \mathcal{O}\left(\frac{\ln^{\varrho_1}(k)}{k^2}\right)\right)v_k + a_k - a_{k+1} + \mathcal{O}\left(\frac{\ln^{\varrho_2-\varrho_1}(k)}{k^2}\right).$$

Finally, by Lemma A.1 we have

$$\lim_{k\to\infty} \frac{k}{c\ln^{\varrho_1}(k)-p}(a_k - a_{k+1}) = 0,$$

so we can apply [79, Lemma 3, p. 45] to conclude that

$$\limsup_{k\to\infty} v_k \le 0,$$

which shows that (60) holds. □

We now use Lemmas A.1 and A.2 to prove Lemma 2.12.

Proof of Lemma 2.12. First, let us simply assume that

$$w_k \le c_0 \frac{\ln^{p_1}(k)}{k^{p_2}}, \quad \text{for all } k \ge 2.$$

We prove (i) first. Let $b \in (0,1)$ be arbitrary. Since $q \in (0,1)$ and $\alpha_k \to 0^+$ we have

$$\alpha_k^{q-1} \to +\infty.$$

Therefore, there exists $k_0 \in \mathbb{N}$ such that for all $k \ge k_0$,

$$\alpha_{k+1} \le b\left(\alpha_k + \frac{\ln^{p_1}(k)}{k^{p_2}}\right).$$

Fix $k \ge k_0$. Iterating the previous inequality from k_0 up to $k-1$, we obtain

$$\begin{aligned} \alpha_k &\le b^{k-k_0}\alpha_{k_0} + \sum_{j=k_0}^{k-1} b^{k-j}\frac{\ln^{p_1}(j)}{j^p} = b^{k-k_0}\alpha_{k_0} + \sum_{i=1}^{k-k_0} b^i \frac{\ln^{p_1}(k-i)}{(k-i)^{p_2}} \\ &\le b^{k-k_0}\alpha_{k_0} + \sum_{i=1}^{k-1} b^i \frac{\ln^{p_1}(k-i)}{(k-i)^{p_2}}, \end{aligned}$$

where in the first equality we set $i = k-j$, and in the second inequality we extend the sum up to $k-1$, since the summands are nonnegative. Now, assume $k \ge 2$ and split the sum as

$$\sum_{i=1}^{k-1} b^i \frac{\ln^{p_1}(k-i)}{(k-i)^{p_2}} = \sum_{i=1}^{\lfloor k/2\rfloor} b^i \frac{\ln^{p_1}(k-i)}{(k-i)^{p_2}} + \sum_{i=\lfloor k/2\rfloor+1}^{k-1} b^i \frac{\ln^{p_1}(k-i)}{(k-i)^{p_2}}, \tag{61}$$

where $\lfloor k/2 \rfloor$ stands for the largest integer smaller than $k/2$. For the first term in the right-hand side of (61), since $k-i \geq k/2$ whenever $1 \leq i \leq \lfloor k/2 \rfloor$ and the logarithm is strictly increasing, we get

$$\sum_{i=1}^{\lfloor k/2 \rfloor} b^i \frac{\ln^{p_1}(k-i)}{(k-i)^{p_2}} \leq 2^{p_2} \frac{\ln^{p_1}(k)}{k^{p_2}} \sum_{i=1}^{\infty} b^i = \left(\frac{2^{p_2} b}{1-b} \right) \frac{\ln^{p_1}(k)}{k^{p_2}}.$$

For the second term in the right-hand side of (61), since $(k-i)^{-p_2} \leq 1$, we have

$$\sum_{i=\lfloor k/2 \rfloor + 1}^{k-1} b^i \frac{\ln^{p_1}(k-i)}{(k-i)^{p_2}} \leq \ln^{p_1}(k) \sum_{i=\lfloor k/2 \rfloor + 1}^{\infty} b^i = \frac{b^{\lfloor k/2 \rfloor + 1}}{1-b} \ln^{p_1}(k).$$

Consequently,

$$\alpha_k \leq b^{k-k_0} \alpha_{k_0} + \left(\frac{2^{p_2} b}{1-b} \right) \frac{\ln^{p_1}(k)}{k^{p_2}} + \frac{b^{\lfloor k/2 \rfloor + 1}}{1-b} \ln^{p_1}(k).$$

Multiplying the above inequality by $k^{p_2} \ln^{-p_1}(k)$, we obtain

$$\alpha_k \frac{k^{p_2}}{\ln^{p_1}(k)} \leq \alpha_{k_0} b^{k-k_0} \frac{k^{p_2}}{\ln^{p_1}(k)} + \frac{2^{p_2} b}{1-b} + \frac{b^{\lfloor k/2 \rfloor + 1}}{1-b} k^{p_2}.$$

Since exponential decay dominates any power, both

$$b^{k-k_0} \frac{k^{p_2}}{\ln^{p_1}(k)} \to 0 \qquad \text{and} \qquad b^{\lfloor k/2 \rfloor + 1} k^{p_2} \to 0.$$

Hence

$$\limsup_{k \to \infty} \alpha_k \frac{k^{p_2}}{\ln^{p_1}(k)} \leq \frac{2^{p_2} b}{1-b}.$$

Because $b \in (0,1)$ was arbitrary, letting $b \downarrow 0$ yields

$$\limsup_{k \to \infty} \alpha_k \frac{k^{p_2}}{\ln^{p_1}(k)} = 0,$$

that is,

$$\alpha_k = o\left(\frac{\ln^{p_1}(k)}{k^{p_2}} \right),$$

which proves (i).

The proof of (ii) is identical, except that in this case the recurrence becomes

$$\alpha_{k+1} \leq \frac{c}{1+c} \alpha_k + \left(\frac{c_0}{1+c} \right) \frac{\ln^{p_1}(k)}{k^{p_2}}.$$

Repeating the same argument gives

$$\limsup_{k \to \infty} \alpha_k \frac{k^{p_2}}{\ln^{p_1}(k)} \leq c_0 2^{p_2},$$

which concludes $\alpha_k = \mathcal{O}\big(k^{-p_2} \ln^{p_1}(k)\big)$, as claimed.

For item (iii), since $q > 1$, the function $h_q(t) = t^q$ is convex on $[0, +\infty)$. Hence, for every $\varepsilon > 0$,

$$\alpha_{k+1}^q \geq \varepsilon^q + q\varepsilon^{q-1}(\alpha_{k+1} - \varepsilon),$$

that is,

$$\alpha_{k+1}^q - \varepsilon^q \geq q\varepsilon^{q-1}(\alpha_{k+1} - \varepsilon).$$

Combining this estimate with (14), we obtain

$$(c + q\varepsilon^{q-1})\alpha_{k+1} \leq c\alpha_k + q\varepsilon^q + c_0 \frac{\ln^{p_1}(k)}{k^{p_2}}. \tag{62}$$

Now, take

$$\varepsilon := \frac{\ln^{\frac{\varrho_1}{q-1}}(k)}{k^{\frac{1}{q-1}}}, \text{ with } \varrho_1 \in (0,1).$$

Then $\varepsilon^q = k^{-\frac{q}{q-1}} \ln^{\varrho_1 \frac{q}{q-1}}(k)$ and $\varepsilon^{q-1} = k^{-1} \ln^{\varrho_1}(k)$, so inequality (62) becomes

$$\alpha_{k+1} \leq \frac{c}{c + qk^{-1} \ln^{\varrho_1}(k)} \alpha_k + \frac{q \ln^{\varrho_1 \frac{q}{q-1}}(k)}{\big(c + qk^{-1} \ln^{\varrho_1}(k)\big) k^{\frac{q}{q-1}}} + \frac{c_0 \ln^{p_1}(k)}{\big(c + qk^{-1} \ln^{\varrho_1}(k)\big) k^{p_2}}.$$

Since $c + qk^{-1} \ln^{\varrho_1}(k) \geq c$, we deduce that

$$\alpha_{k+1} \leq \frac{c}{c + qk^{-1} \ln^{\varrho_1}(k)} \alpha_k + \left(\frac{q + c_0}{c} \right) \frac{(\ln(k))^{\max\{p_1, \varrho_1 \frac{q}{q-1}\}}}{k^{\min\{p_2, \frac{q}{q-1}\}}}.$$

Moreover,

$$\frac{c}{c+qk^{-1}\ln^{\varrho_1}(k)} = 1 - q\frac{1}{\left(c+qk^{-1}\ln^{\varrho_1}(k)\right)}\frac{\ln^{\varrho_1}(k)}{k}.$$

Since

$$q\frac{1}{(c+q\ln^{\varrho_1}(k)k^{-1})} \to \frac{q}{c}, \text{ as } k\to\infty,$$

there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$q\frac{1}{(c+q\ln^{\varrho_1}(k)k^{-1})} \geq \frac{q}{2c}.$$

Therefore, for all $k \geq k_0$,

$$\alpha_{k+1} \leq \left(1 - \frac{q}{2c}\frac{\ln^{\varrho_1}(k)}{k}\right)\alpha_k + \left(\frac{q+c_0}{c}\right)\frac{\left(\ln(k)\right)^{\max\{p_1,\varrho_1\frac{q}{q-1}\}}}{k^{\min\{p_2,\frac{q}{q-1}\}}}.$$

Now, considering the particular case where $p_1 = q$, $\varrho_1^* = q-1$ and $\varrho_2^* := \max\{p_1, \varrho_1\frac{q}{q-1}\} = q$, we simply get the inequality

$$\alpha_{k+1} \leq \left(1 - \frac{q}{2c}\frac{\ln^{\varrho_1^*}(k)}{k}\right)\alpha_k + \left(\frac{q+c_0}{c}\right)\frac{\ln^{\varrho_2^*}(k)}{k^{\min\{p_2,\frac{q}{q-1}\}}}.$$

Now, noticing that $\varrho_2^* < 3\varrho_1^* + 1 \Leftrightarrow q > 1$, Lemma A.2 implies that

$$\alpha_k = \mathcal{O}\left(\frac{\ln(k)}{k^{\min\{p_2,\frac{q}{q-1}\}-1}}\right).$$

This proves item (iii). □