Stochastic Gradient Descent Revisited

Azar Louzi

Stochastic Gradient Descent Revisited

Azar Louzi*

November 27, 2024

Abstract

Stochastic gradient descent (SGD) algorithms have been a go-to solution for nonconvex stochastic optimization problems arising in machine learning. Their theory however often requires a strong framework to guarantee convergence properties. We hereby present a full scope convergence study of nonconvex SGD, including weak convergence, function-value convergence and global convergence, and also provide subsequent convergence rates and complexities, all under relatively mild conditions in comparison with literature.

Keywords. stochastic gradient descent ; weak convergence ; function-value convergence ; local Łojasiewicz condition ; convergence rate ; complexity ; global convergence

MSC. 90C15 ; 90C26 ; 90C60

Introduction

The advent of artificial intelligence (AI) has been rendered possible by the spectacular acceleration of computing chip capacity over the last few decades, and has driven a technological revolution that has not spared any aspect of life, including healthcare, supply chain management, social media, etc. AI describes a set of machine learning methods that abandon any form of structural representation of data and look instead into uncovering data patterns to produce probabilistic relationships between input and output quantities of interest.

While it has significantly improved people's standards of living, AI has nevertheless engendered many operational risks (e.g. by producing undesirable or unexpected outcomes) as well as systemic risks (e.g. the *"Flash Crash"*, whereby a blue-chip company's share price suddenly plummeted and bounced back in the span of minutes [KL13]). To better manage, prevent and mitigate such risks, some level of mathematical insight must be brought in to shed light onto the inner workings of AI, in order to allow practitioners and regulators alike to act upon it in order to increase its efficiency and curb its shortcomings.

SGD is the engine of AI, making it a natural stepping stone toward mathematically explaining AI. Indeed, to capture their intricacies, machine learning problems are often modeled using wide and highly parametrized neural networks [GBC16], which are then solved using SGD or an adaptive variant thereof, namely Adagrad, Adadelta, RMSProp, Adamax or Adam [Rud17]. To approximate a stationary point of a given loss landscape (also referred to as objective or cost function [LZB22; AL24; AMA05]), SGD recursively spawns a trajectory of iterates by factoring in, at each step, a stochastic gradient modulated by a positive learning rate. Whereas classical SGD literature provides convergence guarantees and convergence rates within a (strongly) convex framework [Duf96; BV04; RM51], machine learning models are often highly nonconvex and require new SGD frameworks to better understand and parametrize them. An adequate parametrization can drastically cut down model training time by eliminating some fine-tuning

*Université Paris Cité, CNRS. Laboratoire de Probabilités, Statistique et Modélisation. azar.louzi@lpsm.paris

overhead.

Recent literature attempting to comprehend nonconvex SGD from a theoretical standpoint presents convergence guarantees and convergence rates under assumptions that may be difficult to verify in practice. The frequently studied SGD convergence modes include weak convergence, function-value convergence and global convergence. Consider a differentiable loss landscape F to minimize and related iterates $(\theta_n)_{n\geq 0}$ produced by SGD. By weak convergence, we refer to the almost sure vanishing of $(\nabla F(\theta_n))_{n\geq 0}$ [AMA05], by function-value convergence, the almost sure convergence of $(F(\theta_n))_{n\geq 0}$ [Dav+20], and by global convergence, the almost sure convergence of the iterates themselves to a stationary point [LP15; Let21]. The speed at which each convergence mode occurs, when quantifiable, helps infer optimal iteration amounts that achieve some prescribed tolerance.

Following the salutary work [Łoj84] on gradient flow convergence for analytical functions, [AMA05] introduced a new framework for studying nonconvex gradient descent (GD), termed the Łojasiewicz condition. It has found a large traction in the optimization community as it describes a large class of nonconvex loss landscapes. Under this condition, for a stationary point $\theta_\star$ of a loss landscape F ($\nabla F(\theta_\star) = 0$), one has $\zeta\|\nabla F(\theta)\| \geq |F(\theta) - F(\theta_\star)|^\beta$ on an open neighborhood of $\theta_\star$, for some constant $\zeta > 0$ and exponent $\beta \in (0, 1)$. Initially introduced to prove single-point convergence of GD-type algorithms, it has also found success in deriving GD convergence rates [AB09; FGP15]. Other variants of the original inequality have been explored as well. When $\beta = \frac{1}{2}$, F is said to be Polyak-Łojasiewicz (PŁ) [Pol63]. When the property is satisfied at every point in space, F is dubbed globally Łojasiewicz [GG24]. A variant thereof, known as the Kurdyka-Łojasiewicz (KŁ) condition, stipulates that $\varphi'(F(\theta) - F(\theta_\star))\|\nabla F(\theta)\| \geq 1$ on an open neighborhood of a stationary point $\theta_\star$, where φ is a desingularization function that is increasing, concave and differentiable, with $\varphi(0) = 0$. This property is proven to hold for continuously differentiable subanalytic functions [Kur98]. Although it is considered to be a slight generalization of the Łojasiewicz condition [DK24], reverting to the Łojasiewicz condition, by considering e.g. $\varphi(x) = x^{1-\beta}$ [FGP15], is often necessary to derive convergence rates.

[Lei+20; DK24; CFR23; Dav+20] provide weak and function-value convergence guarantees under various assumptions on the loss landscape. [Lei+20] proves function-value convergence by lifting the SGD formalism into a reproducing kernel Hilbert space, where the loss landscape is assumed to be Hölder-differentiable. [DK24] studies biased SGD for locally Hölder-differentiable loss landscapes, presupposing the SGD iterates almost surely bounded. [Dav+20] focuses on locally Lipschitz functions with possible singularities, but relies on the presumption that SGD iterates are almost surely bounded.

Recent developments on function-value convergence rate build upon varying conditions [Lei+20; AL24; DK24; Wei+24]. [Lei+20] provides convergence speeds in mean under a global PŁ condition. [AL24] extends the framework of [Cha22] to SGD by supposing a local PŁ condition in a region of the space, then derives an exponential convergence rate on the condition that the iterates remain close to a local minimum. In a local Łojasiewicz setting, [DK24] elicits a convergence rate in mean conditionally to suitable events controlling the stochastic gradients. In an identical setting, [Wei+24] provides a convergence rate in high probability, conditionally to the iterates remaining confined around some local minimum.

Global convergence is investigated in [DK24; CFR23; QMM24] from different perspectives. [DK24] proves almost sure iterate convergence for locally Lipschitz-differentiable loss landscapes under a local Łojasiewicz condition, by controlling the behavior of the stochastic gradients. [CFR23] assumes a local KŁ condition on a compact set coupled with a strong growth property. [QMM24] studies momentum (SGD) in a local Łojasiewicz setting, by assuming bounded conditional variances of the stochastic gradients.

[Liu+23] derives high probability convergence rates for SGD iterates when initialized close

enough to a global minimum, under an aiming and an interpolation conditions. [GP23] obtains $L^{2p}(\mathbb{P})$-convergence rates for SGD iterates, $p \geq 1$, for globally Łojasiewicz loss landscapes with Łojasiewicz exponent $\beta \leq \frac{1}{2}$, by supposing uniformly bounded conditional polynomial-exponential moments on the stochastic gradient noises.

In this work,

- Thorem 1.5 obtains weak and function-value convergences under a relaxation of the smoothness framework and typical learning rate behavior without requiring iterate boundedness;
- Theorem 1.7 characterizes the topology of the iterates' accumulation points under a plain coercivity argument;
- Theorem 1.17 derives high probability function-value convergence speeds within both a local and a global Łojasiewicz frameworks;
- Corollary 1.21 quantifies subsequent complexities and recovers optimal iteration amounts;
- Theorem 1.22 and Corollaries 1.24 and 1.26 deduce similar results for single-point limit convergence of SGD iterates.

This paper is organized as follows. Section 1 establishes a baseline SGD framework and enunciates our standing convergence results, followed up with related comments and discussions. The technical proofs of our main results are delineated in Sections 2–4.

1 Stochastic Gradient Descent

We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that is rich enough to support all the ensuing random variables. In a given vector space $\mathbb{R}^p$, $p \geq 1$, $\|\cdot\|$ denotes the L^2 vector norm. Given a measurable random variable $Z\colon (\Omega, \mathcal{F}) \to (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, $p \geq 1$, we write $Z \in L^q(\mathbb{P})$, $q > 0$, to signify that $\mathbb{E}[\|Z\|^q] < \infty$.

We are interested in the stochastic optimization problem

$$\min_{\theta \in \mathbb{R}^m} \big\{F(\theta) = \mathbb{E}[f(X, \theta)]\big\}, \tag{SO}$$

where X is an $\mathbb{R}^d$-valued random variable, $f\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is a function such that $f(X, \theta) \in L^1(\mathbb{P})$, $\theta \in \mathbb{R}^m$, and $d, m \geq 1$. f is often referred to as the loss or utility function.

$\mathcal{H}$1. F is lower bounded, i.e. $\inf F > -\infty$, differentiable, and its derivative ∇F is (L, α)-Hölder continuous, $L > 0$, $\alpha \in (0, 1]$, i.e.

$$\|\nabla F(\theta') - \nabla F(\theta)\| \leq L\|\theta' - \theta\|^\alpha, \quad \theta, \theta' \in \mathbb{R}^m.$$

$\mathcal{H}1$ relaxes the smoothness condition arising when $\alpha = 1$. In spite of the continuous differentiability of F, the function $\theta \in \mathbb{R}^m \mapsto f(x, \theta)$, $x \in \mathbb{R}^d$, does not need to be differentiable everywhere [Tal01].

Example 1.1. Take the μ-quantile regression problem [RU00]: $\min_\theta\{F(\theta) = \mathbb{E}[f(X, \theta)]\}$, where $f(x, \theta) = \theta + \frac{(x-\theta)^+}{1-\mu}$, $x, \theta \in \mathbb{R}$, with $\mu \in (0, 1)$. Let $x \in \mathbb{R}$. $\theta \in \mathbb{R} \mapsto f(x, \theta)$ is differentiable for $\theta \neq x$ and $\partial_\theta f(x, \theta) = 1 - \frac{\mathbb{1}_{x \geq \theta}}{1-\mu}$, $\theta \in \mathbb{R}$, which is discontinuous at x. Under suitable conditions, notably the continuity of the cdf $F_X(\theta) \coloneqq \mathbb{P}(X \leq \theta)$, $\theta \in \mathbb{R}$, F is continuously differentiable, convex and coercive, with $F'(\theta) = \frac{F_X(\theta)-\mu}{1-\mu}$, $\theta \in \mathbb{R}$ [BFP09, Proposition 2.1]. If X admits a bounded pdf f_X, then F is twice differentiable, with $\|F''\|_\infty = \frac{\|f_X\|_\infty}{1-\mu} =: L < \infty$. Hence F is L-Lipschitz-differentiable and fulfills $\mathcal{H}1$.

[Lei+20, Assumption 1] assumes rather that $\theta \in \mathbb{R}^m \mapsto f(x,\theta)$, $x \in \mathbb{R}^d$, is (L,α)-Hölder-differentiable. In practice, L may depend on x, say, L_x, so that the boundedness of $(L_x)_{x\in\mathbb{R}^m}$ underpins such assumption. Since many machine learning problems suppose X compactly supported, the boundedness of $(L_x)_{x\in\mathcal{X}}$ seems plausible. Under [Lei+20, Assumption 1], supposing $\nabla F(\theta) = \mathbb{E}[\nabla f(X,\theta)]$, $\theta \in \mathbb{R}^m$, via Jensen's inequality,

$$\|\nabla F(\theta') - \nabla F(\theta)\| \le \mathbb{E}\big[\|\nabla_\theta f(X,\theta') - \nabla_\theta f(X,\theta)\|\big] \le L\|\theta' - \theta\|^\alpha, \quad \theta,\theta' \in \mathbb{R}^m,$$

thus retrieving the last part of $\mathcal{H}1$.

Assume we have access to innovations $(X_n)_{n\ge1} \overset{\text{iid}}{\sim} X$ and to an oracle $g : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ providing a noisy approximation of ∇F.

$\mathcal{H}2$. For some constant $C > 0$,

$$\mathbb{E}[g(X,\theta)] = \nabla F(\theta), \quad \mathbb{E}\big[\|g(X,\theta)\|^2\big] \le C\big((F(\theta) - \inf F) + 1\big), \quad \theta \in \mathbb{R}^m.$$

The first half of $\mathcal{H}2$ describes an unbiased stochastic gradient. Its second half generalizes the variance transfer formula [GG24, Lemma 4.20].

Example 1.2. In Example 1.1, $|\partial_\theta f(x,\theta)| \le 1 \vee \frac{\mu}{1-\mu}$, $x,\theta \in \mathbb{R}$, hence fulfilling $\mathcal{H}2$.

[Wei+24, Assumption 2.4] (dubbed the ABC or the expected smoothness assumption, in reference to [KR23, Assumption 2]) alternatively postulates that

$$\mathbb{E}[\|g(X,\theta)\|^2] \le C\big((F(\theta) - \inf F) + \|\nabla F(\theta)\|^2 + 1\big),$$

for some constant $C > 0$. However, under $\mathcal{H}1$, by Lemma A.1(ii) and Young's inequality with the adjoint exponents $\big(\frac{1+\alpha}{2\alpha}, \frac{1+\alpha}{1-\alpha}\big)$ if $\alpha \ne 1$,

$$\|\nabla F(\theta)\|^2 \le C(F(\theta) - \inf F)^{\frac{2\alpha}{1+\alpha}} \le C\big((F(\theta) - \inf F) + 1\big) < \infty,$$

for some constant $C > 0$ that may change from line to line. This inequality holds also if $\alpha = 1$. We thus retrieve $\mathcal{H}2$.

$\mathcal{H}2$ is more generic than the interpolation condition $\mathbb{E}[\|g(X,\theta)\|^2] \le C(F(\theta) - \inf F)$, $\theta \in \mathbb{R}^m$ [Liu+23, §1.2]. We refer to Section 1.3 for extensive comments on the limitations posed by this condition.

Under [Lei+20, Assumption 1], which stipulates that $\theta \in \mathbb{R}^m \mapsto f(x,\theta)$, $x \in \mathbb{R}^d$, is (L,α)-Hölder-differentiable, if $\inf_{x,\theta} f > -\infty$, then, by Lemma A.1(ii) and Young's inequality with the adjoint exponents $\big(\frac{1+\alpha}{2\alpha}, \frac{1+\alpha}{1-\alpha}\big)$ if $\alpha \ne 1$,

$$\mathbb{E}\big[\|\nabla_\theta f(X,\theta)\|^2\big] \le C\mathbb{E}\big[\big(f(X,\theta) - \inf f\big)^{\frac{2\alpha}{1+\alpha}}\big] \le C\big((F(\theta) - \inf F) + 1\big) < \infty,$$

for some constant $C > 0$ that may change from line to line. This remains true if $\alpha = 1$. The above inequality matches $\mathcal{H}2$.

Define the stochastic gradient oracles

$$g_n(\theta) \coloneqq g(X_n,\theta), \quad \theta \in \mathbb{R}^m,\ n \ge 1.$$

SGD leverages the stochastic gradient oracles $(g_n)_{n\ge1}$ to approximate solutions of (SO) via the dynamics

$$\theta_n = \theta_{n-1} - \gamma_n g_n(\theta_{n-1}), \quad n \ge 1, \tag{SGD}$$

driven by the innovations $(X_n)_{n\ge1} \overset{\text{iid}}{\sim} X$, initialized at some $\theta_0 \in L^2(\mathbb{P})$ that is independent of $(X_n)_{n\ge1}$, and advancing at the positive learning rate $(\gamma_n)_{n\ge1}$ such that $\sum_{n=1}^\infty \gamma_n = \infty$ and $\lim_{n\to\infty} \gamma_n = 0$.

Remark 1.3. Under $\mathcal{H}1$, using a first order Taylor-Lagrange expansion,

$$|F(\theta)| = \left|F(0) + \theta^\top \int_0^1 \nabla F(t\theta)\mathrm{d}t\right| \leq |F(0)| + \|\nabla F(0)\| + \frac{L\|\theta\|^{1+\alpha}}{1+\alpha} \leq C(1+\|\theta\|^2), \quad \theta \in \mathbb{R}^m, \tag{1.1}$$

for some constant $C > 0$. Thus, via the (SGD) recursion, $\mathcal{H}2$ and Jensen's inequality, it ensues iteratively that $\mathbb{E}[|F(\theta_n)|] + \mathbb{E}[\|\theta_n\|^2] + \mathbb{E}[\|g_n(\theta_{n-1})\|^2] + \mathbb{E}[\|\nabla F(\theta_{n-1})\|^2] < \infty$, $n \geq 1$.

$\mathcal{H}3$. $(\gamma_n)_{n\geq 1}$ is a positive sequence such that

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n^{1+\alpha} < \infty.$$

Note that the Hölder exponent α from $\mathcal{H}1$ shows up in $\mathcal{H}3$. This assumption appears for instance in [Lei+20, Theorem 4(a)]. It translates a desirable trait for (SGD) to find a trade-off between exploration of the parameter space and exploitation of the loss landscape geometry.

Example 1.4. Typical sequences adhering to $\mathcal{H}3$ write as

$$\gamma_n = \frac{\gamma_0}{(n+c)^s}, \quad \gamma_0 > 0,\ c \geq 0,\ s \in \Big(\frac{1}{1+\alpha}, 1\Big],\ n \geq 1. \tag{1.2}$$

Another possible form is

$$\gamma_n = \frac{\gamma_0}{(n+c)\ln^s(1+c'+n)}, \quad \gamma_0 > 0,\ c, c' \geq 0,\ s \in \Big(\frac{1}{1+\alpha}, 1\Big],\ n \geq 1.$$

[Lei+20, Corollary 3(b)] also mentions

$$\gamma_n = \frac{\gamma_0}{((n+c)\ln^s(1+c'+n))^{1/(1+\alpha)}}, \quad \gamma_0 > 0,\ c, c' \geq 0,\ s > 1,\ n \geq 1.$$

1.1 Convergence Analysis

In this section, we seek convergence guarantees for (SGD). Given that global optimization of a nonconvex loss landscape is NP-hard [Dan+22, §2.1], weak convergence ($\nabla F(\theta_n) \to 0$ as $n \to \infty$) seems to be a minimal requirement for (SGD) to be worthwhile. Convergence of the function values $(F(\theta_n))_{n\geq 0}$ is another useful imperative for (SGD), the reason being twofold. First, convergence of $(\theta_n)_{n\geq 0}$ is simply not guaranteed in the nonconvex setting, even when $(F(\theta_n))_{n\geq 0}$ does converge [AMA05]. Second, when a limit $\theta_\star \in \mathbb{R}^m$ to $(\theta_n)_{n\geq 0}$ exists, it is often the case that the behavior of the function value gap $F(\theta_n) - F(\theta_\star)$ is the one reflected in that of the iterate gap $\theta_n - \theta_\star$ rather than the other way around [LZ24, §2.1].

Throughout, $(\mathcal{F}_n)_{n\geq 0}$ designates the filtration recursively spawned by θ_0 and the innovations $(X_n)_{n\geq 1}$, and $\mathbb{E}_n$ denotes the conditional expectation $\mathbb{E}[\,\cdot\,|\mathcal{F}_n]$, $n \geq 0$. We hereby provide our standing convergence results and follow up with discussions thereon. Related proofs are postponed to Section 2.

Theorem 1.5. *Under $\mathcal{H}1$–3, the function values $(F(\theta_n))_{n\geq 0}$ converge $\mathbb{P}$-as to a real-valued random variable $F_\star \in L^1(\mathbb{P})$ and $\nabla F(\theta_n) \to 0$ $\mathbb{P}$-as as $n \to \infty$.*

Note that the above result is independent from any coercivity or iterate boundedness assumption. The $\mathbb{P}$-as convergence of $(F(\theta_n))_{n\geq 0}$ is obtained by applying the Siegmund-Robbins lemma [RS71] to the quasisupermartingale property (2.2). The weak convergence of (SGD) expands on some arguments used in [Lei+20, Theorem 2(c)] for $L^1(\mathbb{P})$ vanishing of $(\nabla F(\theta_n))_{n\geq 0}$ when $\alpha = 1$. It builds upon the premise that it is impossible to simultaneously fulfill $\|\nabla F(\theta_n) -$

$\nabla F(\theta_{n-1})\| \to 0$ as $n \to \infty$ and $\sum_{n=1}^{\infty} \gamma_n \|\nabla F(\theta_{n-1})\|^2 < \infty$, unless $\limsup_{n\to\infty} \|\nabla F(\theta_n)\| = 0$ $\mathbb{P}$-as.

[Dav+20, Theorem 4.2] states the $\mathbb{P}$-as function-value and weak convergences of (SGD) for locally Lipschitz loss landscapes (i.e. almost everywhere differentiable, by Rademacher's theorem), under the assumption that the iterates are $\mathbb{P}$-as bounded. Such an assumption de facto implies that the gradients $(\nabla F(\theta_n))_{n\geq 0}$ are $\mathbb{P}$-as bounded. [Mer+20, §4.1] signals that presupposing iterate boundedness renders the ensuing convergence results of little practical value because (SGD) trajectories can in fact escape to infinity, as already observed for gradient flows [AMA05, Theorem 2.2]. Our convergence guarantees lie within the continuity of recent efforts to eliminate such an assumption [Mer+20; QMM24].

[DK24, Theorem 1.1] reaches similar results for biased (SGD) on locally Hölder-differentiable loss landscapes, conditionally to the iterates remaining bounded and the conditional q-th moments of the stochastic gradient noises being deterministically quantified [DK24, Assumption $\mathbb{M}^{\sigma,q}$]. In our notation, the latter condition writes

$$\mathbb{E}_{n-1}\big[\|g_n(\theta_{n-1}) - \nabla F(\theta_{n-1})\|^q\big] \leq v_n, \quad n \geq 1, \tag{1.3}$$

for some $q \geq 1$ and a deterministic positive sequence $(v_n)_{n\geq 1}$ satisfying some summability assumptions. Let us examine the latter condition more closely. Consider $f(x,\theta) = \frac{1}{2}x^2\theta^2$, $F(\theta) = \frac{1}{2}\mathbb{E}[X^2]\theta^2$, $x, \theta \in \mathbb{R}$, where $X \in L^2(\mathbb{P})$ is a real-valued random variable. Then $g(x,\theta) = \partial_\theta f(x,\theta) = x^2\theta$, $F'(\theta) = \mathbb{E}[X^2]\theta$, $x, \theta \in \mathbb{R}$, and, by Jensen's inequality,

$$\mathbb{E}\big[\big|\partial_\theta f(X,\theta) - F'(\theta)\big|^q\big] \geq \mathbb{E}\big[\big|\partial_\theta f(X,\theta) - F'(\theta)\big|\big]^q = \mathbb{E}\big[\big|X^2 - \mathbb{E}[X^2]\big|\big]^q |\theta|^q, \quad \theta \in \mathbb{R},\ q \geq 1.$$

On this example, the assumption (1.3) entails that $(\theta_n)_{n\geq 0}$ are compactly supported. But in view of the update rule (SGD), so are the stochastic gradients $(g_n(\theta_{n-1}))_{n\geq 1}$. This limits the scope of applicability of the obtained convergence results. [DK24] further references [DK23, Lemma D.1] as a guarantee for iterate boundedness for coercive Lipschitz-differentiable loss landscapes, which however requires [DK24, Assumption $\mathbb{M}^{\sigma,q}$]. As for weak convergence, the technique used in [DK24] relies on the uniform continuity of ∇F on a compact set containing all of the iterates $(\theta_n)_{n\geq 0}$. This argument has roots in [BT00, Proposition 1] (c.f. also [Ora20, Lemma 1]) and can only be utilized if the iterates are guaranteed to be bounded.

[Lei+20, Theorem 2] shows the $\mathbb{P}$-as-function-value convergence via a quasisupermartingale formalism similar to (2.2), under an (L,α)-Hölder-differentiability assumption on $\theta \in \mathbb{R}^m \mapsto f(x,\theta)$, $x \in \mathbb{R}^m$. [Lei+20, Theorem 2(c)] proves that $\mathbb{E}[\|\nabla F(\theta_n)\|]$ vanishes as $n \to \infty$ for $\alpha = 1$, which solely ensures that $\nabla F(\theta_n) \to 0$ in probability as $n \to \infty$. [Lei+20, Theorem 4(a)] eventually proves weak convergence by additionally imposing a global PŁ condition.

Function-value convergence is ensured in [CFR23, Proposition 5.5] for $\alpha = 1$ via a quasisupermartingale formalism close to [Lei+20]. Weak convergence is treated in [CFR23, Proposition 6.3] under the lens of iterate convergence, by assuming the strong growth condition $\mathbb{E}_{n-1}[\|g_n(\theta_{n-1})\|^2] \leq C\|\nabla F(\theta_{n-1})\|^2$, $n \geq 1$, with $C > 0$. This is far from representing a generic learning problem (SO).

To be able to characterize the asymptotics of $(\theta_n)_{n\geq 0}$, we start by formulating a plain coercivity condition.

***H*4.** *F is coercive. That is, $F(\theta) \to \infty$ as $\|\theta\| \to \infty$.*

Lemma 1.6. *Under $\mathcal{H}1$–4, $(\theta_n)_{n\geq 0}$ is $\mathbb{P}$-as bounded.*

$\mathcal{H}4$ is devoid of any extra coercive regularity. Machine learning problems are customarily regularized to achieve generalization and to induce convexity. This practice often involves adding

a regularization term to the loss associated with the learning model: $f(x,\theta) = \ell(x,\theta) + \varphi(\theta) \geq \varphi(\theta)$, $x \in \mathbb{R}^d$, $\theta \in \mathbb{R}^m$, where ℓ is a nonnegative loss and φ is a coercive regularizer.

Lemma 1.6 retrieves the $\mathbb{P}$-as boundedness of the iterates $(\theta_n)_{n\geq 0}$ as a consequence rather than an assumption, inasmuch as the nominal SGD user can only check the satisfiability of $\mathcal{H}1$–4 beforehand.

In [Mer+20, Assumptions 2 & 3], the sublevels of F and $\|\nabla F\|$ are assumed to be bounded, which represents a strong coercive regularity on the loss landscape.

Hereafter, we denote $\nabla F^{-1}(0)$ the set of stationary points of F. The main convergence result of our analysis follows.

Theorem 1.7. *Assume $\mathcal{H}1$–4 hold. Then,*

- **(i)** *$\mathbb{P}$-as, the set $\mathcal{A}$ of accumulation points of $(\theta_n)_{n\geq 0}$ is a nonempty compact connected set, and every element of $\mathcal{A}$ is a stationary point of F;*
- **(ii)** *if $\nabla F^{-1}(0)$ is additionally at most countable, $(\theta_n)_{n\geq 0}$ converges $\mathbb{P}$-as to a stationary point of F.*

Theorem 1.7 is echoed in [CP17, Proposition 2] in a surrogate optimization setting. The compactness of the accumulation points' set is a direct consequence of the iterate boundedness. Its conntectedness however stems from Ostrowski's lemma [Ost74, Theorem 26.1], a classical result of topology based on the premise that it is impossible to simultaneously ensure that $\|\theta_n - \theta_{n-1}\| \to 0$ as $n \to \infty$ and that the set of accumulation points $\cap_{p\geq 0}\overline{\{\theta_n, n \geq p\}}$ of $(\theta_n)_{n\geq 0}$ write as a union of two disjoint compact sets [Gou20, §4, Exercise 6], thus mirroring the main argument for weak convergence.

1.2 Convergence Rate Analysis

In this section, we seek to quantify the convergence speed of $(F(\theta_n))_{n\geq 0}$ and assess the subsequent (SGD) complexity. Convergence rates of nonconvex approximation schemes are often derived under a Łojasiewicz-type condition [AB09; FGP15; DK24; Wei+24] as we put forward below. Related proofs are available in Section 3.

$\mathcal{H}$5. F is Łojasiewicz, i.e. it is continuously differentiable and, for all $\theta_\star \in \nabla F^{-1}(0)$, there exist $\mathcal{V}$ an open neighborhood of $\theta_\star$, $\beta \in (0,1)$ and $\zeta > 0$ such that

$$|F(\theta) - F(\theta_\star)|^\beta \leq \zeta \|\nabla F(\theta)\|, \quad \theta \in \mathcal{V}.$$

Remark 1.8. Unlike [DK24, Definition 1.2], we do not rule out the possibility that $\beta \in \left(0, \frac{1}{2}\right)$. Indeed, via $\mathcal{H}1$ and Lemma A.1(ii), the above property yields $(F(\theta) - \inf F)^\beta \leq \zeta\|\nabla F(\theta)\| \leq C(F(\theta) - \inf F)^{\frac{\alpha}{1+\alpha}}$ for θ close enough to a global optimum, hence $\beta \geq \frac{\alpha}{1+\alpha}$. This lower bound describes $\left(0, \frac{1}{2}\right]$ as α varies through $(0,1]$.

As surveyed in [GG24, Table 1(a)] or [KNS16, Appendix A], the classical assumptions under which nonconvex (SGD) convergence rates are derived often require compactness of the iterate space and/or bounded stochastic gradient noises [LZ24]. Recently, a slew of literature has departed from such frameworks and settled for a Łojasiewicz-type condition. Such a condition (also dubbed gradient domination [Wei+24]) has found success in proving convergence rates and global convergence properties for GD [AMA05] and proximal methods [AB09]. It has been lately adopted in the context of (SGD) [DK24; Wei+24; AL24] for similar purposes.

Łojasiewicz-type conditions help assess the local growth of a function around its stationary points. One such condition has first been considered in [Pol63, Theorem 4] in a global form to obtain an exponential convergence rate for GD. It has then been revisited in a local form

in [AMA05] to prove iterate convergence of GD, which is not guaranteed otherwise (c.f. the intuitive counterexample in [Cur44, §3] and the explicit one in [AMA05, §2]). Łojasiewicz conditions are notably satisfied by analytic functions [AMA05].

Proposition 1.9 ([Łoj65, §18, Proposition 1])**.** *Let $\varphi\colon \mathbb{R}^m \to \mathbb{R}$ be a function that is analytic on a neighborhood of some point $\theta_\star \in \mathbb{R}^m$ satisfying $\varphi(\theta_\star) = 0$. Then, there exists $\beta \in (0,1)$ such that*

$$\|\nabla\varphi(\theta)\| \geq |\varphi(\theta)|^\beta \quad \text{on a neighborhood of } \theta_\star.$$

When $\nabla\varphi(\theta_\star) \neq 0$, the above Łojasiewicz property is straightforward [DK24, Remark 1.3]. The particular case where $\beta = \frac{1}{2}$ is referred to as the Polyak-Łojasiewicz property [GG24, Remark 2.23]. Bear in mind that the Łojasiewicz condition does not require analyticity per se, as $\mathcal{H}5$ encompasses a larger class of functions [AMA05]. Indeed, [Kur98] guarantees it for subanalytic functions of class $\mathcal{C}^1$. We refer for instance to [BM88; BDL07] for developments on subanalytic functions.

Proposition 1.10 ([Kur98, Theorem ŁI])**.** *Let $\varphi\colon \mathcal{V} \to \mathbb{R}$ be a subanalytic which is differentiable in $\mathcal{V} \setminus \varphi^{-1}(0)$, where $\mathcal{V}$ is an open bounded subset of $\mathbb{R}^m$. Then there exist $C > 0$, $\varrho > 0$ and $\beta \in [0,1)$ such that*

$$\|\nabla\varphi(\theta)\| \geq C|\varphi(\theta)|^\beta,$$

for each $\theta \in \mathcal{V}$ such that $|\varphi(\theta)| \in (0, \varrho)$. If in addition $\lim_{\theta\to\vartheta} \varphi(\theta) = 0$ for some $\vartheta \in \mathcal{V}$ (which holds in the classical case, where φ is analytic and $\vartheta \in \mathcal{V}$, $\varphi(\vartheta) = 0$), then the above inequality holds for each $\theta \in \mathcal{V} \subset \varphi^{-1}(0)$ close to ϑ.

Of course, if $\nabla\varphi(\vartheta) = 0$, then $\beta > 0$.

In accordance with the above results, the Łojasiewicz exponent $\beta \in (0,1)$ and the Łojasiewicz constant $\zeta > 0$ in $\mathcal{H}5$ depend on the stationary point $\theta_\star$. When $\mathcal{V} = \mathbb{R}^m$, F is termed globally Łojasiewicz [GG24, Remark 2.23], in which case every stationary point is a global minimizer [KNS16]. The Łojasiewicz parameters $(\beta, \zeta) \in (0,1) \times \mathbb{R}_+^*$ can then be set uniformly across $\nabla F^{-1}(0)$. Otherwise, in the general case where $\mathcal{V} \subset \mathbb{R}^m$, eliciting a convergence rate directly under $\mathcal{H}5$ is rather tedious. To ease such task, we propose below several alternative frameworks to exploit the local Łojasiewicz property more conveniently.

Proposition 1.11. *Assume $\mathcal{H}5$ holds. Then,*

(i) *if $\mathcal{K} \subset \mathbb{R}^m$ is a compact set,*

a. *there exist $\beta \in (0,1)$ and $\zeta > 0$ such that, for all $\theta_\star \in \nabla F^{-1}(0) \cap \mathcal{K}$, there exists an open neighborhood $\mathcal{V}$ of $\theta_\star$ such that*

$$|F(\theta) - F(\theta_\star)|^\beta \leq \zeta\|\nabla F(\theta)\|, \quad \theta \in \mathcal{V}. \tag{1.4}$$

b. *there exists $\beta \in (0,1)$ such that, for all $\theta_\star \in \nabla F^{-1}(0) \cap \mathcal{K}$, there exists an open neighborhood $\mathcal{V}$ of $\theta_\star$ such that*

$$|F(\theta) - F(\theta_\star)|^\beta \leq \|\nabla F(\theta)\|, \quad \theta \in \mathcal{V}. \tag{1.5}$$

(ii) *there exists $\beta \in (0,1]$ such that, for all $\theta_\star \in \nabla F^{-1}(0)$, there exists an open neighborhood $\mathcal{V}$ of $\theta_\star$ such that*

$$|F(\theta) - F(\theta_\star)|^\beta \leq \|\nabla F(\theta)\|, \quad \theta \in \mathcal{V}. \tag{1.6}$$

(iii) *if $\inf F > -\infty$ and F is globally Łojasiewicz, every stationary point of F is a global minimizer, and there exist $\beta \in \left[\frac{\alpha}{1+\alpha}, 1\right)$ and $\zeta > 0$ such that*

$$(F(\theta) - \inf F)^\beta \leq \zeta\|\nabla F(\theta)\|, \quad \theta \in \mathbb{R}^m. \tag{1.7}$$

Remark 1.12. If one assumes the set of stationary points of F bounded, then Proposition 1.11(i) retrieves uniform Łojasiewicz parameters $(\beta, \zeta) \in (0,1) \times \mathbb{R}_+^*$ across all stationary points. We opt out of this extra condition to allow for the possibility of stationary points existing at infinity.

[AB09, Lemma 1] and [Wei+24, Lemma F.1] fix a compact connected set $\mathcal{K} \subset \nabla F^{-1}(0)$ (of local minima in the case of [Wei+24]) and prove the existence of uniform Łojasiewicz parameters $(\beta, \zeta) \in (0,1) \times \mathbb{R}_+^*$ on an open neighborhood of $\mathcal{K}$. The issue with such approach is the additional need to control $\mathbb{P}(\mathcal{A} \subset \mathcal{K})$, which is not a given, unless additional measures are taken to confine the (SGD) trajectory to a small neighborhood of $\mathcal{A} \subset \mathcal{K}$. In Proposition 1.11(i)a, we utilize [DK24, Lemma 4.3] to obtain uniform Łojasiewicz parameters $(\beta, \zeta) \in (0,1) \times \mathbb{R}_+^*$ on an arbitrary compact set $\mathcal{K} \subset \mathbb{R}^m$. The coercivity of F via $\mathcal{H}4$ hints at its levelsets as natural candidate compact sets upon which one can uniformize the Łojasiewicz parameters. Let

$$\mathcal{L}_\ell \coloneqq \{\theta \in \mathbb{R}^m, F(\theta) \leq \ell\}, \quad \ell > \inf F.$$

Remark 1.13. The continuity and coercivity of F, via $\mathcal{H}1$ and $\mathcal{H}4$, entail the compactness of $\mathcal{L}_\ell$, $\ell > \inf F$.

Proposition 1.11(i)b provides a similar framework to Proposition 1.11(i)a, however with a Łojasiewicz constant $\zeta = 1$. The presented advantage is discussed in more detail following Theorem 1.17.

To allow for a global view on the function-value convergence behavior, we develop a novel formulation for the Łojasiewicz condition in Proposition 1.11(ii), justified by the outcome of Lemma A.5. The latter is based on the observation that $\frac{|F(\theta)-F(\theta_\star)|^{\beta_{\theta_\star}}}{\|\nabla F(\theta)\|}$ is locally bounded around each $\theta_\star \in \nabla F^{-1}(0)$ [BDL07]. This new formulation ensures a unique global exponent $\beta \in (0,1]$ and a unique global scaling factor $\zeta = 1$ throughout all stationary points of F.

Eventually, Proposition 1.11(iii) recalls a classical result on globally Łojasiewicz functions.

The following result provides a convergence rate in high probability using $\mathcal{H}3$ as a sole descriptor of the asymptotic behavior of $(\gamma_n)_{n\geq 1}$.

Theorem 1.14. *Assume $\mathcal{H}1$–5 hold. Then,*

(i) *for all $\delta \in (0,1)$, there exists $\ell_\delta > \inf F$ such that $\mathbb{P}(\mathcal{A} \subset \mathcal{L}_{\ell_\delta}) \geq 1-\delta$, and*

- **a.** *there exist $\beta_\delta \in (0,1)$ and $\zeta_\delta > 0$ such that (1.4) holds on $\mathcal{L}_{\ell_\delta}$.*
- **b.** *there exists $\beta_\delta \in (0,1)$ such that (1.5) holds on $\mathcal{L}_{\ell_\delta}$.*

Besides, there exists $C_\delta > 0$ such that, for all $n \geq 1$,

$$\inf_{0\leq k\leq n} F(\theta_k) - F_\star \leq C_\delta \Big(\sum_{k=1}^n \gamma_k\Big)^{-1\wedge \frac{1}{2\beta_\delta}} \quad \text{with probability at least } 1-\delta.$$

(ii) *there exists $\beta \in (0,1]$ such that (1.6) holds. Moreover, for all $\delta \in (0,1)$, there exists $C_\delta > 0$ such that, for all $n \geq 1$,*

$$\inf_{0\leq k\leq n} F(\theta_k) - F_\star \leq C_\delta \Big(\sum_{k=1}^n \gamma_k\Big)^{-1\wedge \frac{1}{2\beta}} \quad \text{with probability at least } 1-\delta.$$

(iii) *if F is globally Łojasiewicz, there exist $\beta \in \big[\frac{\alpha}{1+\alpha}, 1\big)$ and $\zeta > 0$ such that (1.7) holds. Furthermore, there exists $C > 0$ such that, for all $\delta \in (0,1)$ and all $n \geq 1$,*

$$\inf_{0\leq k\leq n} F(\theta_k) - \inf F \leq \frac{C}{\delta^{\frac{1}{\beta}}} \Big(\sum_{k=1}^n \gamma_k\Big)^{-\frac{1}{2\beta}} \quad \text{with probability at least } 1-\delta.$$

To obtain finer convergence rates than to the ones provided above, we describe the asymptotic behavior of $(\gamma_n)_{n\geq 1}$ more precisely.

$\mathcal{H}$6. **(i)** Either of the following holds:

a. $\ln\left(\frac{\gamma_{n-1}}{\gamma_n}\right) = \mathrm{o}(\gamma_n)$;

b. there exists $\gamma_\star > 0$ such that $\ln\left(\frac{\gamma_{n-1}}{\gamma_n}\right) \sim \frac{\gamma_n}{\gamma_\star}$.

(ii) $\sum_{n=1}^{\infty} \frac{\gamma_n}{\sum_{k=1}^n \gamma_k} = \infty$.

(iii) $\gamma_n = \mathrm{O}\big(\big(\sum_{k=1}^n \gamma_k\big)^{-\rho}\big)$, for some $\rho > \frac{1}{\alpha} \geq 1$.

Example 1.15. Suppose that, as in (1.2), $\gamma_n = \Theta(\frac{1}{n^s})$, $n \geq 1$, with $s \in (0,1]$. Then $s < 1$ corresponds to $\mathcal{H}$6(i)a and $s = 1$ to $\mathcal{H}$6(i)b. $\mathcal{H}$6(ii) checks for all $s \in (0,1]$ and $\mathcal{H}$6(iii) applies if $s \in [\frac{\rho}{1+\rho}, 1]$. By the increasing monotony of $x \in \mathbb{R}_+ \mapsto \frac{x}{1+x}$ and the fact that $\rho > \frac{1}{\alpha}$, one has $\frac{\rho}{1+\rho} > \frac{1}{1+\alpha}$, so $\mathcal{H}$3 is automatically verified.

Remark 1.16. $\mathcal{H}$6(i) is standard in the stochastic approximation literature [For15, Assumption C4]. The parameter $\gamma_\star$ in $\mathcal{H}$6(i)b can be interpreted as an initialization parameter, since if $\gamma_n = \frac{\gamma_1}{n}$, $n \geq 1$, one simply has $\gamma_\star = \gamma_1$. $\mathcal{H}$6(ii) appears as early as in [RM51, Equation (26)] for studying stochastic approximation algorithms. $\mathcal{H}$6(iii) generalizes [DK24, Proposition 4.3(c)] by allowing for a flexible choice on ρ. Essential properties satisfied by sequences fulfilling $\mathcal{H}$6 are provided in Lemma A.4.

Theorem 1.17. *Assume $\mathcal{H}$1–6 hold. Then,*

(i) *for all $\delta \in (0,1)$, there exists $\ell_\delta > \inf F$ such that $\mathbb{P}(\mathcal{A} \subset \mathcal{L}_{\ell_\delta}) \geq 1-\delta$, and*

a. *there exist $\beta_\delta \in (0,1)$ and $\zeta_\delta > 0$ such that (1.4) holds on $\mathcal{L}_{\ell_\delta}$. Besides, there exists $C_\delta > 0$ such that, for all $n \geq 1$, with probability at least $1-\delta$,*

$$F(\theta_n) - F_\star \leq C_\delta \begin{cases} \gamma_n^{\alpha \wedge \frac{1}{2}}, & \text{if } \beta_\delta \in (0, \frac{1}{2}] \text{ and } \gamma_\star > \zeta_\delta^{1/\beta_\delta}(\alpha \vee \frac{1}{2}) \text{ under } \mathcal{H}6\text{(i)b}, \\ \big(\sum_{k=1}^n \gamma_k\big)^{-r_\delta}, & \text{if } \beta_\delta \in (\frac{1}{2}, 1), \text{ with } r_\delta = \frac{1}{2\beta_\delta - 1} \wedge (\alpha\rho - 1) \wedge \frac{\rho - 1}{2} > 0. \end{cases}$$

b. *there exists $\beta_\delta \in (0,1)$ such that (1.5) holds on $\mathcal{L}_{\ell_\delta}$. Besides, there exists $C_\delta > 0$ such that, for all $n \geq 1$, with probability at least $1-\delta$,*

$$F(\theta_n) - F_\star \leq C_\delta \begin{cases} \gamma_n^{\alpha \wedge \frac{1}{2}}, & \text{if } \beta_\delta \in (0, \frac{1}{2}] \text{ and } \gamma_\star > \alpha \vee \frac{1}{2} \text{ under } \mathcal{H}6\text{(i)b}, \\ \big(\sum_{k=1}^n \gamma_k\big)^{-r_\delta}, & \text{if } \beta_\delta \in (\frac{1}{2}, 1), \text{ with } r_\delta = \frac{1}{2\beta_\delta - 1} \wedge (\alpha\rho - 1) \wedge \frac{\rho - 1}{2} > 0. \end{cases}$$

(ii) *there exists $\beta \in (0,1]$ such that (1.6) holds. Moreover, for all $\delta \in (0,1)$, there exists $C_\delta > 0$ such that, for all $n \geq 1$, with probability at least $1-\delta$,*

$$F(\theta_n) - F_\star \leq C_\delta \begin{cases} \gamma_n^{\alpha \wedge \frac{1}{2}}, & \text{if } \beta \in (0, \frac{1}{2}] \text{ and } \gamma_\star > \alpha \vee \frac{1}{2} \text{ under } \mathcal{H}6\text{(i)b}, \\ \big(\sum_{k=1}^n \gamma_k\big)^{-r}, & \text{if } \beta \in (\frac{1}{2}, 1], \text{ with } r = \frac{1}{2\beta - 1} \wedge (\alpha\rho - 1) \wedge \frac{\rho - 1}{2} > 0. \end{cases}$$

(iii) *if F is globally Łojasiewicz, there exist $\beta \in \big[\frac{\alpha}{1+\alpha}, 1\big)$ and $\zeta > 0$ such that (1.7) holds. Suppose $\beta \in \big(\frac{\alpha}{1+\alpha}, 1\big) \cup \big\{\frac{1}{2}\big\}$. Then, there exists $C > 0$ such that,*

- *if $\beta \in \big(\frac{\alpha}{1+\alpha}, \frac{1}{2}\big] \cup \big\{\frac{1}{2}\big\}$, assuming*

$$\gamma_\star > \zeta^{2+\lambda\frac{1+\alpha}{\alpha}} \Big(\frac{(1+\alpha)L^{\frac{1}{\alpha}}}{\alpha}\Big)^{\lambda} \alpha, \quad \lambda = \begin{cases} \frac{1-2\beta}{\frac{1+\alpha}{\alpha}\beta - 1}, & \beta < \frac{1}{2}, \\ 0, & \beta = \frac{1}{2}, \end{cases} \quad \text{under } \mathcal{H}6\text{(i)b},$$

for all $\delta \in (0,1)$ and all $n \geq 1$,

$$F(\theta_n) - \inf F \leq \frac{C}{\delta}\gamma_n^{\alpha} \quad \text{with probability at least } 1-\delta.$$

- *if* $\beta \in \big(\frac{1}{2}, 1\big)$*, denoting* $r = \frac{1}{2\beta-1} \wedge (\alpha\rho - 1) > 0$*, for all* $\delta \in (0,1)$ *and all* $n \geq 1$*,*

$$F(\theta_n) - \inf F \leq \frac{C}{\delta}\Big(\sum_{k=1}^{n} \gamma_k\Big)^{-r} \quad \textit{with probability at least } 1 - \delta.$$

Remark 1.18. The convergence rate proof for globally Łojasiewicz loss landscapes with $\beta \leq \frac{1}{2}$ relies on obtaining a positive constant $\mu > 0$ such that $\|\nabla F(\theta)\|^2 \geq \mu(F(\theta) - \inf F)$, $\theta \in \mathbb{R}^m$. This is not possible if $\beta = \frac{\alpha}{1+\alpha}$ and $\alpha \in (0,1)$. Indeed, within the framework of Theorem 1.17(iii), if $\beta = \frac{\alpha}{1+\alpha}$, $\alpha \in (0,1]$, then, by (1.7) and Lemma A.1(ii),

$$\beta L^{-\frac{1}{\alpha}} \|\nabla F(\theta)\|^{\frac{1}{\beta}} \leq F(\theta) - \inf F \leq \zeta^{\frac{1}{\beta}} \|\nabla F(\theta)\|^{\frac{1}{\beta}}, \quad \theta \in \mathbb{R}^m.$$

Assuming $\beta L^{-\frac{1}{\alpha}} \leq \zeta^{\frac{1}{\beta}}$, there exists a function $\phi\colon \mathbb{R}^m \to \big[\beta L^{-\frac{1}{\alpha}}, \zeta^{\frac{1}{\beta}}\big]$ such that $F(\theta) - \inf F = \phi(\theta)\|\nabla F(\theta)\|^{\frac{1}{\beta}}$, $\theta \in \mathbb{R}^m$. But then,

$$\inf_{\theta \in \mathbb{R}^m} \frac{\|\nabla F(\theta)\|^2}{F(\theta) - \inf F} = \frac{1}{\sup_{\theta \in \mathbb{R}^m} \phi(\theta)^{\beta} (F(\theta) - \inf F)^{1-2\beta}},$$

which, given the coercivity of F, is nonnull if and only if $\beta = \frac{\alpha}{1+\alpha} = \frac{1}{2}$, i.e. $\alpha = 1$.

Remark 1.19. For simplicity, denote β the Łojasiewicz exponent within the framework of either of Theorems 1.17(i)a, b, (ii) or (iii). Suppose $\alpha = 1$ and $\gamma_n = \Theta(\frac{1}{n^s})$, $s \in (0,1]$.

(i) Within the frameworks of Theorems 1.17(i)a, b or (ii), if $\beta = \frac{1}{2}$, the (SGD) convergence rate in $\mathrm{O}(n^{-\frac{s}{2}})$ is optimal if $s = 1$, achieving an $\mathrm{O}(n^{-\frac{1}{2}})$. If $\beta > \frac{1}{2}$, the (SGD) convergence rate in $\mathrm{O}\big(n^{-(1-s)(\frac{1}{2\beta-1} \wedge \frac{\rho-1}{2})}\big)$ if $s < 1$, and in $\mathrm{O}\big((\ln n)^{-\frac{1}{2\beta-1} \wedge \frac{\rho-1}{2}}\big)$ if $s = 1$, is optimal for $\rho = \frac{1+2\beta}{2\beta-1}$ and $s = \frac{\rho}{1+\rho} = \frac{1+2\beta}{4\beta}$, scoring an $\mathrm{O}(n^{-\frac{1}{4\beta}})$. Note that this convergence rate coincides with that of $\beta = \frac{1}{2}$ when $\beta \downarrow \frac{1}{2}$, and approaches an $\mathrm{O}(n^{-\frac{1}{4}})$ when $\beta \uparrow 1$.

(ii) As for the globally Łojasiewicz framework of Thorem 1.17(iii), the (SGD) convergence rate in $\mathrm{O}(n^{-s})$ when $\beta = \frac{1}{2}$ is optimal if $s = 1$, yielding an $\mathrm{O}(n^{-1})$. When $\beta > \frac{1}{2}$, it is in $\mathrm{O}\big(n^{-(1-s)(\frac{1}{2\beta-1} \wedge (\rho-1))}\big)$ if $s < 1$ and in $\mathrm{O}\big((\ln n)^{-\frac{1}{2\beta-1} \wedge (\rho-1)}\big)$ if $s = 1$, which is optimal for $\rho = \frac{2\beta}{2\beta-1}$ and $s = \frac{2\beta}{4\beta-1}$, resulting in $\mathrm{O}(n^{-\frac{1}{4\beta-1}})$. This convergence rate is of the same order as the one for $\beta = \frac{1}{2}$ if $\beta \downarrow \frac{1}{2}$, and reaches an $\mathrm{O}(n^{-\frac{1}{3}})$ if $\beta \uparrow 1$.

Let us study the computational complexity of (SGD). Define the reciprocal $\gamma^{-1}\colon \mathbb{R}_+^* \to \mathbb{N}$ by

$$\gamma^{-1}(t) = \inf\{n \geq 1\colon \gamma_n \leq t\}, \quad t > 0.$$

Note that $\gamma_{\gamma^{-1}(t)} \leq t$, $t > 0$. Define also

$$\Sigma_n \coloneqq \sum_{k=1}^{n+1} \gamma_k, \quad n \geq 1.$$

$(\Sigma_n)_{n \geq 1}$ is hence nondecreasing. Let $\Sigma^{-1}\colon \mathbb{R}_+^* \to \mathbb{N}$ be the mapping given by

$$\Sigma^{-1}(t) = \inf\{n \geq 1\colon \Sigma_n \geq t\}, \quad t > 0.$$

Observe that $\Sigma_{\Sigma^{-1}(t)} \geq t$, $t > 0$.

Example 1.20. If $(\gamma_n)_{n \geq 1}$ is given by (1.2), with $\gamma_0 = 1$ and $c = 0$, we have

$$\gamma^{-1}(t) = \Big\lceil \frac{1}{t^{\frac{1}{s}}} \Big\rceil \quad \text{and} \quad \Sigma^{-1}(t) = \begin{cases} \lceil ((1-s)t)^{\frac{1}{1-s}} \rceil, & s \in (0,1), \\ \lceil \mathrm{e}^{t-1} \rceil, & s = 1, \end{cases} \quad t > 0.$$

Corollary 1.21. *Let $\varepsilon > 0$ be a prescribed tolerance. Then,*

(i) a. *within the framework of either Theorem 1.14(i)a or b, by setting*

$$n \propto \Sigma^{-1}\big(\varepsilon^{-1\vee 2\beta_\delta}\big),$$

one has $\inf_{0\le k\le n} F(\theta_k) - F_\star \le C_\delta \varepsilon$ with probability at least $1-\delta$.

b. *within the framework of Theorem 1.14(ii), by setting*

$$n \propto \Sigma^{-1}\big(\varepsilon^{-1\vee 2\beta}\big),$$

one has $\inf_{0\le k\le n} F(\theta_k) - F_\star \le C_\delta \varepsilon$ with probability at least $1-\delta$.

c. *within the framework of Theorem 1.14(iii), by setting*

$$n \propto \Sigma^{-1}(\varepsilon^{-2\beta}),$$

one has $\inf_{0\le k\le n} F(\theta_k) - \inf F \le \frac{C}{\delta^{1/\beta}}\varepsilon$ with probability at least $1-\delta$.

(ii) a. *within the framework of either Theorem 1.17(i)a or b, by setting*

$$n \propto \begin{cases} \gamma^{-1}(\varepsilon^{2\vee\frac{1}{\alpha}}), & \text{if } \beta_\delta \in (0,\tfrac{1}{2}], \\ \Sigma^{-1}(\varepsilon^{-\frac{1}{r_\delta}}), & \text{if } \beta_\delta \in (\tfrac{1}{2},1), \end{cases}$$

one has $F(\theta_n) - F_\star \le C_\delta \varepsilon$ with probability at least $1-\delta$.

b. *within the framework of Theorem 1.17(ii), by setting*

$$n \propto \begin{cases} \gamma^{-1}(\varepsilon^{2\vee\frac{1}{\alpha}}), & \text{if } \beta \in (0,\tfrac{1}{2}], \\ \Sigma^{-1}(\varepsilon^{-\frac{1}{r}}), & \text{if } \beta \in (\tfrac{1}{2},1], \end{cases}$$

one has $F(\theta_n) - F_\star \le C_\delta \varepsilon$ with probability at least $1-\delta$.

c. *within the framework of Theorem 1.17(iii), by setting*

$$n \propto \begin{cases} \gamma^{-1}(\varepsilon^{\frac{1}{\alpha}}), & \text{if } \beta \in (\frac{\alpha}{1+\alpha},\tfrac{1}{2}] \cup \{\tfrac{1}{2}\}, \\ \Sigma^{-1}(\varepsilon^{-\frac{1}{r}}), & \text{if } \beta \in (\tfrac{1}{2},1), \end{cases}$$

one has $F(\theta_n) - F_\star \le C_\delta \varepsilon$ with probability at least $1-\delta$.

All in all, the (SGD) *computational complexity has bound* $\mathrm{Cost} \le Cn$, *for some constant* $C > 0$.

The convergence rates derived in Theorems 1.17(i,ii) impose no condition on the initialization θ_0 and no concentration on the stochastic gradients $(g_n(\theta_{n-1}))_{n\ge 0}$ (we refer to [AL24; Sa+22; JO19; MDB24] for recent developments on (SGD) gradient concentration). Our obtained results concern the function-value convergence speeds toward the random critical level $F_\star$ as described in Theorem 1.5, and not to a predefined deterministic local minimum as in [AL24, Theorem 3.1] or [Wei+24, Theorem 5.1]. This means in particular that the function value gaps $(F(\theta_n)-F_\star)_{n\ge 0}$ that we deal with are permitted to take negative values.

Except for the condition on $\gamma_\star$ in the case of $\beta_\delta \in \big(0,\frac{1}{2}\big]$ and $\mathcal{H}6$(i)b in Theorem 1.17(i)a, to derive a convergence rate at the confidence level $1-\delta$, $\delta \in (0,1)$, our settings are δ-free. Note that the extra constraint on $\gamma_\star$ in Theorem 1.17 when the Łojasiewicz exponent lies in $\big(0,\frac{1}{2}\big]$ and $\mathcal{H}6$(i)b holds is classical in stochastic approximation literature [Duf96]. The advantage of Theorem 1.17(i)b over (i)a is that the initialization condition $\gamma_\star > \frac{1}{2}\vee\alpha$ for the case of $\beta_\delta \in \big(0,\frac{1}{2}\big]$ and $\mathcal{H}6$(i)b becomes free of the knowledge of $\zeta_\delta^{1/\beta_\delta}$, bearing in mind that the exponent β_δ and

the factor C_δ in the convergence rate bound differ. Such a framework will be useful in Section 4 to help extend global convergence to the case $\mathcal{H}6$(i)b.

To obtain high probability convergence rates at the confidence level $1-\delta$, $\delta \in (0,1)$, [AL24, Theorem 3.1] sets the initialization θ_0 in a small δ-dependent neighborhood of a local minimum, and [Wei+24, Theorem 5.1] requires an initialization that is sufficiently close to a local minimum and bounds the learning rates by a small value dependent on δ. These constraints are designed to prevent (SGD) trajectories from escaping a small region of space where a Łojasiewicz property holds. They may however be restrictive as they require a priori knowledge on $\nabla F^{-1}(0)$, which is unavailable, and their settings are δ-dependent. In practice, an expensive computational overhead is expected to meet all the required conditions.

For the sake of simplicity, let β denote the Łojasiewicz exponent irrespectively of the frameworks of Theorems 1.17(i)a, b and (ii). The shift in behavior, demarcated at $\beta = \frac{1}{2}$, is already observed for GD and proximal methods [FGP15; AB09]. When $\beta = \frac{1}{2}$, F has a local quadratic growth [Liu+23, Lemma 2.1], hence promoting faster convergence rates. When $\beta > \frac{1}{2}$, F is locally subquadratic, so (SGD) is slower to converge. The globally Łojasiewicz case in Theorem 1.17(iii) displays faster convergence rates than Theorems 1.17(i,ii), as it possesses a unique critical level – its global minimum – that acts as an attractor to (SGD) trajectories.

[Lei+20, Theorem 4.6(b)] proves a convergence rate on the mean function-value gap $\mathbb{E}[F(\theta_n) - \inf F]$ in $\mathrm{O}(n^{-\alpha})$ for learning rates $\gamma_n = \Theta\big(\frac{1}{n}\big)$, assuming (L,α)-Hölder differentiability of $\theta \in \mathbb{R}^m \mapsto f(x,\theta)$, $x \in \mathbb{R}^d$, and a global PŁ condition on F ($\beta = \frac{1}{2}$). This result comparable to Theorem 1.17(iii).

[DK24, Proposition 4.3] deals with the case where the Łojasiewicz exponent β is in $\big(\frac{1}{2},1\big)$ and presents a convergence rate in mean that is conditional to the existence of deterministic quantifications of stochastic gradients' conditional moments [DK24, Assumption $\mathbb{M}^{\sigma,q}$], and the pertaining of the iterates to a predefined compact set. The proposed convergence rate is in $\mathrm{O}\big(\big(\sum_{k=1}^n \gamma_k\big)^{-\frac{1}{2\beta-1}}\big)$. If $\gamma_n = \Theta\big(\frac{1}{n^s}\big)$, $s \in (0,1]$, [DK24, Remark 4.1] recommends choosing s close enough to 1, lower bounded by parameters governing [DK24, Assumption $\mathbb{M}^{\sigma,q}$].

Assuming the stochastic gradients to be Gaussianly concentrated, [AL24, Theorem 3.1] derives an exponential convergence rate in high probability for locally Lipschitz-differentiable loss landscapes, conditionally to all the iterates pertaining to a closed ball where a local PŁ condition ($\beta = \frac{1}{2}$) holds and to θ_0 being initialized in a small neighborhood of the local minimum.

[Wei+24, Theorem 5.1] presents a high probability convergence rate for locally Lipschitz-differentiable loss landscapes under a local Łojasiewicz condition with $\beta \in \big[\frac{1}{2},1\big]$. This result is conditional to an initialization that is sufficiently close to a local minimum and is obtained for learning rates such that $\gamma_n = \Theta\big(\frac{1}{n^s}\big)$, $s \in (0,1)$, bounded by a small value. The provided rate is unavailable for $s = 1$. Upon closer inspection, the provided result is likely erroneous: in [Wei+24, Lemma F.4], the inequality "$D_{n+1}\mathbb{1}_{\Omega_{n+1}} \leq D_{n+1}\mathbb{1}_{\Omega_n}$" is wrong, as it requires $D_{n+1}\mathbb{1}_{\Omega_n} \geq 0$, which is not guaranteed therein. Using our notation, for a small neighborhood $\mathcal{V}$ around a local minimum $\theta_\star \in \nabla F^{-1}(0)$ such that $F(\theta) \geq F(\theta_\star)$, $\theta \in \mathcal{V}$, the aforementioned inequality reads $(F(\theta_{n+1}) - F(\theta_\star)) \prod_{k=1}^{n+1} \mathbb{1}_{\theta_k \in \mathcal{V}} \leq (F(\theta_{n+1}) - F(\theta_\star)) \prod_{k=1}^{n} \mathbb{1}_{\theta_k \in \mathcal{V}}$, which is false since the nonnegativity of $(F(\theta_{n+1}) - F(\theta_\star)) \prod_{k=1}^{n} \mathbb{1}_{\theta_k \in \mathcal{V}}$ is not guaranteed. [Wei+24, Theorem 4.1] retrieves a convergence rate for Lipschitz-differentiable globally Łojasiewicz loss landscapes with $\beta \in \big[\frac{1}{2},1\big]$ that attains optimally an $\mathrm{o}(n^{-\eta})$, with $\eta < \frac{1}{4\beta-1}$. The optimal convergence rate we recover in Remark 1.19(ii) is sharper and is exactly in $\mathrm{O}(n^{-\frac{1}{4\beta-1}})$. Besides, as in the locally Łojasiewicz case, the provided rate in [Wei+24] is unavailable for $s = 1$.

1.3 Global Convergence Analysis

The goal of this section is to prove the $\mathbb{P}$-as convergence of the iterates $(\theta_n)_{n\geq 0}$ to a single-point limit. Such a property is important to implement a Polyak-Ruppert averaging [DK24], known to increase the speed and numerical stability of (SGD). The proofs of this section are gathered in Section 4.

To simplify matters, we assume

***H*7.** $\gamma_n = \Theta\big(\frac{1}{n^s}\big)$, $s \in (0,1]$.

Theorem 1.22. *Assume $\mathcal{H}1$–7 hold and that $\rho > 3 \vee \frac{2}{\alpha}$, $s \in \big[\frac{\rho}{1+\rho}, 1\big]$ and $\gamma_\star > \frac{1}{2} \vee \alpha$ if $s = 1$. Then $(\theta_n)_{n\geq 0}$ converges $\mathbb{P}$-as to an $\mathbb{R}^m$-valued random variable $\theta_\star$ that is $\mathbb{P}$-as a stationary point of F and satisfies $F(\theta_\star) = F_\star$ $\mathbb{P}$-as.*

Remark 1.23. It ensues from $\rho > 3 \vee \frac{2}{\alpha}$ and $s \in \big[\frac{\rho}{1+\rho}, 1\big]$ that $s \in \big(\frac{3}{4} \vee \frac{2}{2+\alpha}, 1\big]$.

Recalling Lemma 2.1(iv), a quick glance at the decomposition (4.16) reveals that the $\mathbb{P}$-as-summability of $(\gamma_n \|\nabla F(\theta_{n-1})\|)_{n\geq 1}$ translates directly into the global convergence of (SGD). Proving this summability has thus been the focus of most recent literature in this direction [DK24; CFR23; QMM24]. Theorem 1.22 shows that, by choosing ρ large enough and $\gamma_\star$ as well if $s = 1$, then global convergence is ascertained for all $s \in \big[\frac{\rho}{1+\rho}, 1\big]$.

[DK24, Theorem 1.2] proves global convergence for locally Lipschitz-differentiable loss landscapes, with Łojasiewicz exponent $\beta \in \big[\frac{1}{2}, 1\big)$, however under the additional constraint of existence of deterministic quantifications of stochastic gradients' conditional moments [DK24, Assumption $\mathbb{M}^{\sigma,q}$].

[CFR23, Proposition 6.3] considers a Lipschitz-differentiable objective function verifying a KŁ property on a compact set, and shows iterate convergence for (SGD) trajectories converging in this set. The formalism developed therein relies on extending the KŁ property to a Lyapunov function V that is built so as $(V(\theta_n))_{n\geq 0}$ is a converging supermartingale. Note that our approach uses a deterministic vanishing nonincreasing Lyapunov sequence $(G_n)_{n\geq 0}$ (4.6) to help prove the $L^1(\mathbb{P})$-summability of $(\gamma_n \nabla F(\theta_{n-1}))_{n\geq 1}$. Nonetheless, the iterate convergence result obtained in [CFR23] relies on assumptions that may be difficult to adhere to. Firstly, it is supposed that $(V(\theta_n))_{n\geq 0}$ asymptotically never registers a dropdown below its target $V_\star$, i.e. $\mathbb{P}(\liminf_{n\to\infty}\{V(\theta_n) > V_\star\}) = 1$. Secondly, it is assumed that the information revelation on $F_\star$ is lower-bounded by $F_\star$ itself, i.e. $\mathbb{E}[F_\star|\mathcal{F}_n] \geq F_\star$ $\mathbb{P}$-as. Thirdly, it is presumed that $\|\nabla F(\theta_n)\| > 0$, $n \geq 1$, $\mathbb{P}$-as. Finally, it is required that a strong growth condition holds, i.e. $\mathbb{E}_{n-1}[\|g_n(\theta_{n-1})\|^2] \leq C\|\nabla F(\theta_{n-1})\|^2$, $n \geq 1$, with $C > 0$.

[QMM24, Theorem 5.1 & Corollary 5.2] look into an (SGD) variant with momentum, a technique known to help with iterate convergence. These results suppose however the uniform boundedness of the conditional stochastic gradient variances [QMM24, Assumption 2.1].

Corollary 1.24. *Within the framework of Theorem 1.22,*

(i) *for all $\delta \in (0,1)$, there exists $\ell_\delta > \inf F$ such that $\mathbb{P}(\mathcal{A} \subset \mathcal{L}_{\ell_\delta}) \geq 1 - \delta$, and*

- **a.** *there exist $\beta_\delta \in (0,1)$ and $\zeta_\delta > 0$ such that (1.4) holds on $\mathcal{L}_{\ell_\delta}$.*
- **b.** *there exists $\beta_\delta \in (0,1)$ such that (1.5) holds on $\mathcal{L}_{\ell_\delta}$.*

Besides, there exists $C_\delta > 0$ such that,

- *if $\beta_\delta \in \big(0, \frac{1}{2}\big]$, by setting $\sigma \in \big[\frac{1}{2(2s-1)} \vee \frac{\alpha s}{2((1+\alpha)s-1)} \vee \frac{1-((1-\alpha)\vee\frac{1}{2})s}{2(\alpha\wedge\frac{1}{2})s}, 1\big)$ and, if $s = 1$, $\gamma_\star > (1 \vee \zeta_\delta^{1/\beta_\delta})(\alpha \vee \frac{1}{2})$ if (i)a holds and $\gamma_\star > \alpha \vee \frac{1}{2}$ if (i)b holds, then, for all $n \geq 1$,*

$$\|\theta_n - \theta_\star\| \leq \frac{C_\delta}{1-\sigma} n^{-((\alpha\wedge\frac{1}{2})s \wedge \frac{2s-1}{2} \wedge ((1+\alpha)s-1))(1-\sigma)} \quad \text{with probability at least } 1-\delta.$$

- *if* $\beta_\delta \in \big(\frac{1}{2},1\big)$ *and* $s \in \big[\frac{\rho}{1+\rho},1\big)$*, by setting* $\sigma \in \big[\frac{1}{2(2s-1)} \vee \frac{\alpha s}{2((1+\alpha)s-1)} \vee \frac{1+r_\delta}{2r_\delta},1\big)$*, then, for all* $n \geq 1$*,*

$$\|\theta_n - \theta_\star\| \leq \frac{C_\delta}{1-\sigma} n^{-(r_\delta(1-s)\wedge\frac{2s-1}{2}\wedge((1+\alpha)s-1))(1-\sigma)} \quad \text{with probability at least } 1-\delta,$$

recalling that $r_\delta = \frac{1}{2\beta_\delta - 1} \wedge (\alpha\rho - 1) \wedge \frac{\rho-1}{2} > 1$*.*

- *if* $\beta_\delta \in \big(\frac{1}{2},1\big)$ *and* $s = 1$*, by setting* $\sigma \in \big[\frac{1+r_\delta}{2r_\delta},1\big)$*, then, for all* $n \geq 1$*,*

$$\|\theta_n - \theta_\star\| \leq \frac{C_\delta}{1-\sigma} (\ln n)^{-r_\delta(1-\sigma)} \quad \text{with probability at least } 1-\delta,$$

recalling that $r_\delta = \frac{1}{2\beta_\delta - 1} \wedge (\alpha\rho - 1) \wedge \frac{\rho-1}{2} > 1$*.*

(ii) *if* F *is globally Łojasiewicz, there exist* $\beta \in \big[\frac{\alpha}{1+\alpha},1\big)$ *and* $\zeta > 0$ *such that* (1.7) *holds. Suppose* $\beta \in \big(\frac{\alpha}{1+\alpha},1\big) \cup \big\{\frac{1}{2}\big\}$*. Then, there exists* $C > 0$ *such that,*

- *if* $\beta \in \big(\frac{\alpha}{1+\alpha},\frac{1}{2}\big] \cup \big\{\frac{1}{2}\big\}$*, by setting* $\sigma \in \big[\frac{1}{2(2s-1)} \vee \frac{\alpha s}{2((1+\alpha)s-1)},1\big)$ *and, if* $s = 1$*,*

$$\gamma_\star > \zeta^{2+\lambda\frac{1+\alpha}{\alpha}} \Big(\frac{(1+\alpha)L^{\frac{1}{\alpha}}}{\alpha}\Big)^{\lambda} \alpha, \quad \lambda = \begin{cases} \frac{1-2\beta}{\frac{1+\alpha}{\alpha}\beta - 1}, & \beta < \frac{1}{2}, \\ 0, & \beta = \frac{1}{2}, \end{cases}$$

for all $n \geq 1$*,*

$$\|\theta_n - \theta_\star\| \leq \frac{C}{\delta(1-\sigma)} n^{-(\alpha s\wedge\frac{2s-1}{2}\wedge((1+\alpha)s-1))(1-\sigma)} \quad \text{with probability at least } 1-\delta.$$

- *if* $\beta \in \big(\frac{1}{2},1\big)$ *and* $s \in \big[\frac{\rho}{1+\rho},1\big)$*, by setting* $\sigma \in \big[\frac{1}{2(2s-1)} \vee \frac{\alpha s}{2((1+\alpha)s-1)} \vee \frac{1+r}{2r},1\big)$*, for all* $n \geq 1$*,*

$$\|\theta_n - \theta_\star\| \leq \frac{C}{\delta(1-\sigma)} n^{-(r(1-s)\wedge\frac{2s-1}{2}\wedge((1+\alpha)s-1))(1-\sigma)} \quad \text{with probability at least } 1-\delta,$$

recalling that $r = \frac{1}{2\beta - 1} \wedge (\alpha\rho - 1) > 1$*.*

- *if* $\beta \in \big(\frac{1}{2},1\big)$ *and* $s = 1$*, by setting* $\sigma \in \big[\frac{1+r}{2r},1\big)$*, for all* $n \geq 1$*,*

$$\|\theta_n - \theta_\star\| \leq \frac{C}{\delta(1-\sigma)} (\ln n)^{-r(1-\sigma)} \quad \text{with probability at least } 1-\delta,$$

recalling that $r = \frac{1}{2\beta - 1} \wedge (\alpha\rho - 1) > 1$*.*

Overall, the optimal convergence rates are obtained by taking the smallest possible parameter σ*.*

Remark 1.25. Let $(\beta,\zeta) \in (0,1) \times \mathbb{R}_+^*$ denote the Łojasiewicz parameters regardless of the framework of Corollary 1.24(i)a, b or (ii). Note that, by $\mathcal{H}1$ and Proposition 1.11, one has asymptotically

$$\|\nabla F(\theta_n)\| = \|\nabla F(\theta_n) - \nabla F(\theta_\star)\| \leq L\|\theta_n - \theta_\star\|^\alpha \quad \mathbb{P}\text{-as},$$
$$F(\theta_n) - F_\star \leq \zeta^{\frac{1}{\beta}} \|\nabla F(\theta_n)\|^{\frac{1}{\beta}} \leq L^{\frac{1}{\beta}} \zeta^{\frac{1}{\beta}} \|\theta_n - \theta_\star\|^{\frac{\alpha}{\beta}} \quad \mathbb{P}\text{-as}.$$

This offers sharper gradient convergence rates than Lemma 3.2 and alternative, more pessimistic, function-value convergence rates to Theorem 1.17.

Corollary 1.26. *Let* $\varepsilon > 0$ *be a prescribed tolerance. Then,*

(i) *within the framework of either Corollary 1.24(i)a or b, by setting*

$$n \propto \begin{cases} \lceil \varepsilon^{\frac{-1}{1-\sigma}(\frac{2\vee\alpha^{-1}}{s}\vee\frac{2}{2s-1}\vee\frac{1}{(1+\alpha)s-1})} \rceil, & \text{if } \beta_\delta \in (0,\frac{1}{2}], \\ \lceil \varepsilon^{\frac{-1}{1-\sigma}(\frac{1}{r_\delta(1-s)}\vee\frac{2}{2s-1}\vee\frac{1}{(1+\alpha)s-1})} \rceil, & \text{if } \beta_\delta \in (\frac{1}{2},1) \text{ and } s<1, \\ \lceil \exp(\varepsilon^{-\frac{1}{r_\delta(1-\sigma)}}) \rceil & \text{if } \beta_\delta \in (\frac{1}{2},1) \text{ and } s=1, \end{cases}$$

one has $\|\theta_n - \theta_\star\| \leq \frac{C_\delta}{1-\sigma}\varepsilon$ *with probability at least* $1-\delta$.

(ii) *within the framework of Corollary 1.24(ii), by setting*

$$n \propto \begin{cases} \lceil \varepsilon^{\frac{-1}{1-\sigma}(\frac{1}{\alpha s}\vee\frac{2}{2s-1}\vee\frac{1}{(1+\alpha)s-1})} \rceil, & \text{if } \beta \in (0,\frac{1}{2}], \\ \lceil \varepsilon^{\frac{-1}{1-\sigma}(\frac{1}{r(1-s)}\vee\frac{2}{2s-1}\vee\frac{1}{(1+\alpha)s-1})} \rceil, & \text{if } \beta \in (\frac{1}{2},1) \text{ and } s<1, \\ \lceil \exp(\varepsilon^{-\frac{1}{r(1-\sigma)}}) \rceil & \text{if } \beta \in (\frac{1}{2},1) \text{ and } s=1, \end{cases}$$

one has $\|\theta_n - \theta_\star\| \leq \frac{C}{\delta(1-\sigma)}\varepsilon$ *with probability at least* $1-\delta$.

Altogether, the subsequent (SGD) *computational cost has bound:* $\mathrm{Cost} \leq Cn$, *for some constant* $C>0$. *Moreover, the optimal computational cost is recovered by selecting the smallest possible parameter* σ.

[Liu+23] takes a closer look at overparametrized neural networks satisfying a local quadratic growth (reminiscent of a PŁ condition, i.e. $\beta = \frac{1}{2}$), as well as an interpolation and an aiming properties around their global minima. The aiming property predicates that the loss landscape's gradients locally follow the direction of the global optimum. The interpolation assumption stipulates the existence of a deterministic $\theta_\star \in \mathbb{R}^m$ minimizing simultaneously F and $\theta \in \mathbb{R}^m \mapsto f(x,\theta)$ for any $x \in \mathbb{R}^d$. [Liu+23] then derives exponential convergence rates in mean and in high probability for (SGD) iterates conditionally to an initialization that is sufficiently close to the global minimum. The proposed result quantifies rather the distance of the iterates to a subset of the global minimizers. It imposes an initialization in a δ-dependent radius around the global optimum to achieve a convergence rate with confidence $1-\delta$. The smaller the δ, the closer the initialization must be to the global optimum.

Note that the interpolation condition, claimed to hold for overparametrized neural networks, fails notwithstanding to encompass some elementary (SGD) problems. Take the classical problem of approximating $\mathbb{E}[Y|X]$ by a neural network $\mathcal{N}_\theta(X)$ parametrized by θ. The optimal $\theta_\star$ can be retrieved as solution to $\min_\theta \mathbb{E}[(Y-\mathcal{N}_\theta(X))^2]$. But if Y is not X-measurable, no matter how overparametrized $\mathcal{N}_\theta$ could get, canceling $Y - \mathcal{N}_\theta(X)$ almost surely with some deterministic $\theta_\star$, hence perfectly interpolating Y with $\mathcal{N}_{\theta_\star}(X)$, implies that Y is X-measurable, which is absurd.

[GP23, Theorem 5(iii)] exhibits $L^{2p}(\mathbb{P})$-convergence rates for (SGD) iterates, $p \geq 1$, for loss landscapes that are of class C^2, Lipschitz-differentiable, globally Łojasiewicz with Łojasiewicz exponent $\beta \in [0,\frac{1}{2}]$ and possess a unique minimizer (by Proposition 1.11(iii), necessarily, $\beta = \frac{1}{2}$). The proposed result requires however that the stochastic gradient noises have uniformly bounded conditional polynomial-exponential moments [GP23, Assumption $\mathrm{H}^{\phi}_{\Sigma_p}$].

2 Convergence Analysis Proofs

Henceforth, $C>0$ designates a constant that may change from line to line, but is independent of all variables of the problem at hand. Given a variable $\delta \in \mathbb{R}$, $C_\delta > 0$ denotes one such constant that may only depend on δ.

Lemma 2.1. *Under* $\mathcal{H}1$*–3,*

(i) $(F(\theta_n))_{n\geq 0}$ *converges* $\mathbb{P}$*-as to a real-valued random variable* $F_\star$, $\sum_{n=1}^\infty \gamma_n \|\nabla F(\theta_{n-1})\|^2 < \infty$ $\mathbb{P}$*-as and* $\liminf_{n\to\infty} \|\nabla F(\theta_n)\| = 0$ $\mathbb{P}$*-as;*

(ii) $(\mathbb{E}[F(\theta_n)])_{n\geq 0}$ *converges,* $\sum_{n=1}^{\infty}\gamma_n\mathbb{E}[\|\nabla F(\theta_{n-1})\|^2]<\infty$ *and* $\liminf_{n\to\infty}\mathbb{E}[\|\nabla F(\theta_n)\|^2]=0$*;*

(iii) $\sum_{n=1}^{\infty}\|\theta_n-\theta_{n-1}\|^{1+\alpha}$ *is in* $L^1(\mathbb{P})$ *and is* $\mathbb{P}$*-as-finite;*

(iv) *the martingale* $\big(\sum_{k=1}^{n}\gamma_k(g_k(\theta_{k-1})-\nabla F(\theta_{k-1}))\big)_{n\geq 1}$ *converges in* $L^2(\mathbb{P})$ *and* $\mathbb{P}$*-as;*

(v) $F_\star\in L^1(\mathbb{P})$.

Remark 2.2. **(i)** By telescopic summation, the triangle inequality and Hölder's inequality with the adjoint exponents $\big(1+\alpha,\frac{1+\alpha}{\alpha}\big)$,

$$\|\theta_n\|\leq\|\theta_0\|+\Big(\sum_{k=1}^{n}\|\theta_k-\theta_{k-1}\|^{1+\alpha}\Big)^{\frac{1}{1+\alpha}}n^{\frac{\alpha}{1+\alpha}}\leq\|\theta_0\|+Yn^{\frac{\alpha}{1+\alpha}},\quad n\geq 1,$$

where $Y\coloneqq\big(\sum_{k=1}^{\infty}\|\theta_k-\theta_{k-1}\|^{1+\alpha}\big)^{\frac{1}{1+\alpha}}\in L^1(\mathbb{P})$ by Jensen's inequality and Lemma 2.1(iii). This upper bound does not exclude that $(\theta_n)_{n\geq 0}$ may escape to ∞, at most at an $\mathrm{O}_{L^1(\mathbb{P})}(n^{\frac{\alpha}{1+\alpha}})$ speed. This can occur for loss landscapes where the minimum is only attained at infinity. Consequently, recalling that $\theta_0\in L^2(\mathbb{P})$,

$$\mathbb{E}[\|\theta_n\|]\leq\mathbb{E}[\|\theta_0\|]+\mathbb{E}[Y]n^{\frac{\alpha}{1+\alpha}},\quad n\geq 1.$$

(ii) By the update rule (SGD) and $\mathcal{H}2$,

$$\mathbb{E}[\|\theta_n-\theta_{n-1}\|^2]=\gamma_n^2\mathbb{E}\big[\|g_n(\theta_{n-1})-\nabla F(\theta_{n-1})\|^2\big]+\gamma_n^2\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big],\quad n\geq 1.$$

Using that, for some $C>0$,

$$\sup_{n\geq 1}\mathbb{E}[\|g_n(\theta_{n-1})-\nabla F(\theta_{n-1})\|^2]\leq C(1+\sup_{n\geq 1}\mathbb{E}[F(\theta_{n-1})-\inf F])<\infty$$

(by $\mathcal{H}1$, $\mathcal{H}2$ and Lemmas A.1(ii) and 2.1(ii)), and that $\sum_{n=1}^{\infty}\gamma_n^2\mathbb{E}[\|\nabla F(\theta_{n-1})\|^2]<\infty$ (by Lemma 2.1(ii) and $\mathcal{H}3$), we obtain a weaker alternative to Lemma 2.1(iii) predicating that $\sum_{n=1}^{\infty}\|\theta_n-\theta_{n-1}\|^2$ is in $L^1(\mathbb{P})$ and $\mathbb{P}$-as-finite (by $\mathcal{H}3$ and Fubini's theorem).

Proof of Lemma 2.1. **(i)** Let $n\geq 1$. By $\mathcal{H}1$, Lemma A.1(i) and the (SGD) recursion,

$$F(\theta_n)=F\big(\theta_{n-1}-\gamma_n g_n(\theta_{n-1})\big)\leq F(\theta_{n-1})-\gamma_n\nabla F(\theta_{n-1})^\top g_n(\theta_{n-1})+\frac{L}{1+\alpha}\gamma_n^{1+\alpha}\|g_n(\theta_{n-1})\|^{1+\alpha}.\tag{2.1}$$

Since $\inf F>-\infty$, by $\mathcal{H}2$,

$$\mathbb{E}_{n-1}[F(\theta_n)-\inf F]\leq F(\theta_{n-1})-\inf F-\gamma_n\|\nabla F(\theta_{n-1})\|^2+\frac{L}{1+\alpha}\gamma_n^{1+\alpha}\mathbb{E}_{n-1}\big[\|g_n(\theta_{n-1})\|^{1+\alpha}\big].\tag{2.2}$$

Using that $x^{1+\alpha}\leq C(1+x^2)$, $x\geq 0$, and $\mathcal{H}2$,

$$\mathbb{E}_{n-1}\big[\|g_n(\theta_{n-1})\|^{1+\alpha}\big]\leq C\big(1+\mathbb{E}_{n-1}\big[\|g_n(\theta_{n-1})\|^2\big]\big)\leq C\big((F(\theta_{n-1})-\inf F)+1\big).\tag{2.3}$$

Thus

$$\mathbb{E}_{n-1}[F(\theta_n)-\inf F]\leq\big(1+C\gamma_n^{1+\alpha}\big)\big(F(\theta_{n-1})-\inf F\big)-\gamma_n\|\nabla F(\theta_{n-1})\|^2+C\gamma_n^{1+\alpha}.\tag{2.4}$$

Recalling $\mathcal{H}3$, the Siegmund-Robbins lemma [RS71] guarantees that $\sum_{n=1}^{\infty}\gamma_n\|\nabla F(\theta_{n-1})\|^2<\infty$ $\mathbb{P}$-as and that $(F(\theta_n))_{n\geq 0}$ converges $\mathbb{P}$-as to a finite real-valued random variable $F_\star$.

Furthermore, one has $\big(\sum_{n=1}^{\infty}\gamma_n\|\nabla F(\theta_{n-1})\|^2\big)\mathbb{1}_{\liminf_{n\to\infty}\|\nabla F(\theta_n)\|>0}<\infty$ $\mathbb{P}$-as. Given that $\sum_{n=1}^{\infty}\gamma_n=\infty$ under $\mathcal{H}3$, it ensues that $\mathbb{P}(\liminf_{n\to\infty}\|\nabla F(\theta_n)\|>0)=0$.

(ii) Taking the mean in (2.4),

$$\mathbb{E}[F(\theta_n) - \inf F] \leq \big(1 + C\gamma_n^{1+\alpha}\big)\mathbb{E}[F(\theta_{n-1}) - \inf F] - \gamma_n \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big] + C\gamma_n^{1+\alpha}. \tag{2.5}$$

By [BC17, Lemma 5.31], $(\mathbb{E}[F(\theta_n)])_{n\geq 0}$ is convergent and $\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[\|\nabla F(\theta_{n-1})\|^2] < \infty$. In view of the fact that $\sum_{n=1}^{\infty} \gamma_n = \infty$, necessarily $\liminf_{n\to\infty} \mathbb{E}\big[\|\nabla F(\theta_n)\|^2\big] = 0$.

(iii) Using the (SGD) update rule and taking the mean in the inequality (2.3) yield

$$\mathbb{E}[\|\theta_n - \theta_{n-1}\|^{1+\alpha}] = \gamma_n^{1+\alpha}\mathbb{E}\big[\|g_n(\theta_{n-1})\|^{1+\alpha}\big] \leq C\big(\mathbb{E}[F(\theta_{n-1}) - \inf F] + 1\big)\gamma_n^{1+\alpha}.$$

Given that $(\mathbb{E}[F(\theta_n)])_{n\geq 1}$ is convergent, it is bounded and

$$\mathbb{E}[\|\theta_n - \theta_{n-1}\|^{1+\alpha}] \leq C\gamma_n^{1+\alpha}.$$

Thus, by Fubini's theorem, $\mathbb{E}\big[\sum_{n=1}^{\infty} \|\theta_n - \theta_{n-1}\|^{1+\alpha}\big] < \infty$, and $\mathbb{P}\big(\sum_{n=1}^{\infty} \|\theta_n - \theta_{n-1}\|^{1+\alpha} = \infty\big) = 0$, yielding the sought result.

(iv) By Jensen's inequality, the tower rule, $\mathcal{H}2$ and Lemma 2.1(ii),

$$\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\big] \leq \mathbb{E}\big[\|g_k(\theta_{k-1})\|^2\big] \leq C\big(\sup_{n\geq 0} \mathbb{E}[F(\theta_n) - \inf F] + 1\big) =: K^2 < \infty, \quad k \geq 1. \tag{2.6}$$

Define the $L^2(\mathbb{P})$ $(\mathcal{F}_n)_{n\geq 0}$-martingale

$$M_n \coloneqq \sum_{k=1}^{n} \gamma_k\big(g_k(\theta_{k-1}) - \nabla F(\theta_{k-1})\big), \quad n \geq 1. \tag{2.7}$$

Using that $\big(g_k(\theta_{k-1}) - \nabla F(\theta_{k-1})\big)_{k\geq 1}$ are martingale increments according to $\mathcal{H}2$, (2.6) and that $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$ under $\mathcal{H}3$, we get

$$\sup_{n\geq 1} \mathbb{E}[\|M_n\|^2] = \sum_{k=1}^{\infty} \gamma_k^2 \mathbb{E}\big[\|g_n(\theta_{n-1}) - \nabla F(\theta_{n-1})\|^2\big] \leq 2K^2 \sum_{k=1}^{\infty} \gamma_k^2 < \infty.$$

$(M_n)_{n\geq 1}$ is thus bounded in $L^2(\mathbb{P})$, and converges in $L^2(\mathbb{P})$ and $\mathbb{P}$-as.

(v) Using Lemma 2.1(i), Fatou's lemma, the triangle inequality and Lemma 2.1(ii),

$$\mathbb{E}[|F_\star|] = \mathbb{E}\big[\lim_{n\to\infty} |F(\theta_n)|\big] \leq \liminf_{n\to\infty} \mathbb{E}[F(\theta_n) - \inf F] + |\inf F| < \infty.$$

□

Lemma 2.3. *Suppose $\mathcal{H}1$–3 hold. Then $\nabla F(\theta_n) \to 0$ both in $L^2(\mathbb{P})$ and $\mathbb{P}$-as, as $n \to \infty$.*

Proof. We will show separately the $L^2(\mathbb{P})$ and $\mathbb{P}$-as convergences.

▸ *Claim 1. $\nabla F(\theta_n) \to 0$ in $L^2(\mathbb{P})$ as $n \to \infty$.*
▸ *Step 1.1. Reductio ad absurdum.*
In view of $\mathcal{H}1$, Jensen's inequality and Lemma 2.1(iii),

$$\lim_{n\to\infty} \mathbb{E}\big[\|\nabla F(\theta_n) - \nabla F(\theta_{n-1})\|^2\big]^{\frac{1}{2}} \leq L \lim_{n\to\infty} \mathbb{E}[\|\theta_n - \theta_{n-1}\|^{1+\alpha}]^{\frac{\alpha}{1+\alpha}} = 0. \tag{2.8}$$

Recall that $\liminf_{n\to\infty}\mathbb{E}\big[\|\nabla F(\theta_n)\|^2\big]=0$ according to Lemma 2.1(ii). Suppose, by contradiction, that $\limsup_{n\to\infty}\mathbb{E}\big[\|\nabla F(\theta_n)\|^2\big]^{\frac{1}{2}}>\eta$ for some $\eta>0$. Then, via Lemma A.2, there exist extractions $\chi,\psi\colon\mathbb{N}\to\mathbb{N}$ such that, for all $n\geq 0$, $\chi(n)<\psi(n)$,

$$\mathbb{E}\big[\|\nabla F(\theta_{\chi(n)})\|^2\big]^{\frac{1}{2}}<\frac{\eta}{3},\quad \mathbb{E}\big[\|\nabla F(\theta_{\psi(n)})\|^2\big]^{\frac{1}{2}}>\frac{2\eta}{3}, \tag{2.9}$$

$$\frac{\eta}{3}\leq\mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big]^{\frac{1}{2}}\leq\frac{2\eta}{3},\quad \chi(n)<k<\psi(n), \tag{2.10}$$

and, taking into consideration (2.8),

$$\mathbb{E}\big[\|\nabla F(\theta_k)-\nabla F(\theta_{k-1})\|^2\big]^{\frac{1}{2}}\leq\frac{\eta}{6},\quad k\geq\chi(0). \tag{2.11}$$

▶ *Step 1.2. Subsequent properties.*
Let $n\geq 0$. Via (2.9), the triangle inequality, $\mathcal{H}1$, Jensen's inequality and (2.6),

$$\begin{aligned}\Big(\frac{\eta}{3}\Big)^{\frac{1}{\alpha}}&\leq\big|\mathbb{E}\big[\|\nabla F(\theta_{\psi(n)})\|^2\big]^{\frac{1}{2}}-\mathbb{E}\big[\|\nabla F(\theta_{\chi(n)})\|^2\big]^{\frac{1}{2}}\big|^{\frac{1}{\alpha}}\leq L^{\frac{1}{\alpha}}\mathbb{E}[\|\theta_{\psi(n)}-\theta_{\chi(n)}\|^2]^{\frac{1}{2}}\\&\leq L^{\frac{1}{\alpha}}\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k\mathbb{E}\big[\|g_k(\theta_{k-1})\|^2\big]^{\frac{1}{2}}\leq(K\vee 1)L^{\frac{1}{\alpha}}\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k,\end{aligned}$$

hence

$$\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k\geq\frac{1}{K\vee 1}\Big(\frac{\eta}{3L}\Big)^{\frac{1}{\alpha}}. \tag{2.12}$$

Moreover, using (2.11), $|\mathbb{E}[\|\nabla F(\theta_{\chi(n)+1})\|^2]^{\frac{1}{2}}-\mathbb{E}[\|\nabla F(\theta_{\chi(n)})\|^2]^{\frac{1}{2}}|\leq\frac{\eta}{6}$, so that, by (2.10), $\mathbb{E}[\|\nabla F(\theta_{\chi(n)})\|^2]^{\frac{1}{2}}\geq\frac{\eta}{6}$. Thus

$$\mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big]\geq\Big(\frac{\eta}{6}\Big)^2,\quad \chi(n)\leq k<\psi(n). \tag{2.13}$$

▶ *Step 1.3. Contradiction.*
Owing to (2.12) and (2.13),

$$0<\frac{1}{K\vee 1}\Big(\frac{\eta}{3L}\Big)^{\frac{1}{\alpha}}\Big(\frac{\eta}{6}\Big)^2\leq\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\big]\leq\sum_{k=\chi(n)+1}^{\infty}\gamma_k\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\big].$$

Considering Lemma 2.1(ii), the right hand side of the above inequality tends to 0 as n goes to ∞, which is contradictory.

▶ *Claim 2.* $\nabla F(\theta_n)\to 0$ $\mathbb{P}$*-as as* $n\to\infty$.
▶ *Step 2.1. Reductio ad absurdum.*
By Lemma 2.1(i), $(F(\theta_n))_{n\geq 1}$ is $\mathbb{P}$-as bounded. Hence, by Lemma A.1(ii), for all $k\geq 1$,

$$\|\nabla F(\theta_{k-1})\|\leq\Big(\frac{1+\alpha}{\alpha}\Big)^{\frac{\alpha}{1+\alpha}}L^{\frac{1}{1+\alpha}}\sup_{n\geq 0}\big(F(\theta_{n-1})-\inf F\big)^{\frac{\alpha}{1+\alpha}}=:\widetilde{K}<\infty\quad\mathbb{P}\text{-as.} \tag{2.14}$$

Moreover, via $\mathcal{H}1$ and Lemma 2.1(iii),

$$\lim_{n\to\infty}\|\nabla F(\theta_n)-\nabla F(\theta_{n-1})\|\leq L\lim_{n\to\infty}\|\theta_n-\theta_{n-1}\|^\alpha=0\quad\mathbb{P}\text{-as.} \tag{2.15}$$

Recall that, by Lemma 2.1(i), $\liminf_{n\to\infty}\|\nabla F(\theta_n)\| = 0$. Suppose, by contradiction, that $\mathbb{P}(\limsup_{n\to\infty}\|\nabla F(\theta_n)\| > \eta) > 0$ for some $\eta > 0$. Denote $\Lambda = \{\limsup_{n\to\infty}\|\nabla F(\theta_n)\| > \eta\}$. Then, via Lemma A.2, $\mathbb{P}$-as on Λ, there exist extractions $\chi, \psi\colon \mathbb{N}\to\mathbb{N}$ such that, for all $n \geq 0$, $\chi(n) < \psi(n)$,

$$\|\nabla F(\theta_{\chi(n)})\| < \frac{\eta}{3}, \quad \|\nabla F(\theta_{\psi(n)})\| > \frac{2\eta}{3} \tag{2.16}$$

$$\frac{\eta}{3} \leq \|\nabla F(\theta_k)\| \leq \frac{2\eta}{3}, \quad \chi(n) < k < \psi(n), \tag{2.17}$$

and, in view of (2.15) and Lemma 2.1(iv),

$$\|\nabla F(\theta_k) - \nabla F(\theta_{k-1})\| \leq \frac{\eta}{6}, \quad \Big\|\sum_{j=k}^{\infty}\gamma_j\big(g_j(\theta_{j-1}) - F(\theta_{j-1})\big)\Big\| \leq \frac{1}{4L^{\frac{1}{\alpha}}}\Big(\frac{\eta}{3}\Big)^{\frac{1}{\alpha}}, \quad k \geq \chi(0). \tag{2.18}$$

▸ *Step 2.2. Subsequent properties.*
We put ourselves in Λ and consider $n \geq 0$. Via (2.16), the triangle inequality, $\mathcal{H}1$, (2.18) and (2.14),

$$\begin{aligned}\Big(\frac{\eta}{3}\Big)^{\frac{1}{\alpha}} &\leq \big|\|\nabla F(\theta_{\psi(n)})\| - \|\nabla F(\theta_{\chi(n)})\|\big|^{\frac{1}{\alpha}} \leq L^{\frac{1}{\alpha}}\|\theta_{\psi(n)} - \theta_{\chi(n)}\| \\ &\leq L^{\frac{1}{\alpha}}\Big\|\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k\big(g_k(\theta_{k-1}) - \nabla F(\theta_{k-1})\big)\Big\| + L^{\frac{1}{\alpha}}\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k\big\|\nabla F(\theta_{k-1})\big\| \\ &\leq \frac{1}{2}\Big(\frac{\eta}{3}\Big)^{\frac{1}{\alpha}} + (\widetilde{K}\vee 1)L^{\frac{1}{\alpha}}\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k,\end{aligned}$$

hence

$$\sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k \geq \frac{1}{2(\widetilde{K}\vee 1)}\Big(\frac{\eta}{3L}\Big)^{\frac{1}{\alpha}}. \tag{2.19}$$

Besides, via (2.18), $\big|\|\nabla F(\theta_{\chi(n)+1})\| - \|\nabla F(\theta_{\chi(n)})\|\big| \leq \frac{\eta}{6}$, hence, by (2.17), $\|\nabla F(\theta_{\chi(n)})\| \geq \frac{\eta}{6}$. Therefore,

$$\|\nabla F(\theta_k)\|^2 \geq \Big(\frac{\eta}{6}\Big)^2, \quad \chi(n) \leq k < \psi(n). \tag{2.20}$$

▸ *Step 2.3. Contradiction.*
Owing to (2.19) and (2.20),

$$0 < \frac{1}{2(\widetilde{K}\vee 1)}\Big(\frac{\eta}{3L}\Big)^{\frac{1}{\alpha}}\Big(\frac{\eta}{6}\Big)^2 \leq \sum_{k=\chi(n)+1}^{\psi(n)}\gamma_k\|\nabla F(\theta_{k-1})\|^2 \leq \sum_{k=\chi(n)+1}^{\infty}\gamma_k\|\nabla F(\theta_{k-1})\|^2.$$

By Lemma 2.1(i), the right hand side above converges $\mathbb{P}$-as on Λ to 0 as $n\to\infty$, which is absurd. □

Proof of Theorem 1.5. See Lemmas 2.1(i,v) and 2.3. □

Proof of Lemma 1.6. Since, by Lemma 2.1(i), $(F(\theta_n))_{n\geq 0}$ is $\mathbb{P}$-as convergent, it is $\mathbb{P}$-as bounded. There then exists $\mathbb{P}$-as some $\eta > 0$ such that, for all $n \geq 0$, $F(\theta_n) \leq \eta$ $\mathbb{P}$-as. By $\mathcal{H}4$, there exists $\mathbb{P}$-as $\rho > 0$ such that, for all $\theta \in \mathbb{R}^m$ such that $\|\theta\| > \rho$, it holds $F(\theta) > \eta$ $\mathbb{P}$-as. Thus, for all $n \geq 0$, $\|\theta_n\| \leq \rho$ $\mathbb{P}$-as. □

Proof of Theorem 1.7. **(i)** Via Lemmas 2.1(iii) and 1.6, $\lim_{n\to\infty} \theta_n - \theta_{n-1} = 0$ $\mathbb{P}$-as and $(\theta_n)_{n\geq 0}$ is $\mathbb{P}$-as bounded. Thus, by Ostrowski's lemma [Ost74, Theorem 26.1], the set of accumulation points of $(\theta_n)_{n\geq 1}$ is $\mathbb{P}$-as a nonempty compact connected set $\mathcal{A}$.

According to Lemma 2.3, there exists $\Lambda \in \mathcal{F}$, $\mathbb{P}(\Lambda) = 1$, such that for all $\omega \in \Lambda$, $\lim_{n\to\infty} \nabla F(\theta_n(\omega)) = 0$. Let $\omega \in \Lambda$, $\theta_\star \in \mathcal{A}(\omega)$ and an extraction $\psi\colon \mathbb{N} \to \mathbb{N}$ such that $\lim_{n\to\infty} \theta_{\psi(n)}(\omega) = \theta_\star$. Then, by the continuity of ∇F under $\mathcal{H}1$, one has $\nabla F(\theta_\star) = \lim_{n\to\infty} \nabla F(\theta_{\psi(n)}(\omega)) = 0$.

(ii) Since $\mathcal{A} \subset \nabla F^{-1}(0)$, $\mathcal{A}$ is $\mathbb{P}$-as at most countable. Being $\mathbb{P}$-as-compact, connected and at most countable, $\mathcal{A}$ necessarily reduces $\mathbb{P}$-as to a singleton $\{\theta_\star\}$. Finally, being bounded and possessing a unique accumulation point, $(\theta_n)_{n\geq 0}$ converges $\mathbb{P}$-as to $\theta_\star$. □

3 Convergence Rate Proofs

Proof of Proposition 1.11. **(i)a.** By [DK24, Lemma 4.3(i)], the set of critical levels $\{F(\theta), \theta \in \nabla F^{-1}(0) \cap \mathcal{K}\}$ is finite, say, of the form $(\ell_i)_{1\leq i\leq I}$, $I \geq 1$. According to [DK24, Lemma 4.3(ii)], for all $1 \leq i \leq I$, there exist $\mathcal{V}_i \supset F^{-1}(\ell_i) \cap \nabla F^{-1}(0) \cap \mathcal{K}$ an open neighborhood, $\beta_i \in (0,1)$ and $\zeta_i > 0$ such that $|F(\theta) - \ell_i|^{\beta_i} \leq \zeta_i \|\nabla F(\theta)\|$, $\theta \in \mathcal{V}_i$. By the continuity of ∇F, we can assume up to a shrinking of the $(\mathcal{V}_i)_{1\leq i\leq I}$ that $\|\nabla F(\theta)\| \leq 1$, $\theta \in \mathcal{V}_i$, $1 \leq i \leq I$. Set $\beta = \max_i \beta_i \in (0,1)$ and $\zeta = (1 \vee \max_i \zeta_i^{1/\beta_i})^\beta > 0$. Then $|F(\theta) - \ell_i|^\beta \leq \zeta \|\nabla F(\theta)\|$, $\theta \in \mathcal{V}_i$, $1 \leq i \leq I$. Finally, letting $\theta_\star \in \nabla F^{-1}(0) \cap \mathcal{K}$, one has $F(\theta_\star) = \ell_i$ for some $1 \leq i \leq I$, so that $\theta_\star \in \mathcal{V}_i \cap \nabla F^{-1}(0)$ and $|F(\theta) - F(\theta_\star)|^\beta = |F(\theta) - \ell_i|^\beta \leq \zeta \|\nabla F(\theta)\|$, $\theta \in \mathcal{V}_i$, thus the result.

(i)b. The sought property follows by the previous result and Lemma A.5.

(ii) Since F is Łojasiewicz, to any $\theta_\star \in \nabla F^{-1}(0)$, we can associate a triplet $(\mathcal{V}_{\theta_\star}, \beta_{\theta_\star}, \zeta_{\theta_\star}) \in \mathcal{B}(\mathbb{R}^m) \times (0,1) \times \mathbb{R}^+$ such that $\mathcal{V}_{\theta_\star}$ is an open neighborhood of $\theta_\star$ and

$$|F(\theta) - F(\theta_\star)|^{\beta_{\theta_\star}} \leq \zeta_{\theta_\star} \|\nabla F(\theta)\|, \quad \theta \in \mathcal{V}_{\theta_\star}.$$

Take

$$\beta = \frac{1}{2}\big(\sup_{\theta_\star \in \nabla F^{-1}(0)} \beta_{\theta_\star} + 1\big) \in (0,1].$$

Then $\beta > \beta_{\theta_\star}$, $\theta_\star \in \nabla F^{-1}(0)$, hence the sought property by invoking Lemma A.5.

(iii) Let $\theta_\star \in \nabla F^{-1}(0)$. Then $(F(\theta_\star) - \inf F)^\beta \leq \zeta \|\nabla F(\theta_\star)\| = 0$ for some $(\beta, \zeta) \in (0,1) \times \mathbb{R}_+^*$. A similar observation to Remark 1.8 ensures that $\beta \geq \frac{\alpha}{1+\alpha}$. □

Lemma 3.1. *Assume $\mathcal{H}1$–4 hold. Then, for all $\delta \in (0,1)$, there exists $\ell_\delta > \inf F$ such that $\mathbb{P}(\mathcal{A} \subset \mathcal{L}_{\ell_\delta}) \geq 1 - \delta$.*

Proof. ▸ *Step 1. Final hitting time.*
Consider the random time

$$\tau_0 \coloneqq \inf\{n \geq 0 \colon \forall k \geq n, \|\nabla F(\theta_k)\| \leq 1\}, \tag{3.1}$$

with the convention $\inf \varnothing = \infty$. According to Lemma 2.3, $\tau_0 < \infty$ $\mathbb{P}$-as.

Let $\delta \in (0,1)$. The reverse Fatou lemma ensures that

$$\limsup_{n\to\infty} \mathbb{P}(\tau_0 \geq n) \leq \mathbb{P}\big(\limsup_{n\to\infty} \{\tau_0 \geq n\}\big) = 0.$$

Thus, there exists $n_\delta \geq 0$ such that

$$\mathbb{P}(\tau_0 < n_\delta) \geq 1 - \frac{\delta}{2}. \tag{3.2}$$

According to Lemmas 2.1(i) and 1.6 and Theorem 1.7(i), there exists $\Lambda \in \mathcal{F}$ such that $\mathbb{P}(\Lambda) = 1$ and, for all $\omega \in \Lambda$, $F(\theta_n(\omega)) \to F_\star(\omega)$, $(\theta_n(\omega))_{n\geq 0}$ is bounded and its set of accumulation points $\mathcal{A}(\omega)$ is a compact connected set. Let $\omega \in \Lambda$. By the continuity of F via $\mathcal{H}1$, for all $\theta_\star \in \mathcal{A}(\omega)$, $F(\theta_\star) = \lim_{n\to\infty} F(\theta_{\psi(n)}(\omega)) = F_\star(\omega)$, where $\psi\colon \mathbb{N} \to \mathbb{N}$ is an extraction such that $\theta_{\psi(n)}(\omega) \to \theta_\star$ as $n \to \infty$. F is therefore constant on $\mathcal{A}(\omega)$ equal to $F_\star(\omega)$.

Let $\ell > \inf F$. The previous discussion entails

$$\{\mathcal{A} \subset \mathcal{L}_\ell\} = \{F_\star \leq \ell\} = \big\{\forall \varepsilon \in \mathbb{Q}_+^*, \exists p \geq n_\delta, \forall n \geq p, F(\theta_n) \leq \ell + \varepsilon\big\}. \tag{3.3}$$

▸ *Step 2. High probability bound.*
Let $n > n_\delta$. By (3.3) and Markov's inequality,

$$\begin{aligned}\mathbb{P}(\mathcal{A} \not\subset \mathcal{L}_\ell, \tau_0 < n_\delta) &= \mathbb{P}\big(\big\{\exists \varepsilon \in \mathbb{Q}_+^*, \forall p \geq n_\delta, \exists n \geq p, F(\theta_n) > \ell + \varepsilon\big\} \cap \{\tau_0 < n_\delta\}\big) \\ &\leq \inf_{p\geq n_\delta} \frac{\mathbb{E}[\sup_{n\geq p} \sup_{\varepsilon\in\mathbb{Q}_+^*} (F(\theta_n) - \inf F - \varepsilon)^+ \mathbb{1}_{\tau_0<n_\delta}]}{\ell - \inf F} \leq \frac{\mathbb{E}[\sup_{n\geq n_\delta}(F(\theta_n) - \inf F)\mathbb{1}_{\tau_0<n_\delta}]}{\ell - \inf F}.\end{aligned} \tag{3.4}$$

Let $(\Delta_n)_{n\geq 1}$ be the $(\mathcal{F}_n)_{n\geq 0}$-martingale increment sequence given by

$$\Delta_n = \nabla F(\theta_{n-1})^\top \big(\nabla F(\theta_{n-1}) - g_n(\theta_{n-1})\big)\mathbb{1}_{\|\nabla F(\theta_{n-1})\|\leq 1}, \quad n \geq 1. \tag{3.5}$$

Using (2.1),

$$\begin{aligned}&(F(\theta_n) - \inf F)\mathbb{1}_{\tau_0<n_\delta} \\ &\qquad \leq \Big(F(\theta_{n-1}) - \inf F - \gamma_n\|\nabla F(\theta_{n-1})\|^2 + \gamma_n\Delta_n + \frac{L}{1+\alpha}\gamma_n^{1+\alpha}\|g_n(\theta_{n-1})\|^{1+\alpha}\Big)\mathbb{1}_{\tau_0<n_\delta},\end{aligned}$$

hence, by telescoping,

$$\begin{aligned}&(F(\theta_n) - \inf F)\mathbb{1}_{\tau_0<n_\delta} \\ &\qquad \leq \Big(F(\theta_{n_\delta-1}) - \inf F + \sum_{k=n_\delta}^{n} \gamma_k\Delta_k + \frac{L}{1+\alpha}\sum_{k=n_\delta}^{n} \gamma_k^{1+\alpha}\|g_k(\theta_{k-1})\|^{1+\alpha}\Big)\mathbb{1}_{\tau_0<n_\delta}.\end{aligned}$$

Therefore, by Fubini's theorem,

$$\begin{aligned}&\mathbb{E}\Big[\sup_{n\geq n_\delta}(F(\theta_n) - \inf F)\mathbb{1}_{\tau<n_\delta}\Big] \\ &\qquad \leq \sup_{n\geq 0}\mathbb{E}[F(\theta_n) - \inf F] + \mathbb{E}\Big[\sup_{n>n_\delta}\Big|\sum_{k=n_\delta}^{n}\gamma_k\Delta_k\Big|\Big] + \frac{L\sup_{n\geq 1}\mathbb{E}[\|g_n(\theta_{n-1})\|^{1+\alpha}]}{1+\alpha}\sum_{k=1}^{\infty}\gamma_k^{1+\alpha}.\end{aligned} \tag{3.6}$$

By Lemma 2.1(ii), $\sup_{n\geq 0}\mathbb{E}[F(\theta_n) - \inf F] < \infty$. Moreover, using that $x^{1+\alpha} \leq C(1+x^2)$, $x \in \mathbb{R}$, and $\mathcal{H}2$,

$$\sup_{n\geq 1}\mathbb{E}\big[\|g_n(\theta_{n-1})\|^{1+\alpha}\big] \leq C\big(\sup_{n\geq 1}\mathbb{E}[F(\theta_{n-1}) - \inf F] + 1\big) < \infty. \tag{3.7}$$

Besides, by the definition (3.5), using the Cauchy-Schwarz inequality, Jensen's inequality and $\mathcal{H}2$,

$$\sup_{n\geq 1}\mathbb{E}\big[\Delta_n^2\big] \leq 2\big(\sup_{n\geq 1}\mathbb{E}\big[\|g_n(\theta_{n-1})\|^2\big] + 1\big) \leq C\big(\sup_{n\geq 1}\mathbb{E}[F(\theta_{n-1}) - \inf F] + 1\big) < \infty, \quad n \geq 1. \tag{3.8}$$

Hence, by monotone convergence, Jensen's inquality, Doob's maximal inequality (or the Burkholder-Davis-Gundy inequality), (3.8) and $\mathcal{H}3$,

$$\mathbb{E}\Big[\sup_{n>n_\delta}\Big|\sum_{k=n_\delta}^{n}\gamma_k\Delta_k\Big|\Big] = \lim_{N\to\infty}\mathbb{E}\Big[\sup_{n_\delta<n\leq N}\Big|\sum_{k=n_\delta}^{n}\gamma_k\Delta_k\Big|^2\Big]^{\frac{1}{2}} \leq \Big(\sum_{k=1}^{\infty}\gamma_k^2\mathbb{E}[\Delta_k^2]\Big)^{\frac{1}{2}} < \infty. \tag{3.9}$$

Coming back to (3.4), by combining (3.6), (3.7) and (3.9),

$$\mathbb{P}(\mathcal{A} \not\subset \mathcal{L}_\ell, \tau_0 < n_\delta) \leq \frac{C}{\ell - \inf F} \to 0 \quad \text{as } \ell \to \infty. \tag{3.10}$$

▸ *Step 3. Conclusion.*
(3.10) shows that there exists $\ell_\delta > \inf F$ such that

$$\mathbb{P}(\mathcal{A} \not\subset \mathcal{L}_{\ell_\delta}, \tau_0 < n_\delta) \leq \frac{\delta}{2}. \tag{3.11}$$

Consequently, from (3.2) and (3.11),

$$\mathbb{P}(\mathcal{A} \subset \mathcal{L}_{\ell_\delta}) \geq \mathbb{P}(\mathcal{A} \subset \mathcal{L}_{\ell_\delta}, \tau_0 < n_\delta) \geq 1 - \delta.$$

□

The next result deals with the convergence rate of the gradients $(\|\nabla F(\theta_n)\|)_{n\geq 0}$ and requires only $\mathcal{H}1$–3. It is provided here help prove our first (SGD) convergence rate in Theorem 1.14.

Lemma 3.2. *Suppose $\mathcal{H}1$–3 hold. Then,*

(i) *there exists a constant $C > 0$ such that, for all $0 \leq p \leq n$,*

$$\inf_{p\leq k\leq n} \mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big] \leq C\Big(\sum_{k=p+1}^{n+1} \gamma_k\Big)^{-1}.$$

(ii) *there exists a constant $C > 0$ such that, for all $\delta \in (0,1)$, for all $0 \leq p \leq n$,*

$$\inf_{p\leq k\leq n} \|\nabla F(\theta_k)\| \leq \frac{C}{\delta}\Big(\sum_{k=p+1}^{n+1} \gamma_k\Big)^{-\frac{1}{2}} \quad \textit{with probability at least } 1-\delta.$$

The first part of Lemma 3.2 is mirrored for instance in [GG24, Theorem 5.12]. It is somewhat a more generic result than [Lei+20, Theorem 2(a)], as the left end indexation on k starts at an arbitrary integer p, not necessarily null.

Proof of Lemma 3.2. **(i)** From Lemma 2.1(ii),

$$\inf_{p\leq k\leq n} \mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big] \sum_{k=p+1}^{n+1} \gamma_k \leq \sum_{k=1}^{\infty} \gamma_k \mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\big] =: \Xi < \infty. \quad 0 \leq p \leq n.$$

(ii) Let $\delta \in (0,1)$. By the Markov and Jensen inequalities,

$$\mathbb{P}\Big(\inf_{p\leq k\leq n} \|\nabla F(\theta_k)\| \geq \frac{\sqrt{\Xi}}{\delta}\Big(\sum_{k=p+1}^{n+1}\Big)^{-\frac{1}{2}}\Big) \leq \delta.$$

□

Proof of Theorem 1.14. **(i)a.** ▸ *Step 1. Final hitting time.*
Let $\delta \in (0,1)$. According to Lemma 3.1, there exists $\ell_\delta > \inf F$ such that

$$\mathbb{P}(\mathcal{A} \subset \mathcal{L}_{\ell_\delta}) \geq 1 - \frac{\delta}{3}. \tag{3.12}$$

Besides, by Proposition 1.11(i), there exist $(\beta_\delta, \zeta_\delta) \in (0,1)\times\mathbb{R}_+^*$ such that, for all $\theta_\star \in \nabla F^{-1}(0)\cap \mathcal{L}_{\ell_\delta}$, there exists an open neighborhood $\mathcal{V}_\delta(\theta_\star)$ of $\theta_\star$ such that $|F(\theta)-F(\theta_\star)|^{\beta_\delta} \le \zeta_\delta\|\nabla F(\theta)\|$, $\theta \in \mathcal{V}_\delta(\theta_\star)$. Furthermore, by Theorem 1.7(i), there exists $\Lambda \in \mathcal{F}$ such that $\mathbb{P}(\Lambda)=1$ and, for all $\omega\in\Lambda$, the set of accumulation $\mathcal{A}(\omega)$ is a compact connected set. Recalling that $F(\theta_\star)=F_\star(\omega)$, $\theta_\star \in \mathcal{A}(\omega)$, note that

$$|F(\theta)-F_\star(\omega)|^{\beta_\delta} \le \zeta_\delta\|\nabla F(\theta)\|, \quad \theta\in\mathcal{V}_\delta(\theta_\star),\ \theta_\star\in\mathcal{A}(\omega),\ \omega\in\Lambda\cap\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}. \tag{3.13}$$

Define $\mathcal{V}_{\delta,\star}(\omega) \coloneqq \cup_{\theta_\star\in\mathcal{A}(\omega)}\mathcal{V}_\delta(\theta_\star)$, $\omega\in\Lambda\cap\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}$. Note that, for all $\omega\in\Lambda\cap\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}$, $\mathcal{A}(\omega)\subsetneq\mathcal{V}_{\delta,\star}(\omega)$, so that $\mathcal{V}_{\delta,\star}(\omega)\neq\varnothing$. Let

$$\tau_1 \coloneqq \begin{cases} \tau_0 \vee \inf\{n\ge 0\colon \forall k\ge n, \theta_k\in\mathcal{V}_{\delta,\star}\} & \text{on } \{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\},\\ \infty & \text{on } \{\mathcal{A}\not\subset\mathcal{L}_{\ell_\delta}\},\end{cases} \tag{3.14}$$

with the convention $\inf\varnothing=\infty$. Via the $\mathbb{P}$-as-boundedness of $(\theta_n)_{n\ge0}$ by Lemma 1.6, one has $\tau_1<\infty$ $\mathbb{P}$-as on $\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}$. Hence, by the reverse Fatou lemma,

$$\limsup_{n\to\infty}\mathbb{P}(\tau_1\ge n, \mathcal{A}\subset\mathcal{L}_{\ell_\delta}) \le \mathbb{P}\big(\limsup_{n\to\infty}\{\tau_1\ge n\}\cap\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}\big)=0.$$

Via (3.12), there then exists $n_\delta\ge0$ such that

$$\mathbb{P}(\tau_1<n_\delta, \mathcal{A}\subset\mathcal{L}_{\ell_\delta}) \ge \mathbb{P}(\mathcal{A}\subset\mathcal{L}_{\ell_\delta}) - \frac{\delta}{3} \ge 1-\frac{2}{3}\delta. \tag{3.15}$$

▸ *Step 2. $L^1(\mathbb{P})$ bound on $(F(\theta_n))_{n\ge0}$.*
Introduce

$$\mu_\delta \coloneqq \zeta_\delta^{-\frac{1}{\beta_\delta}} > 0 \tag{3.16}$$

and

$$\mathbb{A}_\delta \coloneqq \{\tau_1<n_\delta\}\cap\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}. \tag{3.17}$$

Let $n_\delta\le k\le n$. Observe that, by (3.1) and (3.14), $\|\nabla F(\theta_k)\|\le1$ on $\mathbb{A}_\delta$. Hence, using (3.13),

$$\begin{aligned}
&\mu_\delta\mathbb{E}[|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}]\\
&\qquad\le \mathbb{E}\big[\|\nabla F(\theta_k)\|^{\frac{1}{\beta_\delta}}\mathbb{1}_{\mathbb{A}_\delta}\big]\\
&\qquad\le \begin{cases}\mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big] & \text{if } \beta_\delta\in(0,\frac12], \text{ since } \|\nabla F(\theta_k)\|\le1,\\ \mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big]^{\frac{1}{2\beta_\delta}} & \text{if } \beta_\delta\in(\frac12,1), \text{ via Jensen's inequality,}\end{cases}\\
&\qquad= \mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big]^{1\wedge\frac{1}{2\beta_\delta}}.
\end{aligned}$$

Therefore, via Lemma 3.2(i),

$$\inf_{n_\delta\le k\le n}\mathbb{E}[|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}] \le \mu_\delta^{-1}\inf_{n_\delta\le k\le n}\mathbb{E}\big[\|\nabla F(\theta_k)\|^2\big]^{1\wedge\frac{1}{2\beta_\delta}} \le C_\delta\Big(\sum_{k=n_\delta+1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta_\delta}}.$$

Thus

$$\mathbb{E}\big[\inf_{0\le k\le n}|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}\big] \le \inf_{n_\delta\le k\le n}\mathbb{E}[|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}] \le C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta_\delta}},$$

where we used that $\sum_{k=1}^{n+1}\gamma_k\sim\sum_{k=n_\delta+1}^{n+1}\gamma_k$ as $n\to\infty$.

Finally, up to a modification of C_δ, for all $n \geq 0$,

$$\begin{aligned}
&\mathbb{E}\big[\inf_{0\leq k\leq n}|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}\big]\\
&\quad=\mathbb{E}\big[\inf_{0\leq k\leq n}|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}\big]\mathbb{1}_{n<n_\delta}+\mathbb{E}\big[\inf_{0\leq k\leq n}|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}\big]\mathbb{1}_{n\geq n_\delta}\\
&\quad\leq\Big(\sup_{k\geq 0}\mathbb{E}[|F(\theta_k)-F_\star|]\Big(\sum_{k=1}^{n_\delta}\gamma_k\Big)^{1\wedge\frac{1}{2\beta_\delta}}+C_\delta\Big)\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta_\delta}}\leq C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta_\delta}},
\end{aligned}\tag{3.18}$$

with the convention $\sum_{\varnothing}=0$, where we used that, due to Lemmas 2.1(ii,v),

$$\sup_{k\geq 0}\mathbb{E}[|F(\theta_k)-F_\star|]\leq\sup_{k\geq 0}\mathbb{E}[F(\theta_k)-\inf F]+|\inf F|+\mathbb{E}[|F_\star|]<\infty.\tag{3.19}$$

▸ *Step 3. High probability bound on* $(F(\theta_n))_{n\geq 0}$.
Reusing the constant C_δ from the right hand side of (3.18), by Markov's inequality and (3.17),

$$\begin{aligned}
\mathbb{P}\Big(\inf_{0\leq k\leq n}|F(\theta_k)-F_\star|>\frac{3}{\delta}C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta_\delta}},&\tau_1<n_\delta,\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\Big)\\
&\leq\frac{\delta\mathbb{E}[\inf_{0\leq k\leq n}|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{A}_\delta}]}{3C_\delta\big(\sum_{k=1}^{n+1}\gamma_k\big)^{-1\wedge\frac{1}{2\beta_\delta}}}\leq\frac{\delta}{3},\quad n\geq 0.
\end{aligned}$$

Therefore, recalling (3.15),

$$\begin{aligned}
\mathbb{P}\Big(\inf_{0\leq k\leq n}|F(\theta_k)-F_\star|\leq\frac{3}{\delta}C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta_\delta}}\Big)&\\
\geq\mathbb{P}\Big(\inf_{0\leq k\leq n}|F(\theta_k)-F_\star|\leq\frac{3}{\delta}C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta_\delta}},\tau_1<n_\delta,\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\Big)&\geq 1-\delta,\quad n\geq 0.
\end{aligned}$$

(i)b. The proof is identical to the previous one, once setting μ_δ to 1.

(ii) Reusing notation from the previous proof, let $\omega\in\Lambda$ and $\theta_\star\in\mathcal{A}(\omega)$. According to Proposition 1.11(ii), there exists $\beta\in(0,1]$ independent of ω and an open neighborhood $\mathcal{V}(\theta_\star)$ of $\theta_\star$ such that $|F(\theta)-F(\theta_\star)|^\beta\leq\|\nabla F(\theta)\|$, $\theta\in\mathcal{V}(\theta_\star)$. Define $\mathcal{V}_\star(\omega)\coloneqq\cup_{\theta_\star\in\mathcal{A}(\omega)}\mathcal{V}(\theta_\star)$. Given that $\mathcal{A}(\omega)\subsetneq\mathcal{V}_\star(\omega)$, one has $\mathcal{V}_\star(\omega)\neq\varnothing$. Recalling (3.1), consider

$$\tau_2\coloneqq\tau_0\vee\inf\{n\geq 0\colon\forall k\geq n,\theta_k\in\mathcal{V}_\star\},\tag{3.20}$$

with the convention $\inf\varnothing=\infty$. Since, according to Lemma 1.6, $(\theta_n)_{n\geq 0}$ is $\mathbb{P}$-as bounded, one has $\tau_2<\infty$ $\mathbb{P}$-as.

Let $\delta\in(0,1)$. By the reverse Fatou lemma,

$$\limsup_{n\to\infty}\mathbb{P}(\tau_2\geq n)\leq\mathbb{P}\big(\limsup_{n\to\infty}\{\tau_2\geq n\}\big)=0.$$

Therefore, there exists $n_\delta\geq 0$ such that

$$\mathbb{P}(\tau_2<n_\delta)\geq 1-\frac{\delta}{2}.\tag{3.21}$$

Take

$$\mathbb{B}_\delta\coloneqq\{\tau_2<n_\delta\}.\tag{3.22}$$

Performing the substitutions $\mathbb{A}_\delta \leftarrow \mathbb{B}_\delta$, $\beta_\delta \leftarrow \beta$ and $\mu_\delta \leftarrow 1$ in Step 2 of the preceding proof and applying the same lines of reasoning therein line by line yields

$$\mathbb{E}\big[\inf_{0\le k\le n}|F(\theta_k)-F_\star|\mathbb{1}_{\mathbb{B}_\delta}\big] \le C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta}}, \quad n\ge 0. \tag{3.23}$$

Fixing the C_δ above, via Markov's inequality, (3.23), (3.22) and (3.21),

$$\begin{aligned}\mathbb{P}\Big(\inf_{0\le k\le n}|F(\theta_k)-F_\star| &\le \frac{2}{\delta}C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta}}\Big)\\ &\ge \mathbb{P}\Big(\inf_{0\le k\le n}|F(\theta_k)-F_\star| \le \frac{2}{\delta}C_\delta\Big(\sum_{k=1}^{n+1}\gamma_k\Big)^{-1\wedge\frac{1}{2\beta}}, \tau_2<n_\delta\Big)\ge 1-\delta, \quad n\ge 0.\end{aligned} \tag{3.24}$$

(iii) It ensues from Proposition 1.11(iii) and Theorem 1.7(i) that $F_\star = \inf F$ $\mathbb{P}$-as. The result follows from (1.7) and Lemma 3.2(ii). □

Our proofs for recovering the convergence speeds of Theorem 1.17 are novel to literature and build upon two key ideas. First, the quasisupermartingale property of the function values is decomposed into a drift and a martingale terms (3.29,3.37). Their contributions are reflected in the exponents of the convergence rates displaying minima between multiple quantities. Second, a new concept of final hitting time (3.14,3.20) is leveraged, which, although not a stop time, allows to isolate the asymptotic regime in order to account for the Łojasiewicz property $\mathcal{H}5$. The interplay between the martingale term and the final hitting time is accounted for in the convergence speed (3.32,3.41). In the case of $\beta > \frac{1}{2}$, the convergence rate exponent writes as a minimum of three quantities: $\frac{1}{2\beta-1}$, $\alpha\rho-1$ and $\frac{\rho-1}{2}$. The first two stem from the drift contribution and the last one from the martingale contribution. The double contribution from the drift is due to the correction of order disparity between the quantities $F(\theta_n)-F_\star = \mathrm{O}(\|\nabla F(\theta_n)\|^{1/\beta})$ (as per $\mathcal{H}5$) and $\|\nabla F(\theta_n)\|^2$ resulting from the drift-martingale decomposition (3.36).

Proof of Theorem 1.17. Let $\delta\in(0,1)$.

(i)a. We reposition ourselves in the framework of Theorem 1.14(i)a's proof. Define

$$\mathbb{A}_\delta^n \coloneqq \mathbb{A}_\delta\cap\{F(\theta_n)>F_\star\} = \{\tau_1<n_\delta\}\cap\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}\cap\{F(\theta_n)>F_\star\}. \tag{3.25}$$

Let $n_\delta < k \le n$. Via (2.1) and (3.7),

$$\mathbb{E}[(F(\theta_k)-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}] \le \mathbb{E}[(F(\theta_{k-1})-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}] - \gamma_k\mathbb{E}\big[g_k(\theta_{k-1})^\top\nabla F(\theta_{k-1})\mathbb{1}_{\mathbb{A}_\delta^n}\big] + C\gamma_k^{1+\alpha} \tag{3.26}$$

▸ *Case 1.* $\beta_\delta\in\big(0,\frac{1}{2}\big]$.
▸ *Step 1.1.* $L^1(\mathbb{P})$ *bound on* $(F(\theta_n))_{n\ge0}$.
(3.26) rewrites

$$\begin{aligned}\mathbb{E}[(F(\theta_k)-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}] &\le (1-\mu_\delta\gamma_k)\mathbb{E}[(F(\theta_{k-1})-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}]\\ &\quad+\gamma_k\mathbb{E}\big[\big(\mu_\delta(F(\theta_{k-1})-F_\star)-g_k(\theta_{k-1})^\top\nabla F(\theta_{k-1})\big)\mathbb{1}_{\mathbb{A}_\delta^n}\big]+C\gamma_k^{1+\alpha}.\end{aligned} \tag{3.27}$$

Owing to (3.1) and (3.14), given that $k-1\ge n_\delta$, one has $\|\nabla F(\theta_{k-1})\|\le 1$ on $\mathbb{A}_\delta^n$. Thus, since $\beta_\delta\in\big(0,\frac{1}{2}\big]$, by (3.16) and (3.13),

$$\mu_\delta(F(\theta_{k-1})-F_\star) \le \|\nabla F(\theta_{k-1})\|^{\frac{1}{\beta_\delta}} \le \|\nabla F(\theta_{k-1})\|^2 \quad \text{on } \mathbb{A}_\delta^n.$$

Hence

$$\mathbb{E}\big[\big(\mu_\delta(F(\theta_{k-1})-F_\star)-g_k(\theta_{k-1})^\top\nabla F(\theta_{k-1})\big)\mathbb{1}_{\mathbb{A}_\delta^n}\big]\le\mathbb{E}\big[\nabla F(\theta_{k-1})^\top\big(\nabla F(\theta_{k-1})-g_k(\theta_{k-1})\big)\mathbb{1}_{\mathbb{A}_\delta^n}\big]. \tag{3.28}$$

Combining (3.27) and (3.28), recalling the notation (3.5),

$$\mathbb{E}[(F(\theta_k)-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}]\le(1-\mu_\delta\gamma_k)\mathbb{E}[(F(\theta_{k-1})-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}]+\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]+C\gamma_k^{1+\alpha}. \tag{3.29}$$

Note that, by (3.25), $(F(\theta_n)-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}=(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}$. Besides, since $\gamma_n\to0$ as $n\to\infty$, there exists $n'_\delta\ge0$ such that $1-\mu_\delta\gamma_n>0$, $n\ge n'_\delta$. Thus, via Lemma A.3, unrolling the inequality (3.29) over $n_\delta\vee n'_\delta=:p_\delta<k\le n$ yields

$$\begin{aligned}\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}]&\le\mathbb{E}[|F(\theta_{p_\delta})-F_\star|]\prod_{k=p_\delta+1}^n(1-\mu_\delta\gamma_k)\\&\quad+\sum_{k=p_\delta+1}^n\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]\prod_{j=k+1}^n(1-\mu_\delta\gamma_j)+C\sum_{k=p_\delta+1}^n\gamma_k^{1+\alpha}\prod_{j=k+1}^n(1-\mu_\delta\gamma_j),\end{aligned}$$

hence, using (3.19) and the inequality $1+x\le\mathrm{e}^x$, $x\in\mathbb{R}$,

$$\begin{aligned}\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}]&\le\big(\sup_{k\ge0}\mathbb{E}[|F(\theta_k)-F_\star|]\big)\exp\Big(\mu_\delta\sum_{k=1}^{p_\delta}\gamma_k\Big)\exp\Big(-\mu_\delta\sum_{k=1}^n\gamma_k\Big)\\&\quad+\sum_{k=p_\delta+1}^n\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]\prod_{j=k+1}^n(1-\mu_\delta\gamma_j)+C\sum_{k=1}^n\gamma_k^{1+\alpha}\exp\Big(-\mu_\delta\sum_{j=k+1}^n\gamma_j\Big),\end{aligned} \tag{3.30}$$

with the conventions $\prod_\varnothing=1$ and $\sum_\varnothing=0$.

Assuming that $\mu_\delta\gamma_\star>\alpha$ in the case $\mathcal{H}6$(i)b, Lemma A.4(i) gives

$$\exp\Big(-\mu_\delta\sum_{k=1}^n\gamma_k\Big)+\sum_{k=1}^n\gamma_k^{1+\alpha}\exp\Big(-\mu_\delta\sum_{j=k+1}^n\gamma_j\Big)\le C\gamma_n^\alpha. \tag{3.31}$$

Furthermore, using Jensen's inequality and that $(\Delta_n)_{n\ge1}$ are martingale increments,

$$\begin{aligned}&\sum_{k=p_\delta+1}^n\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]\prod_{j=k+1}^n(1-\mu_\delta\gamma_j)\\&\qquad\le\mathbb{E}\Big[\Big|\sum_{k=p_\delta+1}^n\gamma_k\Delta_k\prod_{j=k+1}^n(1-\mu_\delta\gamma_j)\Big|^2\Big]^{\frac12}=\Big(\sum_{k=p_\delta+1}^n\gamma_k^2\mathbb{E}[\Delta_k^2]\prod_{j=k+1}^n(1-\mu_\delta\gamma_j)^2\Big)^{\frac12}.\end{aligned}$$

Thus, supposing $2\mu_\delta\gamma_\star>1$ in the case of $\mathcal{H}6$(i)b, by (3.8) and Lemma A.4(i)a,

$$\sum_{k=p_\delta+1}^n\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]\prod_{j=k+1}^n(1-\mu_\delta\gamma_j)\le C\Big(\sum_{k=1}^n\gamma_k^2\exp\Big(-2\mu_\delta\sum_{j=k+1}^n\gamma_j\Big)\Big)^{\frac12}\le C_\delta\gamma_n^{\frac12}. \tag{3.32}$$

Gathering (3.30)–(3.32) therefore yields

$$\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}]\le C_\delta\gamma_n^{\alpha\wedge\frac12}. \tag{3.33}$$

By (3.33) and (3.19), up to a modification of C_δ, for $n\ge0$,

$$\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}]\le\big(\sup_{k\ge0}\mathbb{E}[|F(\theta_k)-F_\star|]\sup_{1\le k\le n_\delta}\gamma_k^{-\alpha\wedge\frac12}+C_\delta\big)\gamma_n^{\alpha\wedge\frac12}\le C_\delta\gamma_n^{\alpha\wedge\frac12}. \tag{3.34}$$

▸ *Step 1.2. High probability bound on* $(F(\theta_n))_{n\geq 0}$.
Let $n \geq 0$. Reusing the constant C_δ from the right hand side of (3.34), by Markov's inequality and (3.17),

$$\mathbb{P}\Big((F(\theta_n)-F_\star)^+ > \frac{3}{\delta}C_\delta\gamma_n^{\alpha\wedge\frac{1}{2}}, \tau_1 < n_\delta, \mathcal{A}\subset\mathcal{L}_{\ell_\delta}\Big) \leq \frac{\delta\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}]}{3C_\delta\gamma_n^{\alpha\wedge\frac{1}{2}}} \leq \frac{\delta}{3}.$$

Thus, via (3.15),

$$\mathbb{P}\Big((F(\theta_n)-F_\star)^+ \leq \frac{3}{\delta}C_\delta\gamma_n^{\alpha\wedge\frac{1}{2}}\Big) \geq \mathbb{P}\Big((F(\theta_n)-F_\star)^+ \leq \frac{3}{\delta}C_\delta\gamma_n^{\alpha\wedge\frac{1}{2}}, \tau_1 < n_\delta, \mathcal{A}\subset\mathcal{L}_{\ell_\delta}\Big) \geq 1-\delta.$$

▸ *Case 2.* $\beta_\delta \in \big(\frac{1}{2}, 1\big)$.
▸ *Step 2.1.* $L^1(\mathbb{P})$ *bound on* $(F(\theta_n))_{n\geq 0}$.
Reusing the notation (3.5), we have

$$-\gamma_k\mathbb{E}\big[g_k(\theta_{k-1})^\top\nabla F(\theta_{k-1})\mathbb{1}_{\mathbb{A}_\delta^n}\big] = \gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big] - \gamma_k\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big]. \tag{3.35}$$

Define the deterministic event

$$\mathfrak{S} \coloneqq \Big\{\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big]^{\frac{1}{2}} \geq \Big(\sum_{j=1}^k\gamma_j\Big)^{-\frac{\beta_\delta}{2\beta_\delta-1}}\Big\}.$$

Recalling that $2\beta_\delta > 1$, by (3.13) and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big] &\geq \mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big]^{\frac{2\beta_\delta-1}{2\beta_\delta}} \times \mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big]^{\frac{1}{2\beta_\delta}} \times \mathbb{1}_{\mathfrak{S}} \\
&\geq \frac{1}{\sum_{j=1}^k\gamma_j}\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big]^{\frac{1}{2\beta_\delta}}\mathbb{1}_{\mathfrak{S}} \\
&= \frac{1}{\sum_{j=1}^k\gamma_j}\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big]^{\frac{1}{2\beta_\delta}} - \frac{1}{\sum_{j=1}^k\gamma_j}\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|^2\mathbb{1}_{\mathbb{A}_\delta^n}\big]^{\frac{1}{2\beta_\delta}}\mathbb{1}_{\overline{\mathfrak{S}}} \\
&\geq \frac{\mu_\delta}{\sum_{j=1}^k\gamma_j}\mathbb{E}\big[|F(\theta_{k-1})-F_\star|^{2\beta_\delta}\mathbb{1}_{\mathbb{A}_\delta^n}\big]^{\frac{1}{2\beta_\delta}} - \frac{1}{\big(\sum_{j=1}^k\gamma_j\big)^{\frac{2\beta_\delta}{2\beta_\delta-1}}} \\
&\geq \frac{\mu_\delta}{\sum_{j=1}^k\gamma_j}\mathbb{E}[(F(\theta_{k-1})-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}] - \frac{1}{\big(\sum_{j=1}^k\gamma_j\big)^{\frac{2\beta_\delta}{2\beta_\delta-1}}}.
\end{aligned} \tag{3.36}$$

Therefore, combining (3.26), (3.35) and (3.36),

$$\begin{aligned}
\mathbb{E}[(F(\theta_k)-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}] \leq \Big(1-\frac{\mu_\delta\gamma_k}{\sum_{j=1}^k\gamma_j}\Big)\mathbb{E}[(F(\theta_{k-1})-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n}] \\
+ \frac{\gamma_k}{\big(\sum_{j=1}^k\gamma_j\big)^{\frac{2\beta_\delta}{2\beta_\delta-1}}} + \gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big] + C\gamma_k^{1+\alpha}.
\end{aligned} \tag{3.37}$$

Recall that $(F(\theta_n)-F_\star)\mathbb{1}_{\mathbb{A}_\delta^n} = (F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}$. Moreover, given that $\frac{\gamma_n}{\sum_{k=1}^n\gamma_k} \to 0$ as $n\to\infty$ by $\mathcal{H}3$, for some $n'_\delta \geq 0$, one has $1-\frac{\mu_\delta\gamma_n}{\sum_{k=1}^n\gamma_k} > 0$, $n \geq n'_\delta$. Thus, by Lemma A.3,

iterating (3.37) over $n_\delta \vee n'_\delta =: p_\delta < k \le n$ yields

$$\begin{aligned}\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}] &\le \mathbb{E}[|F(\theta_{p_\delta})-F_\star|]\prod_{k=p_\delta+1}^{n}\Big(1-\frac{\mu_\delta\gamma_k}{\sum_{j=1}^k\gamma_j}\Big)\\ &+C\sum_{k=p_\delta+1}^{n}\frac{\gamma_k}{\big(\sum_{j=1}^k\gamma_j\big)^{\frac{2\beta_\delta}{2\beta_\delta-1}}}\prod_{j=k+1}^{n}\Big(1-\frac{\mu_\delta\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)+\sum_{k=p_\delta+1}^{n}\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]\prod_{j=k+1}^{n}\Big(1-\frac{\mu_\delta\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)\\ &+C\sum_{k=p_\delta+1}^{n}\gamma_k^{1+\alpha}\prod_{j=k+1}^{n}\Big(1-\frac{\mu_\delta\gamma_j}{\sum_{i=1}^j\gamma_i}\Big).\end{aligned}$$

Hence, utilizing (3.19) and the inequality $1+x\le \mathrm{e}^x$, $x\in\mathbb{R}$,

$$\begin{aligned}\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}] &\le \big(\sup_{k\ge0}\mathbb{E}[|F(\theta_k)-F_\star|]\big)\exp\Big(\mu_\delta\sum_{k=1}^{p_\delta}\frac{\gamma_k}{\sum_{j=1}^k\gamma_j}\Big)\exp\Big(-\mu_\delta\sum_{k=1}^{n}\frac{\gamma_k}{\sum_{j=1}^k\gamma_j}\Big)\\ &+C\sum_{k=1}^{n}\frac{\gamma_k}{\big(\sum_{j=1}^k\gamma_j\big)^{\frac{2\beta_\delta}{2\beta_\delta-1}}}\exp\Big(-\mu_\delta\sum_{j=k+1}^{n}\frac{\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)+\sum_{k=p_\delta+1}^{n}\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]\prod_{j=k+1}^{n}\Big(1-\frac{\mu_\delta\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)\\ &+C\sum_{k=1}^{n}\gamma_k^{1+\alpha}\exp\Big(-\mu_\delta\sum_{j=k+1}^{n}\frac{\gamma_j}{\sum_{i=1}^j\gamma_i}\Big),\end{aligned}\tag{3.38}$$

with the conventions $\prod_\varnothing = 1$ and $\sum_\varnothing = 0$.

Under $\mathcal{H}6$(ii,iii), Lemma A.4(ii) gives

$$\exp\Big(-\mu_\delta\sum_{k=1}^{n}\frac{\gamma_k}{\sum_{j=1}^k\gamma_j}\Big)+\sum_{k=1}^{n}\frac{\gamma_k}{\big(\sum_{j=1}^k\gamma_j\big)^{\frac{2\beta_\delta}{2\beta_\delta-1}}}\exp\Big(-\mu_\delta\sum_{j=k+1}^{n}\frac{\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)\le\frac{C_\delta}{\big(\sum_{k=1}^n\gamma_k\big)^{\frac{1}{2\beta_\delta-1}}},\tag{3.39}$$

and also

$$\begin{aligned}&\sum_{k=1}^{n}\gamma_k^{1+\alpha}\exp\Big(-\mu_\delta\sum_{j=k+1}^{n}\frac{\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)\\ &\qquad\le\sum_{k=1}^{n}\frac{\gamma_k}{\big(\sum_{j=1}^k\gamma_j\big)^{\alpha\rho}}\exp\Big(-\mu_\delta\sum_{j=k+1}^{n}\frac{\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)\le\frac{C_\delta}{\big(\sum_{k=1}^n\gamma_k\big)^{\alpha\rho-1}}.\end{aligned}\tag{3.40}$$

Furthermore, using Jensen's inequality, that $(\Delta_n)_{n\ge1}$ are martingale increments, (3.8), $\mathcal{H}6$(ii,iii) and Lemma A.4(ii)a,

$$\begin{aligned}\sum_{k=p_\delta+1}^{n}\gamma_k\mathbb{E}\big[\Delta_k\mathbb{1}_{\mathbb{A}_\delta^n}\big]\prod_{j=k+1}^{n}\Big(1-\frac{\mu_\delta\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)&\le\Big(\sum_{k=1}^{n}\gamma_k^2\mathbb{E}\big[\Delta_k^2\big]\prod_{j=k+1}^{n}\Big(1-\frac{\mu_\delta\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)^2\Big)^{\frac12}\\ &\le C\Big(\sum_{k=1}^{n}\frac{\gamma_k}{\big(\sum_{j=1}^k\gamma_j\big)^{\rho}}\exp\Big(-2\mu_\delta\sum_{j=k+1}^{n}\frac{\gamma_j}{\sum_{i=1}^j\gamma_i}\Big)\Big)^{\frac12}\le\frac{C_\delta}{\big(\sum_{k=1}^n\gamma_k\big)^{\frac{\rho-1}{2}}}.\end{aligned}\tag{3.41}$$

Consequently, combining (3.38)–(3.41),

$$\mathbb{E}[(F(\theta_n)-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}]\le\frac{C_\delta}{\big(\sum_{k=1}^n\gamma_k\big)^{\frac{1}{2\beta_\delta-1}\wedge(\alpha\rho-1)\wedge\frac{\rho-1}{2}}}.$$

Let

$$r_\delta := \frac{1}{2\beta_\delta-1}\wedge(\alpha\rho-1)\wedge\frac{\rho-1}{2}>0,\quad\text{if }\beta_\delta\in\Big(\frac12,1\Big).\tag{3.42}$$

Recalling (3.19), up to a redefinition of C_δ, for all $n \geq 0$,

$$\mathbb{E}[(F(\theta_n) - F_\star)^+ \mathbb{1}_{\mathbb{A}_\delta}] \leq \frac{\sup_{k\geq 0} \mathbb{E}[|F(\theta_k) - F_\star|] \big(\sum_{k=1}^{p_\delta} \gamma_k\big)^{r_\delta} + C_\delta}{\big(\sum_{k=1}^n \gamma_k\big)^{r_\delta}} \leq \frac{C_\delta}{\big(\sum_{k=1}^n \gamma_k\big)^{r_\delta}}. \tag{3.43}$$

▸ *Step 2.2. High probability bound on* $(F(\theta_n))_{n\geq 0}$.
Reusing C_δ from the right hand side of (3.43), by Markov's inequality and (3.17), for all $n \geq 0$,

$$\mathbb{P}\Big((F(\theta_n) - F_\star)^+ > \frac{3}{\delta} C_\delta \Big(\sum_{k=1}^n \gamma_k\Big)^{r_\delta}, \tau_1 < n_\delta, \mathcal{A} \subset \mathcal{L}_{\ell_\delta}\Big) \leq \frac{\delta \mathbb{E}[(F(\theta_n) - F_\star)^+ \mathbb{1}_{\mathbb{A}_\delta}]}{3C_\delta \big(\sum_{k=1}^n \gamma_k\big)^{r_\delta}} \leq \frac{\delta}{3}.$$

All in all, via (3.15),

$$\begin{aligned}\mathbb{P}\Big((F(\theta_n) - F_\star)^+ &\leq \frac{3}{\delta} C_\delta \Big(\sum_{k=1}^n \gamma_k\Big)^{r_\delta}\Big) \\ &\geq \mathbb{P}\Big((F(\theta_n) - F_\star)^+ \leq \frac{3}{\delta} C_\delta \Big(\sum_{k=1}^n \gamma_k\Big)^{r_\delta}, \tau_1 < n_\delta, \mathcal{A} \subset \mathcal{L}_{\ell_\delta}\Big) \geq 1 - \delta.\end{aligned}$$

(i)b. The proof is similar to the prior one, once μ_δ is set to 1.

(ii) We put ourselves in the framework of the proof of Theorem 1.14(ii). Define

$$\mathbb{B}_\delta^n \coloneqq \mathbb{B}_\delta \cap \{F(\theta_n) > F_\star\} = \{\tau_2 < n_\delta\} \cap \{F(\theta_n) > F_\star\}. \tag{3.44}$$

Swapping $\beta_\delta \leftarrow \beta$, $\mu_\delta \leftarrow 1$ and $\mathbb{A}_\delta^n \leftarrow \mathbb{B}_\delta^n$ and setting $\gamma_\star > \alpha \vee \frac{1}{2}$ under $\mathcal{H}6$(i)b, the previous proof line by line shows that, for all $n \geq 0$,

$$\mathbb{E}[(F(\theta_n) - F_\star)^+ \mathbb{1}_{\mathbb{B}_\delta}] \leq C_\delta \begin{cases} \gamma_n^{\alpha \wedge \frac{1}{2}}, & \text{if } \beta \in (0, \frac{1}{2}], \\ \big(\sum_{k=1}^n \gamma_k\big)^{-r}, & \text{if } \beta \in (\frac{1}{2}, 1], \end{cases}$$

where

$$r \coloneqq \frac{1}{2\beta - 1} \wedge (\alpha\rho - 1) \wedge \frac{\rho - 1}{2} > 0, \quad \text{if } \beta \in \Big(\frac{1}{2}, 1\Big]. \tag{3.45}$$

The high probability bound is then retrieved via Markov's inequality similarly to (3.24).

(iii) According to Proposition 1.11(iii) and Theorem 1.7(i), one has $F_\star = \inf F$ $\mathbb{P}$-as. Besides, there exists $(\beta, \delta) \in \big[\frac{\alpha}{1+\alpha}, 1\big) \times \mathbb{R}_+^*$ such that (1.7) holds.

Let $n > 0$ and set

$$\mu = \zeta^{-\frac{1}{\beta}}. \tag{3.46}$$

Taking the mean in (2.2) and using (3.7),

$$\mathbb{E}[F(\theta_n) - \inf F] \leq \mathbb{E}[F(\theta_{n-1}) - \inf F] - \gamma_n \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big] + C\gamma_n^{1+\alpha}. \tag{3.47}$$

Assuming that $\beta \in \big(\frac{\alpha}{1+\alpha}, 1\big) \cup \big\{\frac{1}{2}\big\}$, we distinguish three cases in accordance with the values of α and β.

▸ *Case 1.* $\beta \in \big(\frac{\alpha}{1+\alpha}, \frac{1}{2}\big)$, $\alpha \in (0, 1)$.
Let $\lambda > 0$ and $\theta \in \mathbb{R}^m$ such that $F(\theta) > \inf F$. Then, via $\mathcal{H}1$, Lemma A.1(ii) and (1.7),

$$\begin{aligned}\|\nabla F(\theta)\|^2 &\geq \Big(\frac{\alpha}{(1+\alpha)L^{\frac{1}{\alpha}}}\Big)^\lambda \frac{\|\nabla F(\theta)\|^{2+\lambda\frac{1+\alpha}{\alpha}}}{(F(\theta) - \inf F)^\lambda} \\ &\geq \zeta^{-2-\lambda\frac{1+\alpha}{\alpha}} \Big(\frac{\alpha}{(1+\alpha)L^{\frac{1}{\alpha}}}\Big)^\lambda (F(\theta) - \inf F)^{2\beta + \lambda(\frac{1+\alpha}{\alpha}\beta - 1)}.\end{aligned}$$

Setting

$$\lambda = \frac{1-2\beta}{\frac{1+\alpha}{\alpha}\beta - 1} > 0 \quad \text{and} \quad \widetilde{\mu} = \zeta^{-2-\lambda\frac{1+\alpha}{\alpha}} \Big(\frac{\alpha}{(1+\alpha)L^{\frac{1}{\alpha}}} \Big)^{\lambda}$$

shows that

$$\widetilde{\mu}(F(\theta) - \inf F) \leq \|\nabla F(\theta)\|^2. \tag{3.48}$$

This relation remains true if $F(\theta) = \inf F$. The remaining lines of reasoning coincide with the next case by swapping μ with $\widetilde{\mu}$.

▸ *Case 2.* $\beta = \frac{1}{2}$.
By (3.47), (1.7) and (3.46),

$$\mathbb{E}[F(\theta_n) - \inf F] \leq (1 - \mu\gamma_n)\mathbb{E}[F(\theta_{n-1}) - \inf F] + C\gamma_n^{1+\alpha}.$$

Note that, since $\gamma_n \to 0$ as $n \to \infty$, there exists n_0 such that $\mu\gamma_n < 1$, $n \geq n_0$. Thus, using Lemma A.3,

$$\mathbb{E}[F(\theta_n) - \inf F] \leq \mathbb{E}[F(\theta_{n_0}) - \inf F] \prod_{k=n_0+1}^{n} (1 - \mu\gamma_k) + C \sum_{k=n_0+1}^{n} \gamma_k^{1+\alpha} \prod_{j=k+1}^{n} (1 - \mu\gamma_j),$$

with the convention $\prod_{\varnothing} = 1$. But then, utilizing that $1+x \leq \mathrm{e}^x$, $x \in \mathbb{R}$ and that $\sup_{k\geq 0} \mathbb{E}[F(\theta_k) - \inf F] < \infty$ by Lemma 2.1(ii),

$$\begin{aligned}\mathbb{E}[F(\theta_n) - \inf F] \leq \big(\sup_{k\geq 0} \mathbb{E}[F(\theta_k) - \inf F]\big) \exp\Big(\mu \sum_{k=1}^{n_0} \gamma_k\Big) \exp\Big(-\mu \sum_{k=1}^{n} \gamma_k\Big) \\ + C \sum_{k=1}^{n} \gamma_k^{1+\alpha} \exp\Big(-\mu \sum_{j=k+1}^{n} \gamma_j\Big),\end{aligned}$$

with the convention $\sum_{\varnothing} = 0$. Assuming $\mu\gamma_\star > \alpha$ in the case $\mathcal{H}6$(i)b, invoking Lemma A.4(i) yields

$$\mathbb{E}[F(\theta_n) - \inf F] \leq C\gamma_n^{\alpha}. \tag{3.49}$$

▸ *Case 3.* $\beta \in \big(\frac{1}{2}, 1\big)$.
Let

$$\mathfrak{S} \coloneqq \Big\{ \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big]^{\frac{1}{2}} \geq \Big(\sum_{k=1}^{n} \gamma_k \Big)^{-\frac{\beta}{2\beta-1}} \Big\}.$$

Observe that, by (1.7), $\mathbb{E}[(F(\theta_{n-1}) - \inf F)^{2\beta}] \leq \zeta^2 \mathbb{E}[\|\nabla F(\theta_{n-1})\|^2] < \infty$. Thus, by (1.7), (3.46) and Jensen's inequality,

$$\begin{aligned}\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big] &\geq \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big]^{\frac{2\beta-1}{2\beta}} \times \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big]^{\frac{1}{2\beta}} \times \mathbb{1}_{\mathfrak{S}} \\ &\geq \frac{1}{\sum_{k=1}^{n} \gamma_k} \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big]^{\frac{1}{2\beta}} \mathbb{1}_{\mathfrak{S}} \\ &= \frac{1}{\sum_{k=1}^{n} \gamma_k} \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big]^{\frac{1}{2\beta}} - \frac{1}{\sum_{k=1}^{n} \gamma_k} \mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\big]^{\frac{1}{2\beta}} \mathbb{1}_{\overline{\mathfrak{S}}} \\ &\geq \frac{\mu}{\sum_{k=1}^{n} \gamma_k} \mathbb{E}[(F(\theta_{n-1}) - \inf F)^{2\beta}]^{\frac{1}{2\beta}} - \frac{1}{\big(\sum_{k=1}^{n} \gamma_k\big)^{\frac{2\beta}{2\beta-1}}} \\ &\geq \frac{\mu}{\sum_{k=1}^{n} \gamma_k} \mathbb{E}[F(\theta_{n-1}) - \inf F] - \frac{1}{\big(\sum_{k=1}^{n} \gamma_k\big)^{\frac{2\beta}{2\beta-1}}}.\end{aligned}$$

Thus, using (3.47) and $\mathcal{H}6$(iii),

$$\mathbb{E}[F(\theta_n) - \inf F] \leq \Big(1 - \frac{\mu\gamma_n}{\sum_{k=1}^n \gamma_k}\Big)\mathbb{E}[F(\theta_{n-1}) - \inf F] + \frac{\gamma_n}{\big(\sum_{k=1}^n \gamma_k\big)^{\frac{2\beta}{2\beta-1}}} + \frac{C\gamma_n}{\big(\sum_{k=1}^n \gamma_k\big)^{\alpha\rho}}.$$

Because $\frac{\gamma_n}{\sum_{k=1}^n \gamma_k} \to 0$ as $n \to \infty$, there exists n_0 such that $\frac{\mu\gamma_n}{\sum_{k=1}^n \gamma_k} < 1$, $n \geq n_0$. But then, by Lemma A.3,

$$\begin{aligned}\mathbb{E}[F(\theta_n) - \inf F] &\leq \mathbb{E}[F(\theta_{n_0}) - \inf F] \prod_{k=n_0+1}^n \Big(1 - \frac{\mu\gamma_k}{\sum_{j=1}^k \gamma_j}\Big) \\ &+ \sum_{k=n_0+1}^n \frac{\gamma_k}{\big(\sum_{j=1}^k \gamma_j\big)^{\frac{2\beta}{2\beta-1}}} \prod_{j=k+1}^n \Big(1 - \frac{\mu\gamma_j}{\sum_{i=1}^j \gamma_i}\Big) + \sum_{k=n_0+1}^n \frac{C\gamma_k}{\big(\sum_{j=1}^k \gamma_j\big)^{\alpha\rho}} \prod_{j=k+1}^n \Big(1 - \frac{\mu\gamma_j}{\sum_{i=1}^j \gamma_i}\Big),\end{aligned}$$

with the convention $\prod_\varnothing = 1$. Then, using that $1 + x \leq \mathrm{e}^x$, $x \in \mathbb{R}$ and that $\sup_{k\geq 0} \mathbb{E}[F(\theta_k) - \inf F] < \infty$ by Lemma 2.1(ii),

$$\begin{aligned}\mathbb{E}[F(\theta_n) - \inf F] \leq \big(\sup_{k\geq 0} \mathbb{E}[F(\theta_k) - \inf F]\big) &\exp\Big(\mu \sum_{k=1}^{n_0} \frac{\gamma_k}{\sum_{j=1}^k \gamma_j}\Big) \exp\Big(-\mu \sum_{k=1}^n \frac{\gamma_k}{\sum_{j=1}^k \gamma_j}\Big) \\ &+ \sum_{k=1}^n \frac{\gamma_k}{\big(\sum_{j=1}^k \gamma_j\big)^{\frac{2\beta}{2\beta-1}}} \exp\Big(-\mu \sum_{j=k+1}^n \frac{\gamma_j}{\sum_{i=1}^j \gamma_i}\Big) \\ &+ C \sum_{k=1}^n \frac{\gamma_k}{\big(\sum_{j=1}^k \gamma_j\big)^{\alpha\rho}} \exp\Big(-\mu \sum_{j=k+1}^n \frac{\gamma_j}{\sum_{i=1}^j \gamma_i}\Big),\end{aligned}$$

with the convention $\sum_\varnothing = 0$. Via Lemma A.4(ii), we deduce

$$\mathbb{E}[F(\theta_n) - \inf F] \leq \frac{C}{\big(\sum_{k=1}^n \gamma_k\big)^r}, \tag{3.50}$$

where $r = \frac{1}{2\beta-1} \wedge (\alpha\rho - 1)$.

The subsequent high probability bounds are obtained using Markov's inequality. □

Proof of Corollary 1.21. This is a direct consequence of Theorems 1.14 and 1.17. □

4 Global Convergence Proofs

Henceforward, for a real-valued sequence $(v_n)_{n\geq 0}$, we denote its discrete gradient

$$\Delta v_n \coloneqq v_n - v_{n-1}, \quad n \geq 1.$$

The next lemma is a short interlude providing a useful property.

Lemma 4.1. *Recalling the definition* (3.5)*, under $\mathcal{H}1$–3, the martingale $\big(\sum_{k=1}^n \gamma_k \Delta_k\big)_{n\geq 1}$ converges in $L^2(\mathbb{P})$ and $\mathbb{P}$-as.*

Proof. Considering (3.8), define the $L^2(\mathbb{P})$ $(\mathcal{F}_n)_{n\geq 0}$-martingale

$$M'_n \coloneqq \sum_{k=1}^n \gamma_k \Delta_k, \quad n \geq 1.$$

Since $(\Delta_k)_{k\geq 1}$ are martingale increments, using (3.8) and that $\sum_{k=1}^{\infty}\gamma_k^2<\infty$ by $\mathcal{H}3$,

$$\sup_{n\geq 1}\mathbb{E}[|M'_n|^2]=\sum_{k=1}^{\infty}\gamma_k^2\mathbb{E}[\Delta_k^2]\leq C\sum_{k=1}^{\infty}\gamma_k^2<\infty.$$

$(M'_n)_{n\geq 1}$ is therefore uniformly bounded in $L^2(\mathbb{P})$ and hence converges in $L^2(\mathbb{P})$ and $\mathbb{P}$-as. □

Proof of Theorem 1.22. ▸ *Step 1. Final hitting time.*
Define

$$\tau'\coloneqq\inf\Big\{n\geq 0\colon \forall k\geq n, |F(\theta_k)-F_\star|+\Big|\sum_{j=k+1}^{\infty}\gamma_j\Delta_j\Big|+\frac{L}{1+\alpha}\sum_{j=k+1}^{\infty}\gamma_j^{1+\alpha}\|g_j(\theta_{j-1})\|^{1+\alpha}\leq 1\Big\}. \tag{4.1}$$

By Lemmas 2.1(i,iii) and 4.1,

$$\lim_{k\to\infty}|F(\theta_k)-F_\star|+\Big|\sum_{j=k+1}^{\infty}\gamma_j\Delta_j\Big|+\frac{L}{1+\alpha}\sum_{j=k+1}^{\infty}\gamma_j^{1+\alpha}\|g_j(\theta_{j-1})\|^{1+\alpha}=0\quad\mathbb{P}\text{-as}, \tag{4.2}$$

thus $\tau'<\infty$ $\mathbb{P}$-as.

We reconsider the framework of Theorem 1.17(i)b and, recalling (3.12) and (3.14), we define

$$\widetilde{\tau}\coloneqq\tau_1\vee\tau'. \tag{4.3}$$

Let $\delta\in(0,1)$. Since $\widetilde{\tau}<\infty$ $\mathbb{P}$-as on $\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}$, there exists $n_\delta\geq 0$ such that

$$\mathbb{P}(\widetilde{\tau}<n_\delta,\mathcal{A}\subset\mathcal{L}_{\ell_\delta})\geq 1-\delta. \tag{4.4}$$

▸ *Step 2. Lyapunov sequence.*
Define

$$\mathbb{C}_\delta\coloneqq\{\widetilde{\tau}<n_\delta\}\cap\{\mathcal{A}\subset\mathcal{L}_{\ell_\delta}\}.$$

Let $n>n_\delta$. We hereby define a Lyapunov sequence $(G_n)_{n\geq n_\delta}$ associated to $(F(\theta_n))_{n\geq n_\delta}$ and provide some subsequent properties.

By (2.1), (4.3), (3.14), (3.1) and (3.5),

$$\begin{aligned}&(F(\theta_n)-F_\star)\mathbb{1}_{\mathbb{C}_\delta}\\&\quad\leq\Big(F(\theta_{n-1})-F_\star-\gamma_n\|\nabla F(\theta_{n-1})\|^2+\gamma_n\Delta_n+\frac{L}{1+\alpha}\gamma_n^{1+\alpha}\|g_n(\theta_{n-1})\|^{1+\alpha}\Big)\mathbb{1}_{\mathbb{C}_\delta}\end{aligned} \tag{4.5}$$

Let

$$G_n\coloneqq\mathbb{E}\Big[\Big(F(\theta_n)-F_\star+\sum_{k=n+1}^{\infty}\gamma_k\Delta_k+\frac{L}{1+\alpha}\sum_{k=n+1}^{\infty}\gamma_k^{1+\alpha}\|g_k(\theta_{k-1})\|^{1+\alpha}\Big)\mathbb{1}_{\mathbb{C}_\delta}\Big],\quad n\geq n_\delta, \tag{4.6}$$

which is well defined according to Lemmas 2.1(i,iii) and 4.1. Then, by (4.5),

$$\gamma_n\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\mathbb{1}_{\mathbb{C}_\delta}\big]\leq-\Delta G_n. \tag{4.7}$$

Hence $(G_n)_{n\geq n_\delta}$ is nonincreasing. Via (4.1) and (4.2), by dominated convergence, $G_n\to 0$ as $n\to\infty$. Therefore $G_n\geq 0$, $n\geq n_\delta$.

Note that, since $\widetilde{\tau}\geq\tau_1$ $\mathbb{P}$-as, it follows from (4.4) that $\mathbb{P}(\tau_1<n_\delta,\mathcal{A}\subset\mathcal{L}_{\ell_\delta})\geq 1-\delta$, so that, via (3.33) and (3.43),

$$\mathbb{E}[(F(\theta_{n-1})-F_\star)^+\mathbb{1}_{\mathbb{C}_\delta}]\leq\mathbb{E}[(F(\theta_{n-1})-F_\star)^+\mathbb{1}_{\mathbb{A}_\delta}]\leq C_\delta u_n, \tag{4.8}$$

where

$$u_n \coloneqq \begin{cases} n^{-s(\alpha\wedge\frac{1}{2})}, & \text{if } \beta_\delta \in (0,\frac{1}{2}] \text{ and } s \in (\frac{1}{1+\alpha},1], \text{ with } \gamma_\star > \frac{1}{2}\vee\alpha \text{ if } s=1,\\ n^{-(1-s)r_\delta}, & \text{if } \beta_\delta \in (\frac{1}{2},1) \text{ and } s \in [\frac{\rho}{1+\rho},1),\\ (\ln n)^{-r_\delta}, & \text{if } \beta_\delta \in (\frac{1}{2},1) \text{ and } s=1, \end{cases} \qquad n \geq n_\delta, \quad (4.9)$$

with $\rho > \frac{1}{\alpha}$ and r_δ defined in (3.42).

By monotone convergence, Jensen's inquality, Doob's maximal inequality (or the Burkholder-Davis-Gundy inequality), (3.8) and a series-integral comparison,

$$\begin{aligned} \mathbb{E}\Big[\Big|\sum_{k=n}^{\infty}\gamma_k\Delta_k\Big|\Big] &\leq \mathbb{E}\Big[\sup_{p\geq n}\Big|\sum_{k=n}^{p}\gamma_k\Delta_k\Big|\Big] \\ &= \lim_{N\to\infty}\mathbb{E}\Big[\sup_{n\leq p\leq N}\Big|\sum_{k=n}^{p}\gamma_k\Delta_k\Big|^2\Big]^{\frac{1}{2}} \leq \Big(\sum_{k=n}^{\infty}\gamma_k^2\mathbb{E}[\Delta_k^2]\Big)^{\frac{1}{2}} \leq \frac{C}{n^{\frac{2s-1}{2}}}. \end{aligned} \quad (4.10)$$

Moreover, by Fubini's theorem, (3.7) and a series-integral comparison,

$$\mathbb{E}\Big[\sum_{k=n}^{\infty}\gamma_k^{1+\alpha}\|g_k(\theta_{k-1})\|^{1+\alpha}\Big] \leq \frac{C}{n^{(1+\alpha)s-1}}. \quad (4.11)$$

Thus, gathering (4.8)–(4.11) and recalling (4.6),

$$G_{n-1} \leq C_\delta\Big(u_n + \frac{1}{n^{\frac{2s-1}{2}}} + \frac{1}{n^{(1+\alpha)s-1}}\Big). \quad (4.12)$$

▸ *Step 3. Summability of* $(\gamma_n\|\nabla F(\theta_{n-1})\|)_{n\geq 1}$.
Let $s \in \big[\frac{\rho}{1+\rho},1\big]$, $\rho > \frac{1}{\alpha}$, and $\sigma \in (0,1)$. Assume $\gamma_\star > \frac{1}{2}\vee\alpha$ if $s=1$.

On the one hand, note that if $\mathbb{E}[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}] > 0$, necessarily, $G_{n-1} > 0$. Otherwise, since $0 \leq G_n \leq G_{n-1}$, $G_n = 0$. But then, $\mathbb{E}[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}] \leq \sqrt{\gamma_n^{-1}(-\Delta G_n)} = 0$, which is absurd. Thus, using Jensen's inequality, (4.7) and Lemma A.6, if $\mathbb{E}[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}] > 0$,

$$\begin{aligned} &\gamma_n\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}\big]\mathbb{1}_{\mathbb{E}[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}]>G_{n-1}^\sigma} \\ &\qquad \leq \gamma_n G_{n-1}^{-\sigma}\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\mathbb{1}_{\mathbb{C}_\delta}\big] \leq G_{n-1}^{-\sigma}(-\Delta G_n) \leq \frac{-\Delta G_n^{1-\sigma}}{1-\sigma}, \end{aligned} \quad (4.13)$$

recalling that $-\Delta G_n^{1-\sigma} = G_{n-1}^{1-\sigma} - G_n^{1-\sigma}$. This holds also if $\mathbb{E}[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}] = 0$.

On the other hand, using (4.12) and (4.9),

$$\begin{aligned} &\gamma_n\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}\big]\mathbb{1}_{\mathbb{E}[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}]\leq G_{n-1}^\sigma} \\ &\qquad \leq \gamma_n G_{n-1}^\sigma \leq C_\delta\Big(\frac{u_n^\sigma}{n^s} + \frac{1}{n^{s+\frac{2s-1}{2}\sigma}} + \frac{1}{n^{s+((1+\alpha)s-1)\sigma}}\Big). \end{aligned} \quad (4.14)$$

All in all, gathering the inequalities (4.13) and (4.14),

$$\begin{aligned} &\gamma_n\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}\big] \\ &\leq \begin{cases} \frac{C_\delta}{1-\sigma}\big(n^{-s-s(\alpha\wedge\frac{1}{2})\sigma} + n^{-s-\frac{2s-1}{2}\sigma} + n^{-s-((1+\alpha)s-1)\sigma} - \Delta G_n^{1-\sigma}\big) & \text{if } \beta_\delta \in (0,\frac{1}{2}],\\ \frac{C_\delta}{1-\sigma}\big(n^{-s-(1-s)r_\delta\sigma} + n^{-s-\frac{2s-1}{2}\sigma} + n^{-s-((1+\alpha)s-1)\sigma} - \Delta G_n^{1-\sigma}\big) & \text{if } \beta_\delta \in (\frac{1}{2},1) \text{ and } s<1,\\ \frac{C_\delta}{1-\sigma}\big(n^{-1}(\ln n)^{-r_\delta\sigma} + n^{-1-\frac{\sigma}{2}} + n^{-1-\alpha\sigma} - \Delta G_n^{1-\sigma}\big) & \text{if } \beta_\delta \in (\frac{1}{2},1) \text{ and } s=1. \end{cases} \end{aligned} \quad (4.15)$$

Setting $\rho > 3 \vee \frac{2}{\alpha}$ so that $r_\delta > 1$ if $\beta_\delta \in \big(\frac{1}{2}, 1\big)$ and $s \in \big(\frac{2}{2+\alpha} \vee \frac{3}{4}, 1\big]$, taking

$$\sigma \in \begin{cases} \big(\frac{1-s}{s(\alpha\wedge\frac{1}{2})} \vee \frac{2(1-s)}{2s-1} \vee \frac{1-s}{(1+\alpha)s-1}, 1\big) & \text{if } \beta_\delta \in (0, \frac{1}{2}], \\ \big(\frac{1}{r_\delta} \vee \frac{2(1-s)}{2s-1} \vee \frac{1-s}{(1+\alpha)s-1}, 1\big) & \text{if } \beta_\delta \in (\frac{1}{2}, 1) \text{ and } s < 1, \\ \big(\frac{1}{r_\delta}, 1\big) & \text{if } \beta_\delta \in (\frac{1}{2}, 1) \text{ and } s = 1, \end{cases}$$

and recalling that $G_n \to 0$ as $n \to \infty$, we deduce that $\sum_{n=n_\delta}^\infty \gamma_n \mathbb{E}[\|\nabla F(\theta_{n-1})\| \mathbb{1}_{\mathbb{C}_\delta}] < \infty$.

▸ *Step 4. Conclusion.*
By Fubini's theorem, $\sum_{n=1}^\infty \gamma_n \|\nabla F(\theta_{n-1})\| < \infty$ $\mathbb{P}$-as on $\mathbb{C}_\delta$, thus on $\cup_{p\geq 0} \mathbb{C}_{\delta/2^p}$, which is an event of probability $\mathbb{P}(\cup_{p\geq 0} \mathbb{C}_{\delta/2^p}) \geq \sup_{p\geq 0} \big(1 - \frac{\delta}{2^p}\big) = 1$. Therefore $\sum_{n=1}^\infty \gamma_n \|\nabla F(\theta_{n-1})\| < \infty$ $\mathbb{P}$-as and $(\gamma_n \nabla F(\theta_{n-1}))_{n\geq 1}$ is $\mathbb{P}$-as-absolutely summable. Finally, note that

$$\theta_n = \theta_0 + \sum_{k=1}^n \gamma_k \big(\nabla F(\theta_{k-1}) - g_k(\theta_{k-1})\big) - \sum_{k=1}^n \gamma_k \nabla F(\theta_{k-1}), \quad n \geq 1. \tag{4.16}$$

By Lemma 2.1(iv) and the previous finding, $(\theta_n)_{n\geq 0}$ converges $\mathbb{P}$-as toward a $\mathbb{P}$-as finite random variable $\theta_\star$. Besides, by the continuity of F and ∇F via $\mathcal{H}1$, $\nabla F(\theta_\star) = \lim_{n\to\infty} \|\nabla F(\theta_n)\| = 0$ $\mathbb{P}$-as via Lemma 2.3 and $F(\theta_\star) = \lim_{n\to\infty} F(\theta_n) = F_\star$ $\mathbb{P}$-as via Lemma 2.1(i). □

The telescopic series decomposition (4.17) suggests that the convergence speed of $(\theta_n)_{n\geq 1}$ results from a dual contribution from the martingale residual $\big(\sum_{k=n+1}^\infty \gamma_k (g_k(\theta_{k-1}) - \nabla F(\theta_k))\big)_{n\geq 1}$ and the drift residual $\big(\sum_{k=n+1}^\infty \gamma_k \nabla F(\theta_{k-1})\big)_{n\geq 1}$. A classical interplay between the Jensen and Burkholder-Davis-Gundy inequalities shows that the martingale residual behaves like $\sqrt{\sum_{k=n+1}^\infty \gamma_k^2}$ $= \mathrm{O}(n^{-\frac{2s-1}{2}})$. As for the drift residual, we could rely on the bound (4.15) that has been obtained constructively in Theorem 1.22's proof. The ensuing bound requires however to strike a balance between terms with exponent σ and terms with exponent $1-\sigma$, where $\sigma \in (0,1)$, in order to reach optimality. To avoid such tradeoff, we seek alternative bounds on the trend residual in Corollary 1.24's proof.

Proof of Corollary 1.24. **(i)a.** ▸ *Step 1. Preliminaries.*
We reuse notation from the preceding proof, bearing in mind that we are in the framework of Theorem 1.17(i)a. Via the $\mathbb{P}$-as-absolute summability of $(\gamma_n \nabla F(\theta_{n-1}))_{n\geq 1}$ and Lemma 2.1(iv), by telescopic summation,

$$\theta_n - \theta_\star = \sum_{k=n+1}^\infty \gamma_k \big(\nabla F(\theta_{k-1}) - g_k(\theta_{k-1})\big) + \sum_{k=n+1}^\infty \gamma_k \nabla F(\theta_{k-1}). \tag{4.17}$$

Hence

$$\mathbb{E}\big[\|\theta_n - \theta_\star\| \mathbb{1}_{\mathbb{C}_\delta}\big] \leq \mathbb{E}\Big[\Big\| \sum_{k=n+1}^\infty \gamma_k \big(\nabla F(\theta_{k-1}) - g_k(\theta_{k-1})\big)\Big\|\Big] + \sum_{k=n+1}^\infty \gamma_k \mathbb{E}\big[\|\nabla F(\theta_{k-1})\| \mathbb{1}_{\mathbb{C}_\delta}\big]. \tag{4.18}$$

Note that, by monotone convergence, Jensen's inquality, Doob's maximal inequality (or the Burkholder-Davis-Gundy inequality) and (2.6),

$$\begin{aligned} \mathbb{E}\Big[\Big\| \sum_{k=n+1}^\infty \gamma_k \big(\nabla F(\theta_{k-1}) - g_k(\theta_{k-1})\big)\Big\|\Big] &\leq \mathbb{E}\Big[\sup_{p\geq n+1} \Big\| \sum_{k=n+1}^p \gamma_k \big(\nabla F(\theta_{k-1}) - g_k(\theta_{k-1})\big)\Big\|\Big] \\ &= \lim_{N\to\infty} \mathbb{E}\Big[\sup_{n+1\leq p\leq N} \Big\| \sum_{k=n+1}^p \gamma_k \big(\nabla F(\theta_{k-1}) - g_k(\theta_{k-1})\big)\Big\|^2\Big]^{\frac{1}{2}} \leq C\Big(\sum_{k=n}^\infty \gamma_k^2\Big)^{\frac{1}{2}} \leq \frac{C}{n^{\frac{2s-1}{2}}}. \end{aligned} \tag{4.19}$$

▸ *Step 2. Alternative treatment for* $\big(\sum_{k=n+1}^{\infty}\gamma_k\|\nabla F(\theta_{k-1})\|\big)_{n\geq 1}$.
Let $s\in\big[\frac{\rho}{1+\rho},1\big]$, $\rho>\frac{1}{\alpha}$, and $\sigma\in(0,1)$. Assume

$$\gamma_\star > \begin{cases}(1\vee\zeta_\delta^{1/\beta_\delta})(\frac{1}{2}\vee\alpha) & \text{if } s=1 \text{ and } \beta_\delta\in(0,\frac{1}{2}],\\ \frac{1}{2}\vee\alpha & \text{if } s=1 \text{ and } \beta_\delta\in(\frac{1}{2},1),\end{cases}$$

Define

$$v_n \coloneqq \frac{1}{n^{\frac{2s-1}{2}}}+\frac{1}{n^{(1+\alpha)s-1}},\quad n\geq 1.$$

Let $n>n_\delta$ and $\sigma\in(0,1)$. From (4.12), one gets

$$0<(G_{n-1}+u_{n-1}+v_{n-1})^\sigma \leq C\sqrt{\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|^2\mathbb{1}_{\mathbb{C}_\delta}\big]+u_n^{2\sigma}+v_n^{2\sigma}}.$$

Thus, by Jensen's inequality,

$$\gamma_n\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}\big]\leq C\frac{\gamma_n(\mathbb{E}[\|\nabla F(\theta_{n-1})\|^2\mathbb{1}_{\mathbb{C}_\delta}]+u_n^{2\sigma}+v_n^{2\sigma})}{(G_{n-1}+u_{n-1}+v_{n-1})^\sigma}. \tag{4.20}$$

Observe that, by a first order Taylor-Lagrange expansion, for $\varrho>0$,

$$-\Delta\Big(\frac{1}{n^\varrho}\Big)\sim\frac{\varrho}{n^{1+\varrho}}\quad\text{and}\quad -\Delta\Big(\frac{1}{(\ln n)^\varrho}\Big)\sim\frac{\varrho}{n(\ln n)^{1+\varrho}}\quad\text{as}\quad n\to\infty. \tag{4.21}$$

Hence, taking $\rho>3\vee\frac{2}{\alpha}$ so that $r_\delta>1$ if $\beta_\delta\in\big(\frac{1}{2},1\big)$ and $s\in\big(\frac{3}{4}\vee\frac{2}{2+\alpha},1\big]$, and setting

$$\sigma\in\begin{cases}\big[\frac{1}{2(2s-1)}\vee\frac{\alpha s}{2((1+\alpha)s-1)}\vee\frac{1-((1-\alpha)\vee\frac{1}{2})s}{2(\alpha\wedge\frac{1}{2})s},1\big) & \text{if } \beta_\delta\in(0,\frac{1}{2}],\\ \big[\frac{1}{2(2s-1)}\vee\frac{\alpha s}{2((1+\alpha)s-1)}\vee\frac{1+r_\delta}{2r_\delta},1\big) & \text{if } \beta_\delta\in(\frac{1}{2},1) \text{ and } s<1,\\ [\frac{1+r_\delta}{2r_\delta},1) & \text{if } \beta_\delta\in(\frac{1}{2},1) \text{ and } s=1,\end{cases}$$

one obtains, via (4.20), (4.7), (4.21) and Lemma A.6,

$$\gamma_n\mathbb{E}\big[\|\nabla F(\theta_{n-1})\|\mathbb{1}_{\mathbb{C}_\delta}\big]\leq C_\delta\frac{-\Delta(G_n+u_n+v_n)}{(G_{n-1}+u_{n-1}+v_{n-1})^\sigma}\leq-\frac{C_\delta}{1-\sigma}\Delta(G_n+u_n+v_n)^{1-\sigma}.$$

Thus, using that $G_n+u_n+v_n\to 0$ as $n\to\infty$,

$$\sum_{k=n+1}^{\infty}\gamma_k\mathbb{E}\big[\|\nabla F(\theta_{k-1})\|\mathbb{1}_{\mathbb{C}_\delta}\big]\leq\frac{C_\delta}{1-\sigma}(u_n^{1-\sigma}+v_n^{1-\sigma}). \tag{4.22}$$

▸ *Step 3. Conclusion.*
All in all, gathering (4.18), (4.19) and (4.22) and applying similar reasoning to Steps 1.2 and 2.2 of the proof of Theorem 1.17(i)a yield the sought result.

(i)b, (ii) The proofs are identical to the anteceding one, by applying suitable substitutions. □

Proof of Corollary 1.26. This is a straightforward consequence of Corollary 1.24. □

Conclusion

For α-Hölder-differentiable loss landscapes satisfying a local Łojasiewicz condition with exponent $\beta \in (0,1)$, our analysis shows that, for a learning schedule $(\gamma_n)_{n\geq 1}$ such that $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\lim_{n\to\infty} \gamma_n = 0$, (SGD) converges in high probability in $\mathrm{O}(\gamma_n^{\alpha\wedge\frac{1}{2}})$ if $\beta \in \big(0, \frac{1}{2}\big]$ and in $\mathrm{O}\big(\big(\sum_{k=1}^n \gamma_k\big)^{-\frac{1}{2\beta-1}\wedge(\alpha\rho-1)\wedge\frac{\rho-1}{2}}\big)$ if $\beta \in \big(\frac{1}{2}, 1\big)$, for some adjustable parameter $\rho > \frac{1}{\alpha}$ such that $\gamma_n = \mathrm{O}\big(\big(\sum_{k=1}^n \gamma_k\big)^{-\rho}\big)$. These convergence rates are free from the requirement that the (SGD) trajectories are initialized or remain close to their target. Furthermore, they help uncover optimal iteration amounts to employ in order to help cut down significantly the fine-tuning phase.

Future lines of research could involve extending our results to confined (SGD) trajectories around local minima, biased (SGD) or locally Lipschitz loss landscapes presenting singularities. An outlook on eliciting the dependency of the Łojasiewicz exponent upon a neural network's architecture could also be envisaged.

Acknowledgements. The author would like to thank Guillaume Garrigos, Stéphane Crépey and Noufel Frikha for discussions on SGD and the Łojasiewicz property.

A Ancillary Results

The next result is given in [Lei+20, Lemma 1] in a Hilbert space setting, but we provide a different take on its second part. The first point is often referred to as the descent lemma.

Lemma A.1. *Let $\varphi\colon \mathbb{R}^m \to \mathbb{R}$ be a differentiable function such that its derivative $\nabla\varphi$ is (L,α)-Hölder continuous, with $L > 0$ and $\alpha \in (0,1]$. Then*

(i) *one has*

$$\big|\varphi(\theta') - \varphi(\theta) - \nabla\varphi(\theta)^\top(\theta' - \theta)\big| \leq \frac{L}{1+\alpha}\|\theta' - \theta\|^{1+\alpha}, \quad \theta, \theta' \in \mathbb{R}^m.$$

(ii) *if, additionally, φ is lower bounded,*

$$\|\nabla\varphi(\theta)\|^{\frac{1+\alpha}{\alpha}} \leq \frac{(1+\alpha)L^{\frac{1}{\alpha}}}{\alpha}\big(\varphi(\theta) - \inf\varphi\big), \quad \theta \in \mathbb{R}^m.$$

Proof. **(i)** For $\theta, \theta' \in \mathbb{R}^m$, start with the first order Taylor-Lagrange expansion

$$\varphi(\theta') = \varphi(\theta) + \nabla\varphi(\theta)^\top(\theta' - \theta) + \int_0^1 (\theta' - \theta)^\top\big(\nabla\varphi(\theta + t(\theta' - \theta)) - \nabla\varphi(\theta)\big)\mathrm{d}t,$$

then note that, since $\nabla\varphi$ is (L,α)-Hölder,

$$\left|\int_0^1 (\theta' - \theta)^\top\big(\nabla\varphi(\theta + t(\theta' - \theta)) - \nabla\varphi(\theta)\big)\mathrm{d}t\right| \leq \frac{L}{1+\alpha}\|\theta' - \theta\|^{1+\alpha}.$$

(ii) Let $\theta \in \mathbb{R}^m$. Applying the previous result, for $t \geq 0$,

$$0 \leq \varphi(\theta - t\nabla\varphi(\theta)) - \inf\varphi \leq \varphi(\theta) - \inf\varphi - t\|\nabla\varphi(\theta)\|^2 + \frac{Lt^{1+\alpha}}{1+\alpha}\|\nabla\varphi(\theta)\|^{1+\alpha},$$

so that

$$t\|\nabla\varphi(\theta)\|^2 - \frac{Lt^{1+\alpha}}{1+\alpha}\|\nabla\varphi(\theta)\|^{1+\alpha} \leq \varphi(\theta) - \inf\varphi.$$

Maximizing the left hand side with respect to t leads to the optimal choice

$$t_\star = \frac{1}{L^{\frac{1}{\alpha}}} \|\nabla\varphi(\theta)\|^{\frac{1-\alpha}{\alpha}}.$$

□

The next technical result is invoked but not proven in [Lei+20, Theorem 2(c)].

Lemma A.2. *Let $(u_n)_{n\geq 0}$ be a real sequence and $\ell^-, \ell^+ \in \mathbb{R}$ such that*

$$-\infty \leq \liminf_{n\to\infty} u_n < \ell^- < \ell^+ < \limsup_{n\to\infty} u_n \leq \infty.$$

Then, there exist extractions $\chi, \psi \colon \mathbb{N} \to \mathbb{N}$ such that, for all $n \geq 0$,

$$u_{\chi(n)} < \ell^-, \quad u_{\psi(n)} > \ell^+ \quad \text{and} \quad \ell^- \leq u_k \leq \ell^+, \quad \chi(n) < k < \psi(n),$$

where, potentially, $\psi(n) = \chi(n) + 1$.

Proof. Since $\liminf_{n\to\infty} u_n < \ell^-$, then $u_n < \ell^-$ infinitely often. Let $\mathcal{X} \coloneqq \{n \colon u_n < \ell^-\}$. Thus $\#\mathcal{X} = \infty$, where $\#$ denotes cardinality.

Define $n' = \inf\{p > n \colon u_p > \ell^+\}$, $n \in \mathcal{X}$, with the convention $\inf \varnothing = \infty$. Since $\limsup_{n\to\infty} u_n > \ell^+$, $u_n > \ell^+$ infinitely often so that $n' < \infty$, $n \in \mathcal{X}$. Let $\mathcal{Y} = \{n' \colon n \in \mathcal{X}\}$. Note that $\#\mathcal{Y} = \infty$.

Next, let $n'' \coloneqq \sup\{p < n' \colon u_n < \ell^-\}$, $n \in \mathcal{X}$, with the convention $\sup \varnothing = -\infty$. Observe that, for $n \in \mathcal{X}$, it holds that $n'' \in \mathcal{X}$, that $n'' \geq n > -\infty$ and that $(n'')' = n'$. Let $\mathcal{Z} \coloneqq \{n'' \colon n \in \mathcal{X}\}$. Then $\#\mathcal{Z} = \infty$. Moreover, for $n \in \mathcal{X}$, either $n' = n'' + 1$ or $\ell^- \leq u_k \leq \ell^+$, $n'' < k < n'$. Otherwise, there exists $n \in \mathcal{X}$ such that $n' > n'' + 1$ and there exists $n'' < k < n'$ such that, either $u_k < \ell^-$, or $u_k > \ell^+$. The former possibility contradicts the definition of n'', and the latter contradicts that $(n'')' = n'$.

Finally, since $\mathcal{Z}$ and $\mathcal{Y}$ are ordered, it suffices to take χ and ψ the ordered enumerations of $\mathcal{Z}$ and $\mathcal{Y}$ respectively. □

The following result is a classic in stochastic approximation literature.

Lemma A.3. *Let $(u_n)_{n\geq 0}$ and $(v_n)_{n\geq 0}$ be real-valued sequences and $(\gamma_n)_{n\geq 0}$ a nonnegative sequence such that*

$$u_n \leq (1 - \gamma_n)u_{n-1} + v_n, \quad n \geq 0.$$

If $\gamma_n < 1$ beginning from a certain rank $p \geq 1$, then

$$u_n \leq u_p \prod_{k=p+1}^{n} (1 - \gamma_k) + \sum_{k=p+1}^{n} v_k \prod_{j=k+1}^{n} (1 - \gamma_j),$$

with the convention $\prod_{\varnothing} = 1$.

Proof. The assertion follows by recursion. □

The first part of the below lemma is a special case of [For15, Lemmas 5.8 & 5.9]. Due to a minor mistake in the proof therein and for the sake of completeness, it is reproven below. The second part of the lemma is novel.

Lemma A.4. *Let $\mu > 0$, $\alpha \geq 0$ and $(\gamma_n)_{n\geq 1}$ a positive sequence such that*

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \gamma_n = 0. \tag{A.1}$$

(i) *Assume that either of the following is satisfied:*

C1. $\ln\left(\frac{\gamma_{n-1}}{\gamma_n}\right) = \mathrm{o}(\gamma_n)$;

C2. *there exists* $\gamma_\star > \frac{\alpha}{\mu}$ *such that* $\ln\left(\frac{\gamma_{n-1}}{\gamma_n}\right) \sim \frac{\gamma_n}{\gamma_\star}$.

Then, using the convention $\sum_\varnothing = 0$,

a. $\limsup_{n\to\infty} \gamma_n^{-\alpha} \sum_{k=1}^n \gamma_k^{1+\alpha} \exp\big(-\mu \sum_{j=k+1}^n \gamma_j\big) < \infty$;

b. $\limsup_{n\to\infty} \gamma_n^{-\alpha} \exp\big(-\mu \sum_{k=1}^n \gamma_k\big) = 0$.

(ii) *Let* $\beta \geq 0$. *Suppose that, in addition,*

$$\sum_{n=1}^{\infty} \frac{\gamma_n}{\sum_{k=1}^n \gamma_k} = \infty.$$

Then, with the convention that $\sum_\varnothing = 0$,

a. $\limsup_{n\to\infty} \big(\sum_{k=1}^n \gamma_k\big)^{-\beta} \sum_{k=1}^n \frac{\gamma_k}{(\sum_{j=1}^k \gamma_j)^{1+\beta}} \exp\Big(-\mu \sum_{j=k+1}^n \frac{\gamma_j}{\sum_{i=1}^j \gamma_i}\Big) < \infty$;

b. $\limsup_{n\to\infty} \big(\sum_{k=1}^n \gamma_k\big)^{-\beta} \exp\Big(-\mu \sum_{k=1}^n \frac{\gamma_k}{\sum_{j=1}^k \gamma_j}\Big) = 0$.

Proof. **(i)a.** Define

$$a_n = \gamma_n^{-\alpha} \sum_{k=1}^n \gamma_k^{1+\alpha} \exp\Big(-\mu \sum_{j=k+1}^n \gamma_j\Big), \quad n \geq 1.$$

Then, for $n \geq 2$,

$$a_n = \Big(\frac{\gamma_{n-1}}{\gamma_n}\Big)^\alpha \exp(-\mu\gamma_n) a_{n-1} + \gamma_n = (1-b_n)a_{n-1} + b_n b_n^{-1}\gamma_n,$$

where

$$b_n = 1 - \exp\Big(-\mu\gamma_n + \alpha \ln \frac{\gamma_{n-1}}{\gamma_n}\Big) \sim \begin{cases} \mu\gamma_n, & \text{case C1}, \\ (\mu - \frac{\alpha}{\gamma_\star})\gamma_n, & \text{case C2}, \end{cases} \tag{A.2}$$

$$b_n^{-1}\gamma_n \sim c_\star := \begin{cases} \frac{1}{\mu}, & \text{case C1}, \\ \frac{1}{\mu - \alpha/\gamma_\star}, & \text{case C2}. \end{cases}$$

Let $\varepsilon > 0$ and $c = c_\star + \varepsilon$. Then, there exists $n_0 \geq 0$ such that $b_n^{-1}\gamma_n \leq c$, $n \geq n_0$. Via the convexity of $x \in \mathbb{R} \mapsto x^+$ and Jensen's inequality, for $n \geq n_0$,

$$(a_n - c)^+ \leq (1-b_n)(a_{n-1} - c)^+ + b_n(b_n^{-1}\gamma_n - c)^+ = (1-b_n)(a_{n-1}-c)^+.$$

Considering (A.2), one has $\sum_{n=1}^\infty b_n = \infty$, so that $\lim_{n\to\infty}(a_n - c)^+ = 0$. Hence $\limsup_{n\to\infty} a_n \leq c = c_\star + \varepsilon$ for any $\varepsilon > 0$, thus $\limsup_{n\to\infty} a_n \leq c_\star < \infty$.

(i)b. Let

$$a_n = \gamma_n^{-\alpha} \exp\Big(-\mu \sum_{k=1}^n \gamma_k\Big), \quad n \geq 1.$$

For $n \geq 2$,

$$a_n = \Big(\frac{\gamma_{n-1}}{\gamma_n}\Big)^\alpha \exp(-\mu\gamma_n) a_{n-1} = (1-b_n)a_{n-1},$$

with

$$b_n = 1 - \exp\Big(-\mu\gamma_n + \alpha \ln \frac{\gamma_{n-1}}{\gamma_n}\Big) \sim \begin{cases} \mu\gamma_n, & \text{case C1}, \\ (\mu - \frac{\alpha}{\gamma_\star})\gamma_n, & \text{case C2}. \end{cases}$$

Thus $\sum_{n=1}^{\infty} b_n = \infty$ and $\lim_{n\to\infty} a_n = 0$.

(ii)a. Define

$$a_n = \Big(\sum_{k=1}^{n} \gamma_k\Big)^{-\beta} \sum_{k=1}^{n} \frac{\gamma_n}{\big(\sum_{j=1}^{k} \gamma_j\big)^{1+\beta}} \exp\Big(-\mu \sum_{j=k+1}^{n} \frac{\gamma_j}{\sum_{i=1}^{j} \gamma_i}\Big), \quad n \geq 1.$$

Let $n \geq 2$. Then

$$a_n = \left(\frac{\sum_{k=1}^{n-1} \gamma_k}{\sum_{k=1}^{n} \gamma_k}\right)^{\beta} \exp\Big(-\frac{\mu\gamma_n}{\sum_{k=1}^{n} \gamma_k}\Big) a_{n-1} + \frac{\gamma_n}{\sum_{k=1}^{n} \gamma_k} = (1 - b_n) a_{n-1} + b_n b_n^{-1} \frac{\gamma_n}{\sum_{k=1}^{n} \gamma_k}.$$

Note that

$$b_n = 1 - \exp\left(-\frac{\mu\gamma_n}{\sum_{k=1}^{n} \gamma_k} + \beta \ln \frac{\sum_{k=1}^{n-1} \gamma_k}{\sum_{k=1}^{n} \gamma_k}\right) \sim \frac{(\mu + \beta)\gamma_n}{\sum_{k=1}^{n} \gamma_k}, \tag{A.3}$$

$$b_n^{-1} \frac{\gamma_n}{\sum_{k=1}^{n} \gamma_k} \sim \frac{1}{\mu + \beta} =: c_\star.$$

Let $\varepsilon > 0$ and $c = c_\star + \varepsilon$. Then, there is an $n_0 \geq 0$ such that $b_n^{-1} \frac{\gamma_n}{\sum_{k=1}^{n} \gamma_k} \leq c$, $n \geq n_0$. Hence, for $n \geq n_0$,

$$(a_n - c)^+ \leq (1 - b_n)(a_{n-1} - c)^+ + b_n \Big(b_n^{-1} \frac{\gamma_n}{\sum_{k=1}^{n} \gamma_k} - c\Big)^+ = (1 - b_n)(a_{n-1} - c)^+.$$

By (A.3), $\sum_{n=1}^{\infty} b_n = \infty$, thus $\lim_{n\to\infty} (a_n - c)^+ = 0$. Therefore $\limsup_{n\to\infty} a_n \leq c_\star + \varepsilon$ for all $\varepsilon > 0$, whence $\limsup_{n\to\infty} a_n \leq c_\star < \infty$.

(ii)b. Let

$$a_n = \Big(\sum_{k=1}^{n} \gamma_k\Big)^{-\beta} \exp\Big(-\mu \sum_{k=1}^{n} \frac{\gamma_k}{\sum_{j=1}^{k} \gamma_j}\Big), \quad n \geq 1.$$

For $n \geq 2$,

$$a_n = \left(\frac{\sum_{k=1}^{n-1} \gamma_k}{\sum_{k=1}^{n} \gamma_k}\right)^{\beta} \exp\Big(-\frac{\mu\gamma_n}{\sum_{k=1}^{n} \gamma_k}\Big) a_{n-1} = (1 - b_n) a_{n-1},$$

where b_n satisfies (A.3). Thus $\sum_{n=1}^{\infty} b_n = \infty$ and $\lim_{n\to\infty} a_n = 0$. □

Lemma A.5. *Let $\varphi\colon \mathbb{R}^m \to \mathbb{R}$ be a continuously differentiable function and consider $\theta_\star \in \nabla\varphi^{-1}(0)$. Assume that there exist $\mathcal{V}$ an open neighborhood of $\theta_\star$, $\beta \in (0, 1)$ and $\zeta > 0$ such that*

$$|\varphi(\theta) - \varphi(\theta_\star)|^{\beta} \leq \zeta \|\nabla\varphi(\theta)\|, \quad \theta \in \mathcal{V}.$$

Set $\beta' \in (\beta, 1]$. Then, there exists an open neighborhood $\mathcal{V}'$ of $\theta_\star$ such that

$$|\varphi(\theta) - \varphi(\theta_\star)|^{\beta'} \leq \|\nabla\varphi(\theta)\|, \quad \theta \in \mathcal{V}'.$$

Proof. Note that

$$\frac{|\varphi(\theta) - \varphi(\theta_\star)|^{\beta'}}{\|\nabla\varphi(\theta)\|} \leq \zeta |\varphi(\theta) - \varphi(\theta_\star)|^{\beta' - \beta}, \quad \theta \in \mathcal{V} \setminus \nabla\varphi^{-1}(0).$$

By the continuity of φ, the right hand side above vanishes when $\mathcal{V} \ni \theta \to \theta_\star$. There then exists an open neighborhood $\mathcal{V}' \subset \mathcal{V}$ of $\theta_\star$ such that $\zeta |\varphi(\theta) - \varphi(\theta_\star)|^{\beta' - \beta} \leq 1$, $\theta \in \mathcal{V}'$. Thus

$$|\varphi(\theta) - \varphi(\theta_\star)|^{\beta'} \leq \|\nabla\varphi(\theta)\|, \quad \theta \in \mathcal{V}' \setminus \nabla\varphi^{-1}(0). \tag{A.4}$$

Since $\mathcal{V}' \subset \mathcal{V}$,

$$|\varphi(\theta) - \varphi(\theta_\star)|^{\beta'} \leq \zeta \, \|\nabla\varphi(\theta)\| \, |\varphi(\theta) - \varphi(\theta_\star)|^{\beta'-\beta} = 0, \quad \theta \in \mathcal{V}' \cap \nabla\varphi^{-1}(0). \tag{A.5}$$

Combining (A.4) and (A.5),

$$|\varphi(\theta) - \varphi(\theta_\star)|^{\beta'} \leq \|\nabla\varphi(\theta)\|, \quad \theta \in \mathcal{V}'.$$

□

The concavity inequality below is retrieved in [AB09, Equation (8)] and [CF23, Equation (33)] by applying the tangent line method to the convex functions $x \in \mathbb{R}_+ \mapsto -x^{1-\sigma}$ and $x \in \mathbb{R}_+ \mapsto x^{\frac{1}{1-\sigma}}$, $\sigma \in (0,1)$. We rely here instead on a first order Taylor-Lagrange expansion of $x \in \mathbb{R}_+ \mapsto x^{1-\sigma}$, and complement our result with an assertion for $\sigma = 1$.

Lemma A.6. *Let $0 \leq x \leq y$. Then, for $\sigma \in (0,1)$,*

$$\frac{y-x}{y^\sigma} \leq \frac{y^{1-\sigma} - x^{1-\sigma}}{1-\sigma}.$$

Moreover, if $x > 0$,

$$\frac{y-x}{y} \leq \ln y - \ln x.$$

Proof. Suppose $x < y$. Then, for $\sigma \in (0,1)$,

$$y^{1-\sigma} - x^{1-\sigma} = (y-x) \int_0^1 \frac{(1-\sigma)\mathrm{d}t}{(x + t(y-x))^\sigma} \geq (1-\sigma) y^{-\sigma}(y-x).$$

Assume that, in addition, $0 < x$. Then

$$\ln y - \ln x = (y-x) \int_0^1 \frac{\mathrm{d}t}{x + t(y-x)} \geq y^{-1}(y-x).$$

□

Bibliography

[AB09] H. Attouch and J. Bolte. "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features". In: *Mathematical Programming* 116.1 (2009), pp. 5–16. ISSN: 1436-4646. DOI: 10.1007/s10107-007-0133-5.

[AL24] J. An and J. Lu. *Convergence of stochastic gradient descent under a local Łojasiewicz condition for deep neural networks*. 2024. arXiv: 2304.09221 [cs.LG].

[AMA05] P. A. Absil, R. Mahony, and B. Andrews. "Convergence of the iterates of descent methods for analytic cost functions". In: *SIAM Journal on Optimization* 16.2 (2005), pp. 531–547. DOI: 10.1137/040605266.

[BC17] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017, pp. 91–109. ISBN: 978-3-319-48311-5. DOI: 10.1007/978-3-319-48311-5_5.

[BDL07] J. Bolte, A. Daniilidis, and A. Lewis. "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems". In: *SIAM Journal on Optimization* 17.4 (2007), pp. 1205–1223. DOI: 10.1137/050644641.

[BFP09] O. Bardou, N. Frikha, and G. Pagès. "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling". In: *Monte Carlo Methods and Applications* 15.3 (2009), pp. 173–210. DOI: 10.1515/MCMA.2009.011.

[BM88] E. Bierstone and P. D. Milman. "Semianalytic and subanalytic sets". In: *Publications Mathématiques de l'IHÉS* 67 (1988), pp. 5–42.

[BT00] D. P. Bertsekas and J. N. Tsitsiklis. "Gradient convergence in gradient methods with errors". In: *SIAM Journal on Optimization* 10.3 (2000), pp. 627–642. DOI: 10.1137/S1052623497331063.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CF23] E. Chouzenoux and J.-B. Fest. "Convergence analysis of block majorize-minimize subspace approach". In: *Optimization Letters* (2023). ISSN: 1862-4480. DOI: 10.1007/s11590-023-02055-z.

[CFR23] E. Chouzenoux, J.-B. Fest, and A. Repetti. *A Kurdyka-Łojasiewicz property for stochastic optimization algorithms in a non-convex setting*. 2023. arXiv: 2302.06447 [math.OC].

[Cha22] S. Chatterjee. *Convergence of gradient descent for deep neural networks*. 2022. arXiv: 2203.16462 [cs.LG].

[CP17] E. Chouzenoux and J.-C. Pesquet. "A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation". In: *IEEE Transactions on Signal Processing* 65.18 (2017), pp. 4770–4783. DOI: 10.1109/TSP.2017.2709265.

[Cur44] H. B. Curry. "The method of steepest descent for non-linear minimization problems". In: *Quarterly of Applied Mathematics* 2 (1944), pp. 258–261. DOI: 10.1090/qam/10667.

[Dan+22] M. Danilova, P. Dvurechensky, A. Gasnikov, E. Gorbunov, S. Guminov, D. Kamzolov, and I. Shibaev. "Recent theoretical advances in non-convex optimization". In: *High-Dimensional Optimization and Probability: With a View Towards Data Science*. Springer International Publishing, 2022, pp. 79–163. ISBN: 978-3-031-00832-0. DOI: 10.1007/978-3-031-00832-0_3.

[Dav+20] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. "Stochastic subgradient method converges on tame functions". In: *Foundations of Computational Mathematics* 20.1 (2020), pp. 119–154. ISSN: 1615-3383. DOI: 10.1007/s10208-018-09409-5.

[DK23] S. Dereich and S. Kassing. "Central limit theorems for stochastic gradient descent with averaging for stable manifolds". In: *Electronic Journal of Probability* 28 (2023), pp. 1–48. DOI: 10.1214/23-EJP947.

[DK24] S. Dereich and S. Kassing. "Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes". In: *Journal of Machine Learning* 3.3 (2024), pp. 245–281. ISSN: 2790-2048. DOI: 10.4208/jml.240109.

[Duf96] M. Duflo. *Algorithmes Stochastiques*. Springer, 1996.

[FGP15] P. Frankel, G. Garrigos, and J. Peypouquet. "Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates". In: *Journal of Optimization Theory and Applications* 165.3 (2015), pp. 874–900. ISSN: 1573-2878. DOI: 10.1007/s10957-014-0642-3.

[For15] G. Fort. "Central limit theorems for stochastic approximation with controlled Markov chain dynamics". In: *ESAIM: PS* 19 (2015), pp. 60–80. DOI: 10.1051/ps/2014013.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[GG24] G. Garrigos and R. M. Gower. *Handbook of convergence theorems for (stochastic) gradient methods*. 2024. arXiv: 2301.11235 [math.OC].

[Gou20] X. Gourdon. *Les Maths en Tête: Analyse*. Édition Marketing Ellipses, 2020. ISBN: 9782340056992.

[GP23] S. Gadat and F. Panloup. "Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm". In: *Stochastic Processes and their Applications* 156 (2023), pp. 312–348. ISSN: 0304-4149. DOI: 10.1016/j.spa.2022.11.012.

[JO19] K.-S. Jun and F. Orabona. "Parameter-free online convex optimization with sub-exponential noise". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Vol. 99. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1802–1823.

[KL13] A. A. Kirilenko and A. W. Lo. "Moore's law versus Murphy's law: algorithmic trading and its discontents". In: *Journal of Economic Perspectives* 27.2 (2013), pp. 51–72. DOI: 10.1257/jep.27.2.51.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition". In: *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2016, pp. 795–811. ISBN: 978-3-319-46128-1.

[KR23] A. Khaled and P. Richtárik. "Better theory for SGD in the nonconvex world". In: *Transactions on Machine Learning Research* (2023). ISSN: 2835-8856.

[Kur98] K. Kurdyka. "On gradients of functions definable in o-minimal structures". In: *Annales de l'Institut Fourier* 48.3 (1998), pp. 769–783. DOI: 10.5802/aif.1638.

[Lei+20] Y. Lei, T. Hu, G. Li, and K. Tang. "Stochastic gradient descent for nonconvex learning without bounded gradient assumptions". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.10 (2020), pp. 4394–4400. DOI: 10.1109/TNNLS.2019.2952219.

[Let21] A. Letcher. "On the impossibility of global convergence in multi-loss optimization". In: *International Conference on Learning Representations*. 2021.

[Liu+23] C. Liu, D. Drusvyatskiy, M. Belkin, D. Davis, and Y. Ma. "Aiming towards the minimizers: fast convergence of SGD for overparametrized problems". In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 60748–60767.

[Łoj65] S. Łojasiewicz. *Ensembles Semi-Analytiques*. Institut des Hautes Études Scientifiques, 1965.

[Łoj84] S. Łojasiewicz. *Sur les trajectoires du gradient d'une fonction analytique réelle*. Bologna, 1984.

[LP15] G. Li and T. K. Pong. "Global Convergence of splitting methods for nonconvex composite optimization". In: *SIAM Journal on Optimization* 25.4 (2015), pp. 2434–2460. DOI: 10.1137/140998135.

[LZ24] Z. Liu and Z. Zhou. "Revisiting the last-iterate convergence of stochastic gradient methods". In: *The Twelfth International Conference on Learning Representations*. 2024.

[LZB22] C. Liu, L. Zhu, and M. Belkin. "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks". In: *Applied and Computational Harmonic Analysis* 59 (2022), pp. 85–116. ISSN: 1063-5203. DOI: 10.1016/j.acha.2021.12.009.

[MDB24] L. Madden, E. Dall'Anese, and S. Becker. "High probability convergence bounds for non-convex stochastic gradient descent with sub-Weibull noise". In: *Journal of Machine Learning Research* 25.241 (2024), pp. 1–36.

[Mer+20] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. "On the almost sure convergence of stochastic gradient descent in non-convex problems". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1117–1128.

[Ora20] F. Orabona. *Almost sure convergence of SGD on smooth non-convex functions*. 2020. URL: parameterfree.com.

[Ost74] A. M. Ostrowski. "Solution of equations in Euclidean and Banach spaces". In: *SIAM Review* 16 (1974), pp. 564–564.

[Pol63] B. T. Polyak. "Gradient methods for minimizing functionals". In: *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* 3.4 (1963), pp. 643–653.

[QMM24] J. Qiu, B. Ma, and A. Milzarek. *Convergence of SGD with momentum in the nonconvex case: a novel time window-based analysis*. 2024. arXiv: 2405.16954 [math.OC].

[RM51] H. Robbins and S. Monro. "A stochastic approximation method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407. DOI: 10.1214/aoms/1177729586.

[RS71] H. Robbins and D. Siegmund. "A convergence theorem for non negative almost supermartingales and some applications". In: *Optimizing Methods in Statistics*. Academic Press, 1971, pp. 233–257.

[RU00] R. T. Rockafellar and S. Uryasev. "Optimization of conditional value-at-risk". In: *Journal of Risk* 2.3 (2000), pp. 21–41. DOI: 10.21314/JOR.2000.038.

[Rud17] S. Ruder. *An overview of gradient descent optimization algorithms*. 2017. arXiv: 1609.04747 [cs.LG].

[Sa+22] C. De Sa, S. Kale, J. D. Lee, A. Sekhari, and K. Sridharan. "From gradient flow on population loss to learning with stochastic gradient descent". In: *Advances in Neural Information Processing Systems*. 2022.

[Tal01] E. Talvila. "Necessary and sufficient conditions for differentiating under the integral sign". In: *The American Mathematical Monthly* 108.6 (2001), pp. 544–548. ISSN: 00029890, 19300972.

[Wei+24] S. Weissmann, S. Klein, W. Azizian, and L. Döring. *Almost sure convergence rates of stochastic gradient methods under gradient domination*. 2024. arXiv: 2405.13592 [cs.LG].