A Line Search Proximal Subgradient Method for Nonconvex Stochastic Optimization

Felipe Atenas* Alejandro Jofré† Pedro Pérez-Aros‡ David Torregrosa-Belén§

November 13, 2026

Abstract

This work introduces a proximal sample-average 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 the class of upper-C^2 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 linearization of the sample average that is progressively refined, and the proximal-subgradient stepsize is selected by an Armijo line search enforcing a sufficient-decrease property up to stochastic errors induced by the sample-based approximation. This framework accommodates 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 asymptotic 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, with no prescribed growth rate.

Key words. Stochastic programming, proximal subgradient method, sample average approximation, nonconvex optimization, upper-C^2, Armijo line search

MSC codes. 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 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, [52, 15, 29, 28, 30].

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

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

*Centro de Modelamiento Matemático (CNRS IRL2807), Universidad de Chile, Santiago, Chile (fatenas@cmm.uchile.cl).

†Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático (CNRS IRL2807), Universidad de Chile, Santiago, Chile (ajofre@uchile.cl).

‡Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático (CNRS IRL2807), Universidad de Chile, Santiago, Chile (pperez@dim.uchile.cl).

§Department of Mathematics, University of Alicante, Alicante, Spain (david.torregrosa@ua.es).

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

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

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. 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. Models of the form ($\mathcal{P}$) emerge naturally in stochastic problems where nonsmooth losses and regularizers appear, such as the ones described in Section 3.1.

To tackle problem ($\mathcal{P}$), we propose a method of proximal subgradient type, in which stochastic subgradient 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., [15, 64, 1, 9]), 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., [52]), variance reduction techniques (see, e.g., [37]), and central limit theorems describing the asymptotic behavior of stochastic algorithms (see, e.g., [47] and references therein).

Despite the 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 [29, 28, 30]. 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., [56, 69, 48]. 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 and tame functions; see, e.g., [23, 20, 22, 21].

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, [41, 66, 42, 32]. This approach is especially useful when the sampled problem preserves exploitable deterministic structure. An example of the latter are the stochastic difference-of-convex algorithms proposed in [45], in which both SAA approximations and linearizations using (averaged) subgradients are combined to exploit the difference-of-convex structure. In general, however, solving a sequence of large deterministic problems may 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 [60, 40], and has become one of the central algorithmic frameworks for stochastic optimization; see, e.g., [23, 20, 22, 21]. 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., [57, 52, 44]. However, SA methods suffer from poor sample efficiency to achieve high accuracy.

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., [30, 52, 27, 21]. 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., [54, 53].

Our contribution In this work we propose a novel convergence analysis of a proximal sample-average subgradient algorithm for solving ($\mathcal{P}$) when the integrand is an upper-C^2 function (see Definition 2). At every iteration, our method constructs a surrogate model of ($\mathcal{P}$) where the expected cost is replaced by a linearization of the 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, the method draws a collection $\xi_{n_{k-1}+1}, \xi_{n_{k-1}+2}, \ldots, \xi_{n_k}$ of independent identically distributed (i.i.d.) random variables to construct a first-order approximation of the *empirical average* of the integral in ($\mathcal{P}$). Informally, for $x \in \mathbb{R}^d$ close to x^k,

$$\mathbb{E}_\mu[\varphi(\xi, x)] \approx \frac{1}{n_k}\sum_{j=1}^{n_k}\varphi(\xi_j, x) \approx \frac{1}{n_k}\sum_{j=1}^{n_k}\Big[\varphi(\xi_j, x^k) + \langle \partial\varphi(\xi_j, x^k), x - x^k\rangle\Big]$$

Leveraging the properties of upper-C^2 normal integrands (see Proposition 2), in each iteration k the method constructs a first-order majorization of the empirical average of the integral, namely, for all $x \in \mathbb{R}^d$ close to x^k, for sampled subgradients $v_j^k \in \partial\varphi(\xi_j, x^k)$ it holds that

$$\frac{1}{n_k}\sum_{j=1}^{n_k}\varphi(\xi_j, x) \le \frac{1}{n_k}\sum_{j=1}^{n_k}\Big[\varphi(\xi_j, x^k) + \langle v_j^k, x - x^k\rangle\Big] + \frac{1}{n_k}\sum_{j=1}^{n_k}\ell(\xi_j)\|x - x^k\|^2.$$

In this manner, our method generates an iterative sequence of the form

$$\begin{cases} v^k = \dfrac{1}{n_k}\displaystyle\sum_{j=1}^{n_k} v_j^k, \quad v_j^k \in \partial\varphi(\xi_j, x^k), \\ x^{k+1} \in \operatorname{prox}_{\gamma\psi}(x^k - \gamma v^k) \text{ for some } \gamma > 0, \end{cases}$$

thus combining stochastic (Clarke) subgradient information of φ taken from the empirical approximation with a proximal evaluation of ψ. Observe that the linearization of the empirical approximation prevents our method from suffering from the tractability issues known for SAA subproblems. The stepsize $\gamma > 0$ for this proximal subgradient update is computed by an Armijo-type line search [4] to ensure a sufficient decrease of the sample model of the objective function Φ in ($\mathcal{P}$). This line search replaces the necessity of vanishing stepsizes typical of stochastic (sub)gradient methods, and does not require the prior knowledge of any (sub)gradient Lipschitz bounds. 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. 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. This allows us to combine a stopping time based analysis together with new results on uniform error bounds for average approximations on compact sets to show that our method satisfies a sufficient descent condition perturbed by stochastic errors that vanish along the iteration process, as long as the sequence of iterates remains bounded.

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., [22, 72, 58, 26, 36]. 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 encompasses many relevant classes of constraints and nonconvex regularizers; see, e.g., [25, 71, 49, 51, 6].

Our contributions can be summarized as follows. We introduce the notion of upper-C^2 normal integrands, and show that this class of functions appears naturally in applications. Furthermore, in Theorem 1, we prove that almost surely every accumulation point of bounded sequences generated by our method with line search 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 8 for more details. Furthermore, our new analysis is based on a sufficient decrease condition up to controllable stochastic errors obtained from a line search test that only uses sampled information, in contrast to, e.g., [21, 11, 14, 12], where a differential inclusions approach is taken for methods of SA nature.

Organization of the paper In Section 2, we present the notation and concepts used in this work and show some preliminary results. In Section 3, we introduce the notion of upper-C^2 normal integrands, describe some optimization models with this structure, and show properties pivotal in our analysis. Section 4 presents our proposed method (Algorithm 1) to solve problem ($\mathcal{P}$) and its analysis of almost surely subsequential convergence. Section 5 reports numerical experiments comparing our method with proximal stochastic subgradient methods. 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 $f:\mathbb{R}^d\to\mathbb{R}$, Lipschitz continuous around $x\in\mathbb{R}^d$, the Clarke subdifferential at x is defined as

$$\partial f(x)=\{v\in\mathbb{R}^d:\langle v,d\rangle\le f^\circ(x;d)\ \forall d\in\mathbb{R}^d\},\ \text{where } f^\circ(x;d)=\limsup_{\substack{y\to x\\ t\to 0^+}}\tfrac{f(y+td)-f(y)}{t}$$

is the Clarke subderivative at x in the direction $d\in\mathbb{R}^d$. 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 $\operatorname{prox}_{\gamma f}:\mathbb{R}^d\rightrightarrows\mathbb{R}^d$ defined as the solution set of the optimization problem

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

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

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

The following lemma shows a type of upper-semicontinuity property of the proximal mapping for prox-bounded functions. Its proof follows the same arguments as in [55, Lemma 2.3] and is therefore omitted.

Lemma 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 $\widehat{x}^k \to \bar{x}$.

2.2 Elements from measure theory

In what follows, $(\Xi, \mathcal{A}, \mu)$ will be a complete probability space, and we use the notation L^p_μ for denoting the space of all p-integrable functions. Moreover, we consider a complete 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: $\mathbb{1}_F(\omega) = 1$ if $\omega \in F$, and $\mathbb{1}_F(\omega) = 0$ otherwise. 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 next 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., [13, Lemma 10.10.2]. The proof follows from standard approximation techniques based on simple functions and can be found in [10, Lemma 4.11].

Lemma 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

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

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 1 (locally Lipschitz normal integrand)**.** A normal integrand $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ 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 almost surely $\xi \in \Xi$, $|\varphi_\xi(y) - \varphi_\xi(z)| \leq \kappa(\xi)\|y - z\|$, for all $y, z \in V$.

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 1. 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 almost surely $\xi \in \Xi$

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

2.3 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

$$\mathbf{E}^k_\varphi(x) := \frac{1}{n_k}\sum_{j=1}^{n_k} \varphi(\xi_j, x), \text{ for all } x \in \mathbb{R}^d, \tag{4}$$

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

$$\Phi_k(x) := \mathbf{E}^k_\varphi(x) + \psi(x), \text{ for all } x \in \mathbb{R}^d. \tag{5}$$

In the context of the approximation in (4)-(5), 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 our method, we need to control the error associated with replacing the expected value with the empirical average in (5) during the iteration process. The following lemma furnishes a nonasymptotic bound for the uniform error of the average approximation generated by the sampling.

Lemma 3. Let $(\xi_i)_{i\geq 1}$ be a sampling with sample size $(n_k)_{k\in\mathbb{N}}$ satisfying Assumption 1, and let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ be a normal integrand which is locally Lipschitz around any point $x \in \mathbb{R}^d$ with square integrable modulus. Then, the following hold.

(i) For all $r > 0$, there exists a constant a_r such that

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

(ii) There exists a measurable set $\widehat{\Omega}$ with $\mathbb{P}(\Omega\backslash\widehat{\Omega}) = 0$ such that for all $\omega \in \widehat{\Omega}$ and all $r > 0$ the expected value function $\mathbf{E}_\varphi$ is finite and continuous, and $\mathbf{E}^k_\varphi$ converges uniformly to $\mathbf{E}_\varphi$ on $\mathbb{B}_r[0]$.

Proof. The first assertion follows directly from [43, Theorem 2.1 and Proposition 2.6]. Furthermore, (ii) is a consequence of [67, Theorem 9.60].

□

To conclude this section, we state the following technical lemma which 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 1.

Lemma 4. Let $(n_k)_{k\in\mathbb{N}}$ be a divergent, nondecreasing sequence of positive numbers. Then

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

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

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

Since $n_{k+1} \to +\infty$, we get $\sum_{k=1}^{\infty} \frac{n_{k+1}-n_k}{\sqrt{n_k}n_{k+1}} < +\infty$. □

3 Upper-C^2 normal integrands

Several notions of nonconvexity have been defined in the literature to cover different structures in optimization problems. In particular, the class of lower-C^2 functions [61, 17] generalize both convexity and smoothness. These functions can be locally represented as the maximum of a family of twice continuously differentiable functions parametrized over a compact set, and coincide with the set of locally weakly convex functions [61, Theorem 6]. In this work, we focus on the "mirror" class of the aforementioned functions for integrands, namely, the upper-C^2 integrands that capture value functions of parametric minimization over a compact set. The latter family of functions appear, for instance, in partial-learning tasks [18], trimmed least squares [63], power systems [70], and the models described in Section 3.1 below. Observe that the class of upper-C^2 functions has received comparatively less attention than their lower-C^2 counterpart. We refer to [3, 2, 70, 39, 38] for recent works studying the family of functions of our interest in deterministic settings.

We now introduce the definition of upper-C^2 integrands that will constitute a fundamental structural assumption throughout the remainder of the manuscript.

Definition 2 (upper-C^2 normal integrand)**.** Let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ be a normal integrand. We say that φ is upper-C^2 around $\bar{x}$ if φ is locally Lipschitz around $\bar{x}$ with square integrable modulus and if there exist a compact convex neighborhood V of $\bar{x}$, a nonnegative $\varrho \in L^1_\mu$, a Polish space X, a measurable set-valued mapping $C : \Xi \rightrightarrows X$ with nonempty compact values, and a measurable mapping $g : \Xi \times X \times \mathbb{R}^d \to \mathbb{R}$ such that almost surely $\xi \in \Xi$,

$$\varphi_\xi(x) = \min_{t\in C(\xi)} g(\xi, t, x) \ x \in V, \tag{6}$$

the map $x \mapsto g(\xi, t, x)$ is twice continuously differentiable on an open set containing V, the mappings g, $\nabla_x g$ and $\nabla^2_x g$ are continuous in (t, x), and $\|\nabla^2_x g(\xi, t, x)\| \leq \varrho(\xi)$ for all $\xi \in \Xi$, $t \in C(\xi)$, $x \in V$. 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 1. Definition 2 naturally extends the notion of upper-C^2 functions to the setting of integrands. Recall the classical definition of the upper-C^2 class consisting of those functions that can be locally represented as the minimum over a compact of a parametrized family of twice continuously differentiable functions (see, e.g., [62, Definition 10.29]). This class of functions has been traditionally well-recognized in the variational analysis and generalized differentiation literature, and more recently for the development of numerical methods, as they can be characterized by a nonsmooth counterpart of the descent lemma; see, e.g., [2]. It has been shown that the class of upper-C^2 functions is broad, encompassing structured functions given by the difference between a smooth function with a locally Lipschitz continuous gradient and a possibly nonsmooth convex function; see [2, Proposition 3.2] for the deterministic setting and [7, Corollary 4.1] for the stochastic case.

3.1 Examples of upper-C^2 normal integrands

In this section, we describe some important examples of upper-C^2 normal integrands that appear in machine learning applications.

3.1.1 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, e.g., [31] 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

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

where N is a regularization term that induces a certain structure on the quantization, e.g., sparsity, thus controlling the number of quantization points. See [3, Section 4.2] for a deterministic, empirical-risk formulation of the problem. The integrand above, namely,

$$\varphi_\xi(x) := \min_{t\in\{1,\ldots,s\}} \|x_t - \xi\|^2, \quad \text{for } \xi \in \mathbb{R}^d, x = (x_1, \ldots, x_s) \in \mathbb{R}^{d\times s}, \tag{7}$$

defines an upper-C^2 normal integrand. Indeed, suppose that $\mathbb{E}_\mu[\|\xi\|^4] < \infty$. First, $\varphi_\xi(x) \le 2\|x\|^2 + 2\|\xi\|^2$, which shows that $\varphi_\xi \in L^2_\mu$. Moreover, for a $V \subset \mathbb{R}^{d\times s}$ compact convex neighborhood of a given $\bar{x}$, observe that for all $x, y \in V$,

$$|\varphi_\xi(x) - \varphi_\xi(y)| \;\le\; \kappa_V(\xi)\,\|x - y\|, \qquad \kappa_V(\xi) := 2(\|\xi\| + R_V),$$

where $R_V := \max_{x=(x_1,\ldots,x_s)\in V} \; \max_{1\le j\le s} \|x_j\| < \infty$. Then the assumption on the second moment implies that φ is locally Lipschitz around $\bar{x}$ with square integrable modulus. Furthermore, setting $X = \{1, \ldots, s\}$, $C(\xi) = X$, $g(\xi, t, x) = \|\xi - x_t\|^2$, where for fixed (ξ, t) the map $x \mapsto g(\xi, t, x) = \|x_t - \xi\|^2$ is a quadratic polynomial with $\|\nabla_x^2 g(\xi, t, x)\| = 2$. Then, the function φ in (7) is an upper-C^2 normal integrand around $\bar{x}$. Note that if μ has bounded support, say $\operatorname{supp}\mu \subseteq \mathbb{B}_R[0]$, we may therefore replace the regularizer N by $N + \iota_{(\mathbb{B}_R[0])^s}$ so that $\operatorname{dom}\psi$ is bounded.

3.1.2 Minimum eigenvalue of a stochastic matrix

Given a real symmetric matrix $A \in \mathbb{R}^{p\times p}$, it is well-known that its minimum eigenvalue is the optimal value of the problem $\min_{\|u\|=1}\langle u, Au\rangle$, which defines a deterministic upper-C^2 function. We proceed to describe its stochastic counterpart.

Let $\mathbb{S}^p$ denote the space of real symmetric $p \times p$ matrices, equipped with the spectral norm $\|\cdot\|_2$, and let $A : \Xi \times \mathbb{R}^d \to \mathbb{S}^p$ be a mapping. Suppose that for every $x \in \mathbb{R}^d$, the map $\xi \mapsto A(\xi, x)$ is measurable, and that for every $\xi \in \Xi$, the map $x \mapsto A(\xi, x)$ is C^2, namely, that $D_x A(\xi, x)$, $D_x^2 A(\xi, x)$ are continuous in x. Set

$$\varphi_\xi(x) := \lambda_{\min}(A(\xi, x)) \text{ for all } x \in \mathbb{R}^d \text{ and } \xi \in \Xi, \tag{8}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue operator. The function in (8) defines an upper-C^2 normal integrand. Indeed, let $\bar{x} \in \mathbb{R}^d$, and let V be a compact convex neighborhood of $\bar{x}$. Set

$$\kappa_V(\xi) := \sup_{x\in V} \|D_x A(\xi, x)\|, \qquad \varrho_V(\xi) := \sup_{x\in V} \sup_{\|h\|=1} \left\|D_x^2 A(\xi, x)[h, h]\right\|_2, \tag{9}$$

and suppose that $\varrho_V \in L^1_\mu$ and $\kappa_V, \|A(\cdot, x)\| \in L^2_\mu$ for any $x \in V$. In Definition 2, set $X = \mathcal{S}^{p-1} := \{u \in \mathbb{R}^p : \|u\| = 1\}$, $C(\xi) = \mathcal{S}^{p-1}$, and $g\colon \Xi \times \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ given by $g(\xi, u, x) = \langle u, A(\xi, x)u\rangle$. The Rayleigh quotient characterization of the eigenvalues gives (6). Furthermore, for $\|u\| = \|h\| = 1$, (9) yields $\left|\nabla_x^2 g(\xi, u, x)[h, h]\right| \le \left\|D_x^2 A(\xi, x)[h, h]\right\|_2 \|u\|^2 \le \varrho_V(\xi)$. Then Definition 2 holds true.

Remark 2. A particular setting in which the constants in (9) are, respectively, integrable and square integrable, is the following. Consider $B_0, B_1, \ldots, B_q : \mathbb{R}^d \to \mathbb{S}^p$ of class C^2, and $w = (w_1, \ldots, w_q) : \Xi \to \mathbb{R}^q$ measurable, such that $\mathbb{E}[\|w(\xi)\|_1^2] < \infty$. Define

$$A(\xi, x) := B_0(x) + \sum_{j=1}^{q} w_j(\xi)\, B_j(x), \quad \text{for } x \in \mathbb{R}^d \text{ and } \xi \in \Xi.$$

For $\bar{x} \in \mathbb{R}^d$, V as above, and κ_V and ϱ_V given in (9), it holds that $\kappa_V(\xi) \leq \alpha_0 + \bar{\alpha}\,\|w(\xi)\|_1$ and $\varrho_V(\xi) \leq \beta_0 + \bar{\beta}\,\|w(\xi)\|_1$, where $\alpha_j := \sup_{x\in V} \|D_x B_j(x)\|$ and $\beta_j := \sup_{x\in V} \sup_{\|h\|=1} \|D_x^2 B_j(x)[h,h]\|_2$ for $j = 0, 1, \ldots, q$, with $\bar{\alpha} := \max_{1\leq j\leq q} \alpha_j$ and $\bar{\beta} := \max_{1\leq j\leq q} \beta_j$.

3.1.3 Dictionary learning

The (sparse) dictionary learning problem aims to find a representation of data vectors as sparse linear combinations of basis elements. Formally, consider a (dictionary) matrix $D = [d_1, \ldots, d_K] \in \mathbb{R}^{p\times K}$ whose columns d_j, called *atoms*, form a redundant spanning family of $\mathbb{R}^p$, that is, $K > p$ holds. Let the data $\xi \in \Xi \subseteq \mathbb{R}^p$ be distributed according to a probability measure μ. Each realization ξ carries its own admissible set of coefficient vectors $T(\xi) := \{\alpha \in \mathbb{R}^K : \ \|\alpha\|_1 \leq \tau(\xi)\}$, where $\tau : \Xi \to [0, +\infty)$ is measurable. Given $\lambda \geq 0$, the objective function of the dictionary learning problem is determined by the integrand

$$\varphi_\xi(D) \ = \ \min_{\alpha\in T(\xi)} \ \Big\{ \tfrac{1}{2}\|\xi - D\alpha\|_2^2 + \lambda\|\alpha\|_1 \Big\}, \tag{10}$$

resulting in the following stochastic program

$$\min_{D\in\mathbb{R}^{p\times K}} \ \ \mathbb{E}_\mu[\varphi_\xi(D)] + \iota_{\mathcal{D}}(D), \tag{11}$$

where $\iota_{\mathcal{D}}$ is the indicator of the set $\mathcal{D} := \{D \in \mathbb{R}^{p\times K} : \ \|d_j\|_2 \leq 1, \ j = 1, \ldots, K\}$. This constraint is necessary to prevent the solutions from being invariant under (positive) rescaling. The constraint set $\mathcal{D}$ satisfies $\|D\|_F \leq \sqrt{K}$ for every $D \in \mathcal{D}$, so $\operatorname{dom} \iota_{\mathcal{D}}$ is bounded. By replacing μ with its empirical approximation and taking τ constant, (11) yields the deterministic formulation of the dictionary learning problem [50].

We now verify that (10) defines an upper-C^2 normal integrand. Suppose that

$$\mathbb{E}_\mu\Big[\tau(\xi)^2\big(\|\xi\|_2 + \tau(\xi)\big)^2\Big] < \infty \quad \text{and} \quad \mathbb{E}_\mu[\|\xi\|_2^4] < \infty,$$

and set $X = \mathbb{R}^K$, $C(\xi) = T(\xi)$, $g(\xi, \alpha, D) = \frac{1}{2}\|\xi - D\alpha\|_2^2 + \lambda\|\alpha\|_1$. Observe that the set-valued mapping C is measurable [62, Example 14.15]. Moreover, as $\nabla_D^2 g(\xi, \alpha, D) = (\alpha\alpha^\top) \otimes I_p$, where $\otimes$ is the Kronecker product and I_p is the identity matrix of dimension p, it follows that $\|\nabla_D^2 g(\xi, \alpha, D)\|_2 = \|\alpha\|_2^2 \leq \|\alpha\|_1^2 \leq \tau(\xi)^2$. Then $\varrho := \tau^2 \in L_\mu^1$ follows from the moments assumption and that $\tau^2 \leq \tau^2(\|\xi\|_2 + \tau)^2$ whenever $\|\xi\|_2 + \tau \geq 1$, and $\tau^2 \leq \tau^4 + 1$ otherwise. Next, let $V \subset \mathbb{R}^{p\times K}$ be a compact convex neighborhood of a given $\bar{D}$ and $R_V := \max_{D\in V} \|D\|_F < \infty$. Then for all $\xi \in \Xi$ and all $D, D' \in V$,

$$\big|\varphi_\xi(D) - \varphi_\xi(D')\big| \ \leq \ \kappa_V(\xi)\,\|D - D'\|_F, \quad \kappa_V(\xi) := \tau(\xi)\big(\|\xi\|_2 + R_V\,\tau(\xi)\big),$$

with $\kappa_V \in L_\mu^2$ under the moments assumption. Then the function φ given by (10) is an upper-C^2 normal integrand.

3.2 Some properties of upper-C^2 normal integrands

We proceed to establish some properties of the family of upper-C^2 normal integrands. First, we formalize that upper-C^2 normal integrands satisfy a (random) local version of the descent lemma, which will be used throughout the development of our method.

Proposition 2. Let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ be a finite-valued normal integrand, and let $\bar{x} \in \mathbb{R}^d$. Assume that φ is locally Lipschitz around $\bar{x}$ with square integrable modulus. If φ is upper-C^2 around $\bar{x}$, then there exist a neighborhood V of $\bar{x}$ and (a nonnegative) $\ell \in L^1_\mu$ such that for almost surely $\xi \in \Xi$, $x \in V$ and $v \in \partial\varphi_\xi(x)$, it holds

$$\varphi_\xi(y) \leq \varphi_\xi(x) + \langle v, y - x\rangle + \ell(\xi)\|y - x\|^2, \quad \text{for all } y \in V. \tag{12}$$

Proof. Throughout, V, ϱ, X, C and g denote the objects introduced in Definition 2. After replacing V by a smaller neighborhood, we may suppose that every φ_ξ is Lipschitz continuous on an open set containing V. Furthermore, let us denote $\widehat{\Xi} \subseteq \Xi$ a measurable set such that $\mu(\widehat{\Xi}) = 1$ for which the representation (6) holds. Fix $\xi \in \widehat{\Xi}$, two points $x, y \in V$, and an arbitrary $v \in \partial\varphi_\xi(x)$. First, by the Clarke calculus rule for minima and maxima (see, e.g., [16, Theorem 2.8.2]) there exist an integer $n \geq 1$, indices $t_1, \ldots, t_n \in C(\xi)$ and weights $\lambda_1, \ldots, \lambda_n \geq 0$ with $\sum_{i=1}^n \lambda_i = 1$ such that

$$v = \sum_{i=1}^n \lambda_i \nabla_x g(\xi, t_i, x), \text{ and } g(\xi, t_i, x) = \varphi_\xi(x) \text{ for all } i = 1, \ldots, n.$$

Now fix $i \in \{1, \ldots, n\}$. As φ_ξ is an infimum over $C(\xi)$ and $t_i \in C(\xi)$, we have $\varphi_\xi(y) \leq g(\xi, t_i, y)$, while the standard descent lemma applied to the smooth map $g(\xi, t_i, \cdot)$ yields

$$g(\xi, t_i, y) \leq g(\xi, t_i, x) + \langle \nabla_x g(\xi, t_i, x), y - x\rangle + \frac{\varrho(\xi)}{2}\|y - x\|^2.$$

Moreover, t_i is active, so $g(\xi, t_i, x) = \varphi_\xi(x)$. Putting these three facts together, we obtain

$$\varphi_\xi(y) \leq \varphi_\xi(x) + \langle \nabla_x g(\xi, t_i, x), y - x\rangle + \frac{\varrho(\xi)}{2}\|y - x\|^2, \qquad i = 1, \ldots, n.$$

Therefore, multiplying the above inequality by λ_i and summing over i, we get (12). □

To end this section we study the convergence of the Clarke sample-based subgradients for upper-C^2 normal integrands. Results of similar nature can be found in [19, 65, 68].

Lemma 5. Let $(\xi_i)_{i\geq 1}$ be a sampling with sample size $(n_k)_{k\in\mathbb{N}}$ satisfying Assumption 1, and let $\varphi : \Xi \times \mathbb{R}^d \to \mathbb{R}$ 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 $f(x, h) := (\mathbf{E}_\varphi)^\circ(x; h)$ and $f^k(x, h) := (\mathbf{E}^k_\varphi)^\circ(x; h)$. By the Clarke interchange formula for subdifferentiation under the integral [16, 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 $-\varphi_\xi$ [16, 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 [5, 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 $\langle v^{\nu_k}, h\rangle \leq$

$(\mathbf{E}_\varphi^{\nu_k})^\circ(x^{\nu_k};h) = f^{\nu_k}(x^{\nu_k},h)$, for all $k\in\mathbb{N}$. Since $(x^{\nu_k},h)\to(x,h)$ and f^k hypographically converges to f, [62, Eq. 7(9)] implies

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

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. □

4 Proximal stochastic subgradient algorithm with line search

In this section, we formulate and analyze our method for solving the nonconvex problem ($\mathcal{P}$), corresponding to Algorithm 1. Throughout this section, 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;

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

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

3: Define $v^k:=\frac{1}{n_k}\sum_{i=1}^{n_k}v_i^k$ and set Φ_k using (5). ▷ Subgradient estimate

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

5: Compute ▷ Proximal line search

$$\widehat{x}^k\in\operatorname{prox}_{\gamma_k\psi}\left(x^k-\gamma_k v^k\right). \tag{14}$$

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 (14).

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

Algorithm 1: Proximal sample-average subgradient algorithm for problem ($\mathcal{P}$)

Remark 3. 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, $\mathbb{E}[\nabla_x\varphi(\xi_i,x)]=\nabla\mathbf{E}_\varphi(x)$ 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 [20], 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, $\sup_{x\in U}\mathbb{E}_\mu[\|g(\xi,x)\|^2]<+\infty$. In contrast, since we assume only local Lipschitz continuity of φ in Definition 1, 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 1. 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., [30, 28] and the references therein. More recently, this condition has been weakened by allowing multiplicative noise [37]. 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 differentiable 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 1 provided that V is an open convex set and that there exists $\widehat{x} \in \mathbb{R}^d$ such that $\varphi(\cdot, \widehat{x}) \in L^2_\mu$. For nonconvex normal integrands, relations of this type appear in [7, Corollary 5.3 & Remark 5.4].

The method proposed in Algorithm 1 approximates the (sub-)gradient of the expected cost function 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 empirical average is linearized through its stochastic (sub-)gradient to solve the proximal subproblem, and 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. Moreover, each line search adjustment does not require additional stochastic subgradient evaluations, and a single sample-average subgradient suffices for the entire line search procedure. This is in contrast with other line searches proposed for stochastic methods, which require fresh gradient computations at every trial step; see, e.g., [54, 53].

4.1 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. 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 $v \in \partial \mathbf{E}^k_\varphi(x)$, one has

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

provided that

$$w \in \operatorname{prox}_{\gamma\psi}(x - \gamma v), \quad \text{with } \gamma \in (0, \vartheta). \tag{15}$$

Proof. Let $\bar{x} \in \operatorname{dom}\psi$ and $\sigma > 0$, and take $r \in \mathbb{N}$ such that $\|\bar{x}\| \leq r - 1$. By Proposition 1, there exist $\kappa_r \in L^2_\mu$ and a measurable set $\widehat{\Xi}_1 \subseteq \Xi$ with $\mu(\widehat{\Xi}_1) = 1$, such that (3) holds on $\mathbb{B}_r[0]$ for all $\xi \in \widehat{\Xi}_1$. In addition, from Proposition 2, there exist $V \subseteq \mathbb{B}_r[0]$, $\ell_r \in L^1_\mu$ and a measurable set $\widehat{\Xi}_2 \subseteq \Xi$ with $\mu(\widehat{\Xi}_2) = 1$, such that (12) holds for all $\widehat{\Xi}_2$. By the strong law of large numbers, we can assume that there exists a measurable set $\Omega_r \subseteq \Omega$ with $\mathbb{P}(\Omega_r) = 1$ such that for all $\omega \in \Omega_r$

$$\varsigma_r := \sup_{k\in\mathbb{N}} \frac{1}{n_k}\sum_{i=1}^{n_k} \ell_r(\xi_i) < +\infty \quad \text{and} \quad \widetilde{\varsigma}_r := \sup_{k\in\mathbb{N}} \frac{1}{n_k}\sum_{i=1}^{n_k} \kappa_r(\xi_i) < +\infty. \tag{16}$$

Set $\widehat{\Omega} := \Omega_r \cap \bigcap_{i=1}^\infty \xi_i^{-1}(\widehat{\Xi}_1 \cap \widehat{\Xi}_2)$. Fix $\omega \in \widehat{\Omega}$. For any $x \in V$ and $v \in \partial \mathbf{E}^k_\varphi(x)$, in view of (13) and [16, Proposition 2.1.2], there exist $v_i \in \partial\varphi_{\xi_i}(x)$, $i = 1, \ldots, n_k$, such that

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

Hence, the set $\{v \in \partial \mathbf{E}^k_\varphi(x) : x \in V,\ k \in \mathbb{N}\}$ is bounded. Then, by Lemma 1, we can consider a neighborhood $W \subseteq V$ of $\bar{x}$, a positive constant ϑ_0 so that any vector w given by (15) with

$x \in W$ and $\gamma \in (0, \vartheta_0)$ belongs to V. By resorting to (12), we have that, for all $k \in \mathbb{N}$ and all ξ_i, $i = 1, \ldots, n_k$, the following inequality holds

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

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

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

By (15), 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

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

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

A direct consequence of Proposition 3 is that the line search procedure in Algorithm 1 is well-defined, as shown in the next result.

Corollary 1 (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 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 $\Phi_k(\widehat{x}^k) \leq \Phi_k(x^k) - \sigma\|\widehat{x}^k - x^k\|^2$. □

Remark 4. In view of the proof of Proposition 3, when the function ℓ in the condition (12) is known globally, for example, when φ_ξ is smooth with deterministic globally Lipschitz gradient, the stepsize γ_k can be taken in each iteration in a way that the descent condition in Step 6 of Algorithm 1 is satisfied after one trial.

4.2 Subsequential convergence analysis of Algorithm 1

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 $j > n_k$. Furthermore, we assume that x^k is $\mathcal{F}_{k-1}$-measurable, for $k \geq 1$.

Remark 5 (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 3. 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 (14) 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 (14). 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 4.2.1, we present some technical lemmas instrumental for the proof of the theorem, which will be developed in Section 4.2.2.

Theorem 1 (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 $\left(\Phi_k(x^k)\right)_{k\in\mathbb{N}}$ and $\left(\Phi(x^k)\right)_{k\in\mathbb{N}}$ 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 $\Phi(\bar{x}) = \lim_{k\to\infty} \Phi_k(x^k) = \lim_{k\to\infty} \Phi(x^k) \in \mathbb{R}$*;*

(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 $k \to \infty$.

Remark 6 (On the boundedness of the iterates)**.** Theorem 1 is valid under boundedness of the trajectory $(x^k)_{k\in\mathbb{N}}$, which holds particularly when $\operatorname{dom}\psi$ is bounded. Indeed, by Step 7 and (14) one has $x^{k+1} \in \operatorname{prox}_{\gamma_k\psi}(x^k - \gamma_k v^k) \subseteq \operatorname{dom}\psi$ for all $k \in \mathbb{N}$, so that

$$\sup_{k\in\mathbb{N}} \|x^k\| \leq \sup_{x\in\operatorname{dom}\psi} \|x\| < +\infty.$$

Observe that $\operatorname{dom}\psi$ is bounded in our examples of Section 3.1.

4.2.1 Technical results for Theorem 1

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

Lemma 6. Suppose that Assumption 3 holds. Given $r > 0$ such that $\|x^0\| \leq r$, define

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

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

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

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

In the next lemma, we show that the stopped sequence defined in Lemma 6 satisfies a sufficient decrease condition up to errors.

Lemma 7. Suppose that the assumptions of Theorem 1 hold. In the setting of Lemma 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

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

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

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

By (4)-(5), then

$$\sigma\|x^{k+1} - x^k\|^2 \leq \Phi_k(x^k) - \Phi_k(x^{k+1}) = \Delta\,\Phi_k + \mathbf{E}_\varphi^{k+1}(x^{k+1}) - \mathbf{E}_\varphi^k(x^{k+1}), \tag{22}$$

where $\Delta\,\Phi_k := \Phi_k(x^k) - \Phi_{k+1}(x^{k+1})$ for all $k \in \mathbb{N}$. Given a fixed $k \in \mathbb{N}$, note that

$$\|x_r^{k+1} - x_r^k\| = \begin{cases} 0, & \text{if } \tau_r \leq k, \\ \|x^{k+1} - x^k\|, & \text{if } \tau_r \geq k+1, \end{cases} \tag{23}$$

and

$$\Phi_k^r = \begin{cases} \Phi_{\tau_r}(x^{\tau_r}), & \text{if } \tau_r \leq k, \\ \Phi_k(x^k), & \text{if } \tau_r \geq k+1. \end{cases} \tag{24}$$

Hence, on the one hand, using (22) and (23) 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 have $\sigma\|x_r^{k+1} - x_r^k\|^2 = 0$ and $\Phi_k^r - \Phi_{k+1}^r + W_k^r = 0$, so we conclude that (20) holds. □

The estimate shown in Lemma 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 8. Let us assume the setting of Lemma 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 $Z_k^r := \mathbb{E}[W_k^r|\mathcal{F}_k]$. Then $\sum_{k\in\mathbb{N}} \mathbb{E}[|Z_k^r|] < +\infty$.

Proof. Let us consider an arbitrary $r > 0$ such that $\|x^0\| \leq r$. From Lemma 6, $\|x_r^k\| \leq r$ for all $k \in \mathbb{N}$. Furthermore, by Proposition 1, there exist square integrable functions κ and δ such that for all $\xi \in \Xi$,

$$|\varphi_\xi(u)| \leq \kappa(\xi)\|u\| + \delta(\xi), \quad \text{for all } u \in \mathbb{B}_r[0]. \tag{25}$$

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, $\mathbb{E}[|W_k^r|] \leq 2\mathbb{E}[\kappa]r + 2\mathbb{E}[\delta]$. 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 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 6 that the function x_r^{k+1} is $\mathcal{F}_k$-measurable, and by (25) the function $\omega \mapsto \varphi_{\xi_j(\omega)}(x_r^{k+1}(\omega))$ is integrable, so using Lemma 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 3(i) that there exists a constant $a_r > 0$, independent of k, such that

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

By Lemma 4, this yields $\sum_{k\in\mathbb{N}} \mathbb{E}[|Z_k^r|] < +\infty$, 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 9. Suppose the assumptions of Lemma 8 hold. Then, $(\Phi_k^r)_{k\in\mathbb{N}}$ converges a.s. to some finite value and

$$\sum_{k=0}^{\infty} \|x_r^{k+1} - x_r^k\|^2 < +\infty \text{ a.s.} \tag{26}$$

Proof. Let κ and δ be the functions in (25) (which exist in view of Proposition 1). Let us define

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

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 (25),

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

which particularly implies that Φ_{k+1}^r is integrable in view of Lemma 8. Next, we introduce the process $Y_k^r := \mathbb{E}\left[s_{k+1}^r - s_k^r \,\middle|\, \mathcal{F}_k \right]$ for all $k \in \mathbb{N}$. Notice that (20) yields, for all $k \in \mathbb{N}$,

$$\mathbb{E}[\Phi_{k+1}^r + s_{k+1}^r \mid \mathcal{F}_k] \leq \Phi_k^r + s_k^r - \sigma\|x_r^{k+1} - x_r^k\|^2 + |Y_k^r| + |Z_k^r|. \tag{27}$$

A direct computation yields

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

Hence, by Lemma 3, 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 a constant depending on the second moment of the square integrable function $\xi \mapsto r\kappa(\xi) + \delta(\xi)$. In this manner, Lemma 4 implies $\sum_{k\in\mathbb{N}} \mathbb{E}\big[|Y_k^r|\big] < +\infty$. Therefore, by Lemma 8, the nonnegativity of the sequence $\big(\Phi_k^r + s_k^r\big)_{k\in\mathbb{N}}$, and (27), it follows from the well-known Robbins–Siegmund supermartingale convergence theorem (see [59, Theorem 1]) that $\big(\Phi_k^r + s_k^r\big)_{k\in\mathbb{N}}$ converges and (26) 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.

□

4.2.2 Proof of Theorem 1

We are now in a position to prove each assertion of Theorem 1.

Proof of Theorem 1. We consider a measurable set $\widehat{\Omega}$ of full measure such that for all $\omega \in \widehat{\Omega}$, the conclusions of Proposition 3 and the following conditions hold: for all $r \in \mathbb{N}$ with $\|x^0\| \leq r$,

- a_ω) from Lemma 9,

$$\sum_{k\in\mathbb{N}} \|x_r^{k+1}(\omega) - x_r^k(\omega)\|^2 < +\infty \tag{28}$$

  and Φ_k^r converges to some random variable;

- b_ω) by Lemma 3(ii) we get $\Big|\mathbf{E}_\varphi(x_r^k) - \mathbf{E}_\varphi^k(x_r^k)\Big| \to 0$;

- 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 1 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 (23) and (24), we get that

$$\Phi_k(x^k) = \Phi_k^r \text{ and } \|x^{k+1} - x^k\| = \|x_r^{k+1} - x_r^k\| \text{ for all } k \in \mathbb{N}.$$

Therefore, $\big(\Phi_k(x^k)\big)_{k\in\mathbb{N}}$ converges and (19) holds due to (28). Moreover, by b_ω), (23) 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 $x^{\nu_i} \to \bar{x}$. Observe that by lower semicontinuity of Φ and convergence of $\big(\Phi(x^k)\big)_{k\in\mathbb{N}}$, we have $\bar{x} \in \operatorname{dom}\psi$. 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 $v_j^{\nu_i} \in \partial\varphi_{\xi_j}(x^{\nu_i})$ for all $j = 1, \ldots, n_{\nu_i}$. 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, 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 $x^{k_j}(\omega) \to \bar{x} \in \operatorname{dom}\psi$ 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 $\gamma_{k_j} \to \bar{\gamma} \in (0, +\infty]$. By (28), 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 5, we have that $\bar{v} \in \partial \mathbf{E}_\varphi(\bar{x})$. In addition, by (14) and the definition of the proximal operator, for all $u \in \mathbb{R}^d$, we have

$$\psi(x^{k_j+1}) + \langle v^{k_j}, x^{k_j+1}\rangle + \frac{1}{2\gamma_{k_j}}\|x^{k_j+1} - x^{k_j}\|^2 \leq \psi(u) + \langle v^{k_j}, u\rangle + \frac{1}{2\gamma_{k_j}}\|u - x^{k_j}\|^2. \quad (29)$$

Using the inequality above with $u = \bar{x}$ and lower semicontinuity of ψ, we get that

$$\limsup_{j\to\infty} \psi(x^{k_j+1}) \leq \psi(\bar{x}) \leq \liminf_{j\to\infty} \psi(x^{k_j+1}),$$

from which we obtain $\lim_{j\to\infty} \psi(x^{k_j+1}) = \psi(\bar{x})$. Note that Lemma 3(ii) particularly implies

$$\Phi_{k_j+1}(x^{k_j+1}) \to \Phi(\bar{x}). \quad (30)$$

In this manner, fixing $u \in \mathbb{R}^d$ in the right-hand side of (29), 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 $\bar{\gamma} = \infty$. In the latter case, the same estimate holds replacing $\bar{\gamma}$ with any $\gamma > 0$. 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 the limit $\lim_{k\to\infty} \Phi_k(x^k) = \lim_{k\to\infty} \Phi(x^k)$ exists, and thus combined with (30) 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 (28), 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., [24, Theorem 8.3.9 and Proposition 8.3.10]). □

Remark 7. It is worth mentioning that if one requires that (a.s.) $\|x^k\| \le r$ 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 (27) we get that $\sum_{k\in\mathbb{N}} \mathbb{E}\|x^{k+1} - x^k\|^2 < +\infty$ and that $(\Phi_k(x^k))_{k\in\mathbb{N}}$ converges to an integrable random variable.

Remark 8 (Sampling budget)**.** A distinctive feature of Theorem 1 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., [27, 46, 45, 37, 47]. Specifically, within the smooth, convex setting, the reduced-variable stochastic algorithm in [37] 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 [30] are assumed to scale at least linearly, and no trajectory-level convergence guarantees are obtained. Central limit theorems and asymptotic properties studied in [47] similarly rely on polynomial sample-size growth schedules of the form $n_k^{-1} = \mathcal{O}(k^\nu)$ with $\nu \ge 1$ to enforce noise stabilization. For nonconvex settings, even more demanding polynomial configurations are typically assumed; for instance, the stochastic difference-of-convex framework in [45] requires $n_k^{-1} = \mathcal{O}(k^\nu)$ with a power of at least $\nu \ge 2$, while related nonconvex schemes [46] obtain convergence guarantees with strictly increasing polynomial or geometric sample paths. In the infinite-dimensional Hilbert space setting, the convergence result in [27] needs the summability condition $\sum_{k\in\mathbb{N}} 1/n_k < +\infty$.

Moreover, since Step 2 of Algorithm 1 evaluates n_k subgradients in iteration k, the cumulative number of subgradient evaluations up to iteration K is $\sum_{k\le K} n_k$. As no growth rate is imposed in Theorem 1, we may take, for instance, $n_k := \max\{1, \lceil \log k \rceil\}$ or $n_k = \lceil k^\alpha \rceil$, $\alpha \in (0, 1)$, which are nondecreasing and divergent. As summarized in Table 1, the schedules mandated by the frameworks discussed above force larger budgets.

Covered by	Sampling schedule	$\sum_{k\le K} n_k$
[37]	$n_k \gtrsim k^3 \ln k$	$\mathcal{O}(K^4 \ln K)$
[27, 30, 45]	$n_k \gtrsim k^n$, $n \in \{1, 2\}$	$\mathcal{O}(K^{1+n})$
This work	$n_k = \lceil k^\alpha \rceil$, $\alpha \in (0, 1)$	$\mathcal{O}(K^{1+\alpha})$
This work	$n_k = \max\{1, \lceil \log k \rceil\}$	$\mathcal{O}(K \log K)$

Table 1: Comparison of cumulative subgradient evaluations for different sample-size regimes

5 Numerical illustrations

We illustrate the behavior of Algorithm 1 on the regularized minimum eigenvalue problem with integrand (8) for destabilization analysis. The goal is to identify a bounded and sparse perturbation of the system parameters that minimizes the expected smallest eigenvalue, starting from a nominal configuration where that eigenvalue is positive. As in Remark 2, we take $A(\xi, x) = B_0(x) + \sum_{j=1}^q w_j(\xi) B_j(x)$ with $B_j(x) = C_j + \sum_{i=1}^d x_i G_{j,i} + c_j S_j(x)^\top S_j(x)$ and $S_j(x) = \sum_{i=1}^d x_i H_{j,i}$, so that each B_j depends quadratically on the decision variable. We set $p = 20$, $d = 40$ and $q = 4$. The matrices C_j, $G_{j,i} = G_{j,i}^\top$ and $H_{j,i}$ are drawn at random, and the weights $w_j(\xi)$ are independent Student-t variables with 3 degrees of freedom. Thus $\ell(\xi)$ in (12) is random and unknown, the local Lipschitz moduli grow with $\|x\|$, and no global second-moment bound on the stochastic subgradients is available; see Remark 3. The regularizer is $\psi = \lambda \ell_0 + \iota_{[-1,1]^d}$ with $\lambda = 0.02$, and thus $\operatorname{dom} \psi$ is bounded as in Remark 6. Fig. 1 reports the relative error of the function values (using a reference value computed by SAA with linesearch until $\|x^{k+1} - x^k\|/\gamma_k \le 10^{-10}$) and the stationarity measure $\|x^{k+1} - x^k\|/\gamma_k$, against the number of oracle calls, as medians over ten initial

points with interquartile bands. Algorithm 1 (left panels) decreases both quantities under all four sampling regimes, and outperforms PSS (middle panel). PSS with large constant stepsizes and growing mini-batches (right panels) are competitive in function values only, while tuned constant stepsizes are more competitive in the stationarity measure alone but tend to oscillate more. The line search reaches a better accuracy without any tuning and without knowledge of the curvature modulus, keeping the stepsize values away from zero.

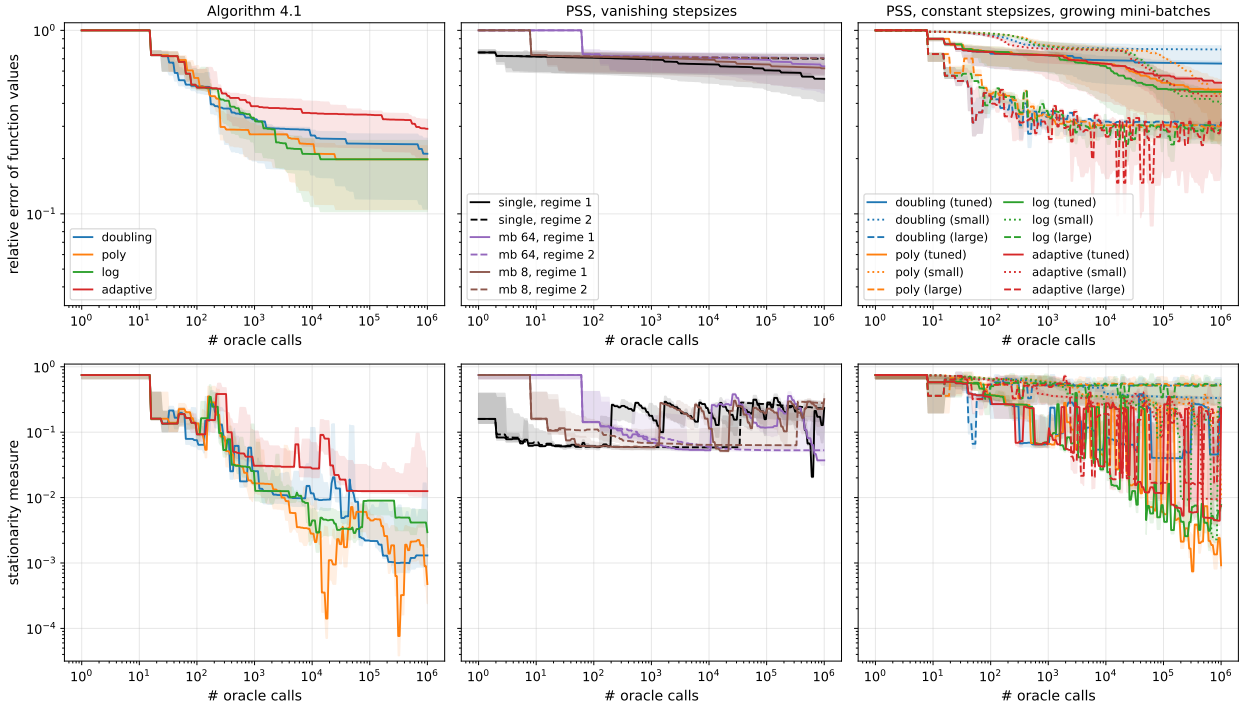


Figure 1: Comparison between Algorithm 1 and the proximal stochastic subgradient methods (PSS), with mini-batches (mb) of size 8 and 64, and without mini-batches (single). Algorithm 1 is run with $\sigma = 10^{-4}$, $\rho = 1/2$, and initial trial stepsize $\bar{\gamma}_0 = 1$. The middle panels show two vanishing stepsize regimes $\gamma_k = (k+1)^{-a}$, $a \in \{0.75, 1\}$ (regimes 1 and 2, resp.). The right panels show the PSS method with growing mini-batches for tuned, small and large constant stepsizes. Four sample-size regimes are implemented: doubling every $T = 10$ iterations, polynomial of degree $1/2$ (poly), logarithmic (log), and an adaptive variant that enlarges the sample size whenever the decrease achieved is below a threshold. The latter is a heuristic in the context of Algorithm 1 since the sample sizes are not deterministic.

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 (cf. [8]), 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 new 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. Our numerical illustrations suggest that Algorithm 1 with line search can be more robust than proximal stochastic subgradient methods with vanishing stepsizes, particularly in guaranteeing precision. A more in-depth numerical analysis is required to fully exploit the capabilities of our framework, including specific sample-size regimes (which, in turn, enables quantifying complexity bounds and rates of convergence), adaptive sampling strategies, and boosted line search procedures [3] to improve the computational cost per iteration.

Several questions remain open for future research. 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 variance-reduction mechanisms within the present nonsmooth nonconvex function setting and the extension of the variational inequality framework developed in previous works [33, 34, 35].

Acknowledgments

Funding. 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 grants Fondecyt Regular 1240335, Fondecyt Regular 1240120, and Fondecyt Regular 1261728. The fourth author was partially supported by grants PID2022-136399NB-C21 and PID2025-168246NB-I00 funded by ERDF/EU and by MICIU/AEI/10.13039/501100011033. The last two authors were supported by grant CIGE/2026/80 by Conselleria d'Educació, Cultura i Universitats, Generalitat Valenciana.

References

[1] W. S. van Ackooij and W. L. de Oliveira. "Methods of Nonsmooth Optimization in Stochastic Programming". In: *International Series in Operations Research and Management Science* (2025).

[2] F. J. Aragón-Artacho, R. Campoy, P. Pérez-Aros, and D. Torregrosa-Belén. "Nonmonotone subgradient methods based on a local descent lemma". In: *Comput. Optim. Appl.* (2026).

[3] F. J. Aragón-Artacho, P. Pérez-Aros, and D. Torregrosa-Belén. "The Boosted Double-proximal Subgradient Algorithm for nonconvex optimization". In: *Math. Program.* 214.1-2, Ser. A (2025), pp. 491–537.

[4] L. Armijo. "Minimization of functions having Lipschitz continuous first partial derivatives". In: *Pacific J. Math.* 16 (1966), pp. 1–3.

[5] Z. Artstein and R. J.-B. Wets. "Consistency of minimizers and the SLLN for stochastic programs". In: *J. Convex Anal.* 2.1-2 (1995), pp. 1–17.

[6] F. Atenas. "Understanding the Douglas–Rachford splitting method through the lenses of Moreau-type envelopes". In: *Comput. Optim. Appl.* 90.3 (2025), pp. 881–910.

[7] F. Atenas, A. Jofré, and P. Pérez-Aros. "Lower- and upper-C^2 integrands: a structured class of functions for stochastic nonconvex optimization". In: *arXiv preprint arXiv:2610.37629* (2026).

[8] F. Atenas, C. Sagastizábal, P. J. Silva, and M. Solodov. "A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods". In: *SIAM J. Optim.* 33.1 (2023), pp. 89–115.

[9] F. Bach. *Learning theory from first principles*. MIT Press, 2024.

[10] P. Baldi. *Probability—an introduction through theory and exercises*. Universitext. Springer, Cham, [2023] ©2023, pp. ix+389.

[11] M. Benaïm, J. Hofbauer, and S. Sorin. "Stochastic approximations and differential inclusions". In: *SIAM J. Control Optim.* 44.1 (2005), pp. 328–348.

[12] P. Bianchi, W. Hachem, and S. Schechtman. "Convergence of constant step stochastic gradient descent for non-smooth non-convex functions". In: *Set-Valued Var. Anal.* 30.3 (2022), pp. 1117–1147.

[13] V. I. Bogachev. *Measure theory. Vol. I, II*. Springer-Verlag, Berlin, 2007, Vol. I: xviii+500 pp., Vol. II: xiv+575.

[14] J. Bolte, T. Le, and E. Pauwels. "Subgradient sampling for nonsmooth nonconvex minimization". In: *SIAM J. Optim.* 33.4 (2023), pp. 2542–2569.

[15] L. Bottou, F. E. Curtis, and J. Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM Rev.* 60.2 (2018), pp. 223–311.

[16] F. H. Clarke. *Optimization and nonsmooth analysis*. Second. Vol. 5. Classics in Applied Mathematics. SIAM, Philadelphia, PA, 1990, pp. xii+308.

[17] F. H. Clarke, R. J. Stern, and P. R. Wolenski. "Proximal smoothness and the lower-C^2 property". In: *J. Convex Anal.* 2.1-2 (1995), pp. 117–144.

[18] T. Cour, B. Sapp, and B. Taskar. "Learning from Partial Labels". In: *J. Mach. Learn. Res.* 12.42 (2011), pp. 1501–1536.

[19] D. Davis and D. Drusvyatskiy. "Graphical convergence of subgradients in nonconvex optimization and learning". In: *Math. Oper. Res.* 47.1 (2022), pp. 209–231.

[20] D. Davis and D. Drusvyatskiy. "Stochastic model-based minimization of weakly convex functions". In: *SIAM J. Optim.* 29.1 (2019), pp. 207–239.

[21] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. "Stochastic subgradient method converges on tame functions". In: *Found. Comput. Math.* 20.1 (2020), pp. 119–154.

[22] D. Davis and B. Grimmer. "Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems". In: *SIAM J. Optim.* 29.3 (2019), pp. 1908–1930.

[23] J. C. Duchi and F. Ruan. "Stochastic methods for composite and weakly convex optimization problems". In: *SIAM J. Optim.* 28.4 (2018), pp. 3229–3259.

[24] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II*. Springer-Verlag, New York, 2003, i–xxxiv, 625–1234 and II1–II57.

[25] J. Fan and R. Li. "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *J. Amer. Statist. Assoc.* 96.456 (2001), pp. 1348–1360.

[26] I. Fatkhullin, N. He, and Y. Hu. "Stochastic optimization under hidden convexity". In: *SIAM J. Optim.* 35.4 (2025), pp. 2544–2571.

[27] C. Geiersbach and T. Scarinci. "Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces". In: *Comput. Optim. Appl.* 78.3 (2021), pp. 705–740.

[28] S. Ghadimi and G. Lan. "Accelerated gradient methods for nonconvex nonlinear and stochastic programming". In: *Math. Program.* 156.1-2 (2016), pp. 59–99.

[29] S. Ghadimi and G. Lan. "Stochastic first- and zeroth-order methods for nonconvex stochastic programming". In: *SIAM J. Optim.* 23.4 (2013), pp. 2341–2368.

[30] S. Ghadimi, G. Lan, and H. Zhang. "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization". In: *Math. Program.* 155.1-2 (2016), pp. 267–305.

[31] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Vol. 1730. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2000, pp. x+230.

[32] T. Homem-de-Mello. "On rates of convergence for stochastic optimization problems under non-IID sampling". In: *SIAM J. Optim.* 19.2 (2008), pp. 524–551.

[33] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. "Extragradient method with variance reduction for stochastic variational inequalities". In: *SIAM J. Optim.* 27.2 (2017), pp. 686–724.

[34] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. "Variance-based extragradient methods with line search for stochastic variational inequalities". In: *SIAM J. Optim.* 29.1 (2019), pp. 175–206.

[35] A. N. Iusem, A. Jofré, and P. Thompson. "Incremental constraint projection methods for monotone stochastic variational inequalities". In: *Math. Oper. Res.* 44.1 (2019), pp. 236–263.

[36] Z. Jia and B. Grimmer. "First-order methods for nonsmooth nonconvex functional constrained optimization with or without Slater points". In: *SIAM J. Optim.* 35.2 (2025), pp. 1300–1329.

[37] A. Jofré and P. Thompson. "On variance reduction for stochastic smooth convex optimization with multiplicative noise". In: *Math. Program.* 174.1-2 (2019), pp. 253–292.

[38] C. Kanzow, J. Krüger, and L. Lehmann. *Projected Subgradient Methods for a Class of Nonsmooth and Nonconvex Optimization Problems*. 2026.

[39] C. Kanzow and L. Lehmann. "A Nonmonotone Descent Method for Optimization Problems Defined by Upper-$\mathcal{C}^2$ Functions over Submanifolds". In: (2026). Preprint, available at: arXiv:2605.26909.

[40] J. Kiefer and J. Wolfowitz. "Stochastic Estimation of the Maximum of a Regression Function". In: *Ann. Math. Statist.* 23.3 (1952), pp. 462–466.

[41] A. J. King and R. T. Rockafellar. "Asymptotic theory for solutions in statistical estimation and stochastic programming". In: *Math. Oper. Res.* 18.1 (1993), pp. 148–162.

[42] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello. "The sample average approximation method for stochastic discrete optimization". In: *SIAM J. Optim.* 12.2 (2002), pp. 479–502.

[43] V. Krätschmer. "Nonasymptotic upper estimates for errors of the sample average approximation method to solve risk-averse stochastic programs". In: *SIAM J. Optim.* 34.2 (2024), pp. 1264–1294.

[44] G. Lan. "An optimal method for stochastic composite optimization". In: *Math. Program.* 133.1-2, Ser. A (2012), pp. 365–397.

[45] H. A. Le Thi, V. N. Huynh, T. Pham Dinh, and H. P. H. Luu. "Stochastic difference-of-convex-functions algorithms for nonconvex programming". In: *SIAM J. Optim.* 32.3 (2022), pp. 2263–2293.

[46] J. Lei and U. V. Shanbhag. "Asynchronous variance-reduced block schemes for composite non-convex stochastic optimization: block-specific steplengths and adapted batch-sizes". In: *Optim. Methods Softw.* 37.1 (2022), pp. 264–294.

[47] J. Lei and U. V. Shanbhag. "Variance-reduced accelerated first-order methods: central limit theorems and confidence statements". In: *Math. Oper. Res.* 50.2 (2025), pp. 1364–1397.

[48] Z. Li. "Simple and optimal stochastic gradient methods for nonsmooth nonconvex optimization". In: *J. Mach. Learn. Res.* 23.239 (2022), pp. 1–61.

[49] P.-L. Loh and M. J. Wainwright. "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima". In: *J. Mach. Learn. Res.* 16 (2015), pp. 559–616.

[50] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. “Online dictionary learning for sparse coding”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML ’09. Association for Computing Machinery, 2009, pp. 689–696.

[51] M. R. Metel and A. Takeda. “Stochastic proximal methods for non-smooth non-convex constrained sparse optimization”. In: *J. Mach. Learn. Res.* 22.115 (2021), pp. 1–36.

[52] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. “Robust stochastic approximation approach to stochastic programming”. In: *SIAM J. Optim.* 19.4 (2009), pp. 1574–1609.

[53] L. M. Nguyen, K. Scheinberg, and T. H. Tran. “Stochastic ISTA/FISTA adaptive step search algorithms for convex composite optimization”. In: *J. Optim. Theory Appl.* 205.1 (2025), p. 10.

[54] C. Paquette and K. Scheinberg. “A stochastic line search method with expected complexity analysis”. In: *SIAM J. Optim.* 30.1 (2020), pp. 349–376.

[55] P. Pérez-Aros and D. Torregrosa-Belén. “Randomized block proximal method with locally Lipschitz continuous gradient”. In: *Numerical Algorithms* (2026). Published online.

[56] N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh. “ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization”. In: *J. Mach. Learn. Res.* 21.110 (2020), pp. 1–48.

[57] B. T. Polyak and A. B. Juditsky. “Acceleration of stochastic approximation by averaging”. In: *SIAM J. Control Optim.* 30.4 (1992), pp. 838–855.

[58] S. Pougkakiotis and D. Kalogerias. “A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization”. In: *SIAM J. Sci. Comput.* 45.5 (2023), A2679–A2702.

[59] H. Robbins and D. Siegmund. “A convergence theorem for non negative almost supermartingales and some applications”. In: *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)*. Academic Press, New York-London, 1971, pp. 233–257.

[60] H. Robbins and S. Monro. “A Stochastic Approximation Method”. In: *Ann. Math. Statist.* 22.3 (1951), pp. 400–407.

[61] R. T. Rockafellar. “Favorable classes of Lipschitz continuous functions in subgradient optimization”. In: *In: Non-Differentiable Optimization, E. Nurminski (ed.), Pergamon Press, New York* (1982).

[62] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Vol. 317. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1998, pp. xiv+733.

[63] P. J. Rousseeuw and K. Van Driessen. “Computing LTS regression for large data sets”. In: *Data Min. Knowl. Disc.* 12.1 (2006), pp. 29–45.

[64] J. O. Royset and R. J.-B. Wets. *An optimization primer*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, [2021] ©2021, pp. xviii+676.

[65] F. Ruan. “On the Uniform Convergence of Subdifferentials in Stochastic Optimization and Learning”. In: *Math. Oper. Res.* 51.3 (2026), pp. 2323–2353.

[66] A. Shapiro. “Asymptotic behavior of optimal solutions in stochastic programming”. In: *Math. Oper. Res.* 18.4 (1993), pp. 829–845.

[67] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming—modeling and theory*. Third. Vol. 28. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2021, pp. xv+525.

[68] L. Tian and J. O. Royset. “Lipschitzian SLLNs for random functions”. In: *arXiv preprint arXiv:2607.20411* (2026).

[69] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. "A hybrid stochastic optimization framework for composite nonconvex optimization". In: *Math. Program.* 191.2 (2022), pp. 1005–1071.

[70] J. Wang and C. G. Petra. "A sequential quadratic programming algorithm for nonsmooth problems with upper-$\mathcal{C}^2$ objective". In: *SIAM J. Optim.* 33.3 (2023), pp. 2379–2405.

[71] C.-H. Zhang. "Nearly unbiased variable selection under minimax concave penalty". In: *Ann. Statist.* 38.2 (2010), pp. 894–942.

[72] J. Zhang and L. Xiao. "Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization". In: *Math. Program.* 195.1 (2022), pp. 649–691.