On the almost sure convergence of stochastic gradient descent in non-convex problems

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, Volkan Cevher

ON THE ALMOST SURE CONVERGENCE OF STOCHASTIC GRADIENT DESCENT IN NON-CONVEX PROBLEMS

PANAYOTIS MERTIKOPOULOS[*,◇,c], NADAV HALLAK[♯], ALI KAVIS[♯], AND VOLKAN CEVHER[*]

Abstract. This paper analyzes the trajectories of stochastic gradient descent (SGD) to help understand the algorithm's convergence properties in non-convex problems. We first show that the sequence of iterates generated by SGD remains bounded and converges with probability 1 under a very broad range of step-size schedules. Subsequently, going beyond existing positive probability guarantees, we show that SGD avoids strict saddle points/manifolds with probability 1 for the entire spectrum of step-size policies considered. Finally, we prove that the algorithm's rate of convergence to Hurwicz minimizers is $\mathcal{O}(1/n^p)$ if the method is employed with a $\Theta(1/n^p)$ step-size. This provides an important guideline for tuning the algorithm's step-size as it suggests that a cool-down phase with a vanishing step-size could lead to faster convergence; we demonstrate this heuristic using ResNet architectures on CIFAR.

1. Introduction

Owing to its simplicity and empirical successes, stochastic gradient descent (SGD) has become the de facto method for training a wide range of models in machine learning. This paper examines the properties of SGD in non-convex problems with the aim of answering the following questions:

(Q1) Does SGD *always* converge?

(Q2) Does SGD *always* avoid spurious critical regions, such as non-isolated saddle points, etc.?

(Q3) How fast does SGD converge to local minima as a function of the method's step-size policy?

We provide the following precise answers to these questions:

[*] Univ. Grenoble Alpes, CNRS, Inria, LIG, 38000, Grenoble, France.

[◇] Criteo AI Lab.

[♯] École Polytechnique Fédérale de Lausanne (EPFL).

[c] Corresponding author.

E-mail addresses: panayotis.mertikopoulos@imag.fr, nadav.hallak@epfl.ch, ali.kavis@epfl.ch, put.email@here.

2020 *Mathematics Subject Classification.* Primary 90C26, 62L20; secondary 90C30, 90C15, 37N40.

Key words and phrases. Non-convex optimization; stochastic gradient descent; stochastic approximation.

This research was partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET). P. Mertikopoulos is also grateful for financial support by the French National Research Agency (ANR) under grant no. ANR–16–CE33–0004–01 (ORACLESS). N. Hallak and A. Kavis were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 725594 - time-data). V. Cevher gratefully acknowledges the support of the Swiss National Science Foundation (SNSF) under grant № 200021–178865 / 1, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 725594 - time-data), and 2019 Google Faculty Research Award.

On (Q1): Under mild conditions for the function to be optimized, and allowing for a wide range of step-size schedules of the form $\Theta(1/n^p)$ for $p \in (0,1]$, the iterate sequence X_n of SGD converges with probability 1. In contrast to existing mean squared error guarantees of the form $\mathbb{E}[\|\nabla f(X_n)\|^2] \to 0$ (where f is the problem's objective), this is a stronger, trajectory convergence result: It is not a guarantee that holds on average, but a convergence certificate that applies with probability 1 to *any* instantiation of the algorithm.

On (Q2): With probablity 1, the trajectories of SGD avoid all strict saddle manifolds – i.e., sets of critical points x^* with at least one negative Hessian eigenvalue ($\lambda_{\min}(\nabla^2 f(x^*)) < 0$). Such manifolds include ridge hypersurfaces and other connected sets of non-isolated saddle points that are common in the loss landscapes of overparametrized neural networks [27]. In this way, our result complements and extends a series of saddle avoidance results for *deterministic* gradient descent [10, 11, 23, 24, 34], and with *high probability* [12] or *in expectation* [43] for stochastic gradient descent.

On (Q3): If SGD is run with a step-size schedule of the form $\gamma_n = \Theta(1/n^p)$ for some $p \in (0,1]$, it converges at a rate of $\mathbb{E}[\|X_n - x^*\|^2] = \mathcal{O}(1/n^p)$ to local minimizers that are regular in the sense of Hurwicz (i.e., $\nabla^2 f(x^*) \succ 0$). We stress here that this is a "last iterate" convergence guarantee; neither ergodic, nor of a mean-squared gradient norm type. This is crucial for real-world applications because, in practice, SGD training is based on the last generated point.

Taken together, the above suggests that a vanishing step-size policy has significant theoretical benefits: almost sure convergence, avoidance of spurious critical points (again with probability 1), and fast stabilization to local minimizers. We explore these properties in a range of standard non-convex test functions and by training a ResNet architecture for a classification task over CIFAR.

The linchpin of our approach is the *ODE method* of stochastic approximation as pioneered by Benveniste et al. [5], Kushner and Yin [22], Ljung [28], and Benaïm [2]. As such, our analysis combines a wide range of techniques from the theory of dynamical systems along with a series of martingale limit theory tools originally developed by Pemantle [35] and Brandière and Duflo [8].

Related work. Ever since the seminal paper of Robbins and Monro [38], SGD has given rise to a vast corpus of literature that we cannot hope to do justice here. We discuss below only those works which – to the best of our knowledge – are the most relevant to the contributions outlined above.

The first result on the convergence of SGD trajectories is due to Ljung [28, 29], who proved the method's convergence under the boundedness assumption $\sup_n \|X_n\| < \infty$. Albeit intuitive, this assumption is fairly difficult to establish from first principles and the problem's primitives. Because of this, boundedness has persisted in the stochastic approximation literature as a condition that needs to be enforced "by hand", see e.g., Benaïm [2], Borkar [7], Kushner and Yin [22], and references therein. To rid ourselves of this condition, we resort to a series of shadowing arguments that interpolate between continuous and discrete time. Our results also improve on a more recent result by Bertsekas and Tsitsiklis [6] who use a completely different analysis to dispense of boundedness via the use of more restrictive, rapidly decaying step-size policies.

On the issue of saddle-point avoidance, Pemantle [35] and Brandière and Duflo [8] showed that SGD avoids *hyperbolic* saddle points ($\lambda_{\min}(\nabla^2 f(x^*)) < 0$, $\det \nabla^2 f(x^*) \neq 0$) with probability 1. More recently, and under different assumptions, Ge et al. [12] showed that SGD avoids *strict* saddle points ($\lambda_{\min}(\nabla^2 f(x^*)) < 0$) with high probability, whereas the work of Vlaski and Sayed [43] guarantees escape from strict saddles in expectation. By

comparison, our paper shows that strict saddles are avoided *with probability* 1, thus providing the missing link between these two threads; for completenes, we review these results in detail in Section 4.3.

The papers mentioned above should be disjoined from an extensive literature on saddle-point avoidance results for *deterministic* gradient descent [10, 11, 20, 23, 24, 33, 34]. Given that these works focus exclusively on deterministic methods, they have no bearing on our work here.

Finally, regarding the rate of convergence of SGD in non-convex problems, Ghadimi and Lan [13, 14] established a series of bounds of the form $\mathbb{E}[\|\nabla f(X_R)\|^2] = \mathcal{O}(1/\sqrt{T})$, where R is drawn randomly from the running horizon $\{1,\dots,T\}$ of the process. More recently, Lei et al. [26] provided a non-asymptotic rate analysis for α-Holder smooth functions, without a bounded gradient assumption; specifically, Lei et al. [26] proved that, for some T, $\min_{n\leq T}\mathbb{E}[\|\nabla f(X_n)\|^2] = \mathcal{O}(T^{p-1})$ with stepsize $\gamma_n = \gamma/n^p$ and $p\in(1/(1+\alpha),1)$. There is no overlap of our results or analysis with these works, and we are not aware of convergence guarantees similar to our own in the literature.

Notation. In the rest of our paper, $\mathbb{R}^d$ denotes a d-dimensional Euclidean space. We also write $\langle\cdot\,,\cdot\rangle$ for the inner product on $\mathbb{R}^d$, $\|\cdot\|$ for the induced norm, and $\mathbb{S}^{d-1} = \{x\in\mathbb{R}^d : \|x\| = 1\}$ for the unit hypersphere of $\mathbb{R}^d$. Since the space is Euclidean, we make no distinction between primal and dual vectors (or norms).

2. Problem setup and assumptions

2.1. **Problem setup.** Throughout the sequel, we focus on the non-convex optimization problem

$$\text{minimize}_{x\in\mathbb{R}^d}\ f(x), \tag{Opt}$$

where $f\colon\mathbb{R}^d\to\mathbb{R}$ is a d-times differentiable function satisfying the following blanket assumptions.

Assumption 1. f is *G-Lipschitz and L-smooth*, i.e.,

$$\|\nabla f(x)\|\leq G \quad\text{and}\quad \|\nabla f(x')-\nabla f(x)\|\leq L\|x'-x\| \ \text{ for all } x,x'\in\mathbb{R}^d. \tag{1}$$

Assumption 2. The *sublevels* $L_c\equiv\{x\in\mathbb{R}^d : f(x)\leq c\}$ of f are bounded for all $c<\sup f$.

Assumption 3. The *gradient sublevels* $M_\varepsilon\equiv\{x\in\mathbb{R}^d : \|\nabla f(x)\|\leq\varepsilon\}$ of f are bounded for some $\varepsilon>0$.

Assumptions 1–3 are fairly standard in non-convex analysis and optimization. Taken individually, Assumption 1 is a basic regularity requirement for f; Assumption 2 guarantees the existence of solutions to (Opt) by ruling out vacuous cases like $f(x)=-x$; and, finally, Assumption 3 serves to exclude objectives with near-critical behavior at infinity such as $f(x)=-e^{-x^2}$.[1] Taken together, Assumptions 1–3 further imply that the critical set

$$\mathcal{X}^*\equiv\operatorname{crit}(f)=\{x\in\mathbb{R}^d : \nabla f(x)=0\} \tag{2}$$

of f is nonempty, a fact that we use freely in the sequel.

Typical examples of (Opt) in machine learning comprise neural networks with sigmoid activation functions, underdetermined inverse problems, empirical risk minimization models, etc. In such problems, obtaining accurate gradient input is impractical, so to solve (Opt), we often rely on *stochastic gradient* information, obtained for example by taking a mini-batch of training instances.

[1] Note that Assumption 3 only concerns near-critical points, not regions where $\|\nabla f(x)\|$ may be large.

2.2. **Assumptions on the oracle.** With this in mind, we will assume throughout that the optimizer can access ∇f via a *stochastic first-order oracle* (SFO). Formally, this is a black-box feedback mechanism which, when queried at an input point $x \in \mathbb{R}^d$, returns a random vector $\mathsf{V}(x;\omega)$ with ω drawn from some (complete) probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In more detail, decomposing the oracle's output at x as

$$\mathsf{V}(x;\omega) = \nabla f(x) + Z(x;\omega)\,, \tag{SFO}$$

we make the following assumption.

Assumption 4. The *error term* $Z(x;\omega)$ of (SFO) has

(*a*) *Zero mean:* $\mathbb{E}[Z(x;\omega)] = 0$ (3a)

(*b*) *Finite q-th moments:* $\mathbb{E}[\|Z(x;\omega)\|^q] \leq \sigma^q$ for some $q \geq 2$ and $\sigma \geq 0$. (3b)

Assumption 4 is standard in stochastic optimization and is usually stated with $q = 2$, i.e., as a "finite variance" condition, cf. Benaïm [2], Juditsky et al. [21], Nesterov [32], Polyak [37], and many others. Allowing values of q greater than 2 provides more flexibility in the choice of step-size policies, so we keep (3b) as a blanket assumption throughout. We also formally allow the value $q = \infty$ in (3b), in which case we will say that the noise is *bounded in* L^∞; put simply, this corresponds to the standard assumption that the noise in (SFO) is bounded almost surely.

2.3. **Stochastic gradient descent.** With all this in hand, the stochastic gradient descent (SGD) algorithm can be written as

$$X_{n+1} = X_n - \gamma_n V_n. \tag{SGD}$$

In the above, $n = 1, 2, \dots$ is the algorithm's iteration counter, γ_n is the algorithm's step-size, and V_n is a sequence of gradient signals of the form

$$V_n = \mathsf{V}(X_n;\omega_n) = \nabla f(X_n) + Z_n. \tag{4}$$

Each gradient signal V_n is generated by querying the oracle at X_n with some random seed ω_n. For concision, we write $Z_n \equiv Z(X_n,\omega_n)$ for the gradient error at the n-th iteration and $\mathcal{F}_n = \sigma(X_1,\dots,X_n)$ for the natural filtration of X_n; in this notation, ω_n and V_n are *not* $\mathcal{F}_n$-measurable.

All our results for (SGD) are stated in the framework of the basic assumptions above. The price to pay for this degree of generality is that the analysis requires an intricate interplay between martingale limit theory and the theory of stochastic approximation; we review the relevant notions below.

3. Stochastic approximation

Asymptotic pseudotrajectories. The departure point for our analysis is to rewrite the iterates of (SGD) as $(X_{n+1} - X_n)/\gamma_n = \nabla(f(X_n)) + Z_n$. In this way, (SGD) can be seen as a Robbins–Monro discretization of the continuous-time *gradient dynamics*

$$\dot{x}(t) = -\nabla f(x(t)). \tag{GD}$$

The main motivation for this comparison is that f is a strict Lyapunov function for (GD), indicating that its solution orbits converge to the critical set $\mathcal{X}^*$ of f (see the supplement for a formal statement and proof of this fact). As such, if the trajectories of (SGD) are "good enough" approximations of the solutions of (GD), one would expect (SGD) to enjoy similar convergence properties.

To make this idea precise, we first connect continuous and discrete time by letting $\tau_n = \sum_{k=1}^{n} \gamma_k$ denote the time that has "elapsed" for (SGD) up to iteration counter n

(inclusive); that is, a step-size in discrete time is translated to elapsed time in the continuous case, and vice-versa. We may then define the continuous-time interpolation of an iterate sequence X_n of (SGD) as

$$X(t) = X_n + [(t-\tau_n)/(\tau_{n+1}-\tau_n)]\,(X_{n+1}-X_n) \qquad \text{for all } t \in [\tau_n, \tau_{n+1}]. \tag{5}$$

To compare this trajectory to the solutions of (GD), we further need to define the "flow" of (GD) which describes how an ensemble of initial conditions evolves over time. Formally, we let $\Phi\colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ denote the map which sends an initial $x \in \mathbb{R}^d$ to the point $\Phi_t(x) \in \mathbb{R}^d$ by following for time $t \in \mathbb{R}_+$ the solution of (GD) starting at x. We then have the following notion of "asymptotic closeness" between a sequence generated by (SGD) and the flow of the dynamics (GD):

Definition 1. We say that $X(t)$ is an *asymptotic pseudotrajectory* (APT) of (GD) if, for all $T > 0$:

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|X(t+h) - \Phi_h(X(t))\| = 0, \tag{6}$$

The notion of an APT is due to Benaïm and Hirsch [4] and essentially posits that $X(t)$ tracks the flow of (GD) with arbitrary accuracy over windows of arbitrary length as $t \to \infty$. When this is the case, we will slightly abuse terminology and say that the sequence X_n itself comprises an APT of (GD).

With all this in hand, the formal link between (GD) and (SGD) is as follows:

Proposition 1. *Suppose that Assumptions 1 and 4 hold and (SGD) is employed with a step-size sequence such that $\sum_{n=1}^\infty \gamma_n = \infty$ and $\sum_{n=1}^\infty \gamma_n^{1+q/2} < \infty$ with $q \geq 2$ as in Assumption 4. Then, with probability 1, X_n is an APT of (GD).*

Corollary 1. *Suppose that (SGD) is run with $\gamma_n = \Theta(1/n^p)$ for some $p \in (2/(q+2), 1]$ and assumptions as in Proposition 1. Then, with probability 1, X_n is an APT of (GD).*

This comparison result plays a key role in the sequel because it delineates the range of step-size policies under which the discrete-time system (SGD) is well-approximated by the continuous-time dynamics (GD). We discuss this issue in detail in the next section.

4. Convergence analysis

Heuristically, the goal of approximating (SGD) via (GD) is to reduce the difficulty of the direct analysis of the former by leveraging the strong convergence properties of the latter. of the latter, which is relatively straightforward to analyze, to the former, which is much more difficult. However, the notion of an APT does not suffice in this regard: in the supplement, we provide an example where a discrete-time APT has a completely different behavior relative to the underlying flow. As such, a considerable part of our analysis below focuses on tightening the guarantees provided by the APT approximation scheme.

4.1. Boundedness and stability of the approximation.

The basic point of failure in the stochastic approximation approach is that APTs may escape to infinity, rendering the whole scheme useless, cf. [2, 7] and references therein. It is for this reason that a large part of the literature on SGD explicitly assumes that the trajectories of the process are bounded (precompact), i.e.,

$$\sup_{t\geq 0} \|X(t)\| < \infty \quad (a.s.). \tag{7}$$

However, this is a prohibitively strong assumption for (Opt): unless certified ahead of time, any theoretical result relying on this assumption would be of limited practical value.

Our first result below provides exactly this certification by establishing that (7) is solely an implication of our underlying Assumptions 1–4. It is a non-trivial outcome which provides the key to unlocking the potential of stochastic approximation techniques in the sequel.

Theorem 1. *Suppose that Assumptions 1–4 hold and (SGD) is run with a variable step-size sequence of the form $\gamma_n \propto 1/n^p$ for some $p \in (2/(q+2), 1]$. Then, with probability 1, every APT $X(t)$ of (GD) that is induced by (SGD) has $\sup_{t\geq 0}\|X(t)\| < \infty$.*

Because of the generality of our assumptions, the proof of Theorem 1 involves a delicate combination of non-standard techniques; for completeness, we provide a short sketch below and refer the reader to the supplement for the details.

Sketch of proof of Theorem 1. The main reasoning evolves along the following lines:

Step 1. We first show that, under the stated assumptions, there exists a (possibly random) subsequence X_{n_k} of X_n that converges to $\mathcal{X}^*$; formally, $\liminf_{n\to\infty} \operatorname{dist}(X_n, \mathcal{X}^*) = 0$ (a.s.). As a result, $X(t)$ eventually reaches a sublevel set L_ε whose elements are arbitrarily close to $\mathcal{X}^*$, i.e., there exists some $t_\varepsilon > 0$ such that $X(t_\varepsilon) \in L_\varepsilon$.

Step 2. By a technical argument relying on the regularity assumptions for f (cf. Assumption 1), it can be shown that there exists some *uniform* time window τ such that $X(t)$ remains within uniformly bounded distance to L_ε for all $t \in [t_\varepsilon, t_\varepsilon + \tau]$. Thus, once $X(t)$ gets close to L_ε, it will not escape too far within a fixed length of time.

Step 3. An additional technical argument reveals that, under the stated assumptions for f, the trajectories of (GD) either descend the objective by a uniform amount, or they have reached a neighborhood of the critical set $\mathcal{X}^*$ where further descent is impossible (or irrelevant).

Step 4. By combining the two previous steps, we conclude that $X(t_\varepsilon + \tau) \in L_\varepsilon$ at the end of said window. This argument may then be iterated ad infinitum to show inductively that $X(t) \in L_\varepsilon$ for all intervals of the form $[t_\varepsilon + k\tau, t_\varepsilon + (k+1)\tau]$.

Since L_ε is bounded (by Assumption 2), we conclude that $X(t)$ remains in a compact set for all $t \geq 0$, i.e., $X(t)$ is precompact. The conclusion of Theorem 1 then follows by Corollary 1. ∎

4.2. **Almost sure convergence.** By virtue of Theorem 1, we are now in a position to state our almost sure convergence result:

Theorem 2. *Suppose that Assumptions 1–4 hold and (SGD) is run with a variable step-size sequence of the form $\gamma_n = \Theta(1/n^p)$ for some $p \in (2/(q+2), 1]$. Then, with probability 1, X_n converges to a (possibly random) connected component $\mathcal{X}^*_\infty$ of $\mathcal{X}^*$ over which f is constant.*

Corollary 2. *With assumptions as in Theorem 2, we have the following:*

(1) $f(X_n)$ converges (a.s.) to some critical value f_∞.

(2) Any limit point of X_n is (a.s.) a critical point of f.

Theorem 2 extends a range of existing treatments of (SGD) under explicit boundedness assumptions of the form (7), cf. [2, 7, 28] and references therein. It also improves on a similar result by Bertsekas and Tsitsiklis [6] who use a completely different analysis to dispense of boundedness requirements via the use of more restrictive step-size policies. Specifically, Bertsekas and Tsitsiklis [6] require the Robbins–Monro summability conditions $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$ under a bounded variance assumption. In this regard, our analysis extends to more general step-size policies, while that of Bertsekas and Tsitsiklis [6] cannot because of its reliance on the Robbins-Siegmund theorem for almost-supermartingales [39]. Among other benefits, this added degree of flexibility is a key advantage of the APT approach.

The heavy lifting in the proof of Theorem 2 is provided by Theorem 1. Thanks to this boundedness certificate, the total chain of implications is relatively short, so we provide it in full below.

Proof of Theorem 2. Under the stated assumptions, f is a strict Lyapunov function for (GD) in the sense of Benaïm [2, Chap. 6.2]. Specifically, this means that $f(\Phi_t(x))$ is strictly decreasing in t unless x is a stationary point of (GD). Furthermore, by Sard's theorem [31, Chap. 2], the set $f(\mathcal{X}^*)$ of critical values of f has Lebesgue measure zero – and hence, empty topological interior. Therefore, applying Theorem 5.7 and Proposition 6.4 of Benaïm [2] in tandem, we conclude that any precompact asymptotic pseudotrajectory of (GD) converges to a connected component $\mathcal{X}^*_\infty$ of $\mathcal{X}^*$ over which f is constant. Since Theorem 1 guarantees that the APTs of (GD) induced by (SGD) are bounded with probability 1, our claim follows. ∎

4.3. **Avoidance analysis.** Theorem 2 represents a strong convergence guarantee but, at the same time, it does not characterize the component of $\mathcal{X}^*$ to which X_n converges. The rest of this section is devoted to showing that X_n does not converge to a component of $\mathcal{X}^*$ that only consists of saddle points (a *saddle-point manifold*). Specifically, we will make precise the following informal statement:

(SGD) *avoids strict saddles – and sets thereof – with probability* 1.

To set the stage for the analysis to come, we begin by reviewing some classical and recent results on the avoidance of saddle points. We then present our general results towards the end of the section.

To begin, a crucial role will be played in the sequel by the *Hessian matrix* of f, viz.

$$H(x) \equiv \nabla^2 f(x) \equiv (\partial_i \partial_j f(x))_{i,j=1,\dots,d}. \tag{8}$$

Since $H(x)$ is symmetric, all of its eigenvalues are real. If x^* is a critical point of f and $\lambda_{\min}(H(x^*)) < 0$, we say that x^* is a *strict saddle point* [23, 24].

By standard results in center manifold theory [42], the space around strict saddle points admits a decomposition into a *stable*, *center* and *unstable* manifold (each of the former two possibly of dimension zero; the latter of dimension at least 1 given that $\lambda_{\min}(H(x^*)) < 0$). Heuristically, under the continuous-time dynamics (GD), directions along the stable manifold of x^* are attracted to x^* at a linear rate, while those along the unstable manifold are repelled (again at a linear rate); the dynamics along the center manifold could be considerably more complicated, but, in the presence of unstable directions, they only emerge from a measure zero of initial conditions. As a result, if x^* is a strict saddle point of f, it stands to reason that (SGD) should "probably" avoid it as well.

In the case of *deterministic* gradient descent with step-size $\gamma < 1/L$, this intuition was made precise by Lee et al. [23, 24] who proved that all but a measure zero of initializations of gradient descent avoid strict saddles. As we discussed in the introduction, this result was then extended to various deterministic settings, with different assumptions for the gradient oracle, the method's step-size, or the structure of the saddle-point manifold, see e.g., [10, 11, 20, 24, 33, 34] and the references therein.

In the stochastic regime, the situation is considerably more involved. Pemantle [35] and Brandière and Duflo [8] were the first to establish the avoidance of hyperbolic unstable equilibria in general stochastic approximation schemes. However, a key requirement in the analysis of these works is that of *hyperbolicity*, which in our setting amounts to asking that $H(x^*)$ is *invertible*. In particular, this means the saddle point in question cannot be isolated, nor can it have a center manifold: both hypotheses are too stringent for applications of SGD to contemporary machine learning models, such as deep net training, so their results do not apply in many cases of practical interest.

More relevant for our purposes is the recent result of Ge et al. [12], who provided the following guarantee. Suppose that f is $(\alpha, \beta, \varepsilon, \delta)$-*strict saddle*, i.e., for all $x \in \mathbb{R}^d$, one of the following holds: (*i*) $\|\nabla f(x)\| \geq \varepsilon$; (*ii*) $\lambda_{\min}(H(x)) \leq -\beta$; or (*iii*) x is δ-close to a local minimum x_c around which f is α-strongly convex. Suppose further that f is bounded,

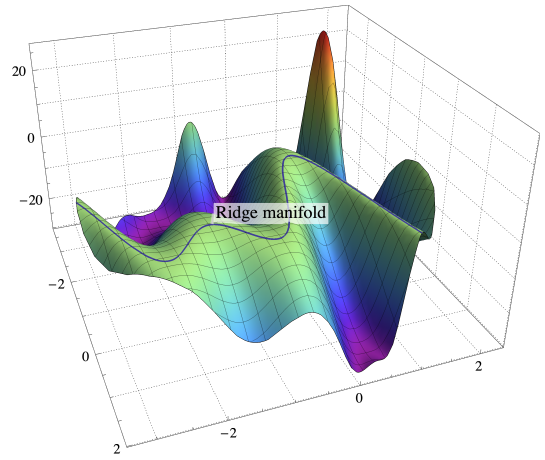


Figure 1: A strict saddle manifold (a ridge), typical of ResNet loss landscapes [27].

L-Lipschitz smooth, and $H(x)$ is ρ-Lipschitz continuous; finally, assume that the noise in the gradient oracle (SFO) is finite (a.s.) and contains a component uniformly sampled from the unit sphere. Then, given a confidence level $\zeta > 0$, and assuming that (SGD) is run with *constant* step-size $\gamma = \mathcal{O}(1/\log(1/\zeta))$, the algorithm produces after a given number of iterations a point which is $\mathcal{O}(\sqrt{\gamma \log(1/(\gamma\zeta))})$-close to x_c, and hence away from any strict saddle of f, with probability at least $1 - \zeta$.

In a more recent paper, Vlaski and Sayed [43] examined the convergence of (SGD) to second-order stationary points. More precisely, they showed that (SGD) guarantees *expected* descent for strict saddle points in a finite number of iterations, and with high probability, (SGD) iterates reach a set of *approximate* second-order stationary points in finite time.

The theory of Pemantle [35] and the result of Ge et al. [12] paint a complementary picture to the above: Pemantle [35] shows that saddle points are avoided with probability 1, provided they are hyperbolic (i.e., $\det \nabla^2 f(x^*) \neq 0$); on the other hand, Ge et al. [12] require much less structure on the saddle point, but they only provide a result with high probability (and ζ cannot be taken to zero because the range of allowable step-sizes would also vanish).[2] Our objective in the sequel is to provide a result that combines the "best of both worlds", i.e., almost sure avoidance of strict saddle points (and sets thereof) with probability 1.

To that end, we make the following assumption for the noise:

Assumption 5. The error term $Z \equiv Z(x;\omega)$ of (SFO) is *uniformly exciting*, i.e., there exists some $c > 0$ such that

$$\mathbb{E}[\langle Z(x;\omega), u\rangle^+] \geq c \tag{9}$$

for all $x \in \mathbb{R}^d$ and all unit vectors $u \in \mathbb{S}^{d-1}$.

This assumption simply means that the average projection of the noise along every ray in $\mathbb{R}^d$ is uniformly positive; in other words, Z "excites" all directions uniformly – though not necessarily *isotropically*. As such, Assumption 5 is automatically satisfied by noisy gradient dynamics (e.g., as in Ge et al. [12]), generic finite sum objectives with at least d summands, etc.

With all this in hand, we say that $\mathcal{S}$ is a *strict saddle manifold* of f if it is a smooth connected component of $\mathcal{X}^*$ such that:

[2] Pemantle [35] employs a vanishing step-size, which is more relevant for us: (SGD) with persistent noise and a constant step-size is an irreducible ergodic Markov chain whose trajectories do not converge *anywhere* [2].

(1) Every $x^* \in \mathcal{S}$ is a strict saddle point of f (i.e., $\lambda_{\min}(H(x^*)) < 0$).
(2) There exist $c_-, c_+ > 0$ such that, for all $x^* \in \mathcal{S}$, all negative eigenvalues of $H(x^*)$ are bounded from above by $-c_- < 0$, and any positive eigenvalues (*if* they exist) are bounded from below by c_+.

Somewhat informally, the definition of a strict saddle manifold implies that the eigenspaces of $H(x^*)$ corresponding to zero, positive, and negative eigenvalues decompose smoothly along $\mathcal{S}$ and $\mathcal{S}$ can be seen as an "integral manifold" of the nullspace of the Hessian of f.

With all this in hand, we are finally in a position to state our main avoidance result.

Theorem 3. *Suppose that* (SGD) *is run with a variable step-size sequence of the form* $\gamma_n \propto 1/n^p$ *for some* $p \in (0, 1]$. *If Assumptions 1–5 hold (with* $q = \infty$ *for Assumption 4), and* $\mathcal{S}$ *is a strict saddle manifold of* f, *we have* $\mathbb{P}(X_n \to \mathcal{S} \text{ as } n \to \infty) = 0$.

Theorem 3 is the formal version of the avoidance principle that we stated in the beginning of this section. Importantly, it makes *no* assumptions regarding the initialization of (SGD) and holds for *any* initial condition.

The proof of Theorem 3 relies on two basic components. The first is a probabilistic estimate, originally due to Pemantle [35], that shows that a certain class of stochastic processes avoid zero with probability 1. The second is a differential-geometric argument, building on Benaïm and Hirsch [3] and Benaïm [2], and relying on center manifold theory to isolate the center/stable and unstable manifolds of $\mathcal{S}$. Combining these two components, it is possible to show that even ambulatory random walks along the stable manifold of $\mathcal{S}$ will eventually be expelled from a neighborhood of $\mathcal{S}$. We provide the details of this argument in the paper's supplement.

4.4. **Rate of convergence.** We conclude our analysis of (SGD) by establishing the algorithm's rate of convergence, as stated in Theorem 4 below. Since f is non-convex, any convergence rate analysis of this type must be a fortiori local; in view of this, we will examine the algorithm's convergence to local minimizers $x^* \in \mathcal{X}^*$ that are regular in the sense of Hurwicz, i.e., $H(x^*) \succ 0$.

Because we are primarily interested in the convergence of the algorithm's trajectories, we focus here on the distance $D_n = \|X_n - x^*\|^2/2$ between the iterates of (SGD) and a local minimizer of f. In this light, our rate guarantee (which we state below), differs substantially from other results in the literature, in both scope and type, as it does not concern the ergodic average $\bar{X}_n = n^{-1}\sum_{k=1}^n X_k$ or the "best iterate" $X_n^{\text{best}} = \arg\min_{k=1,\dots,n}\|\nabla f(X_k)\|$ of (SGD): the former has very weak convergence in convex settings (if at all), while the latter cannot be calculated with access to perfect gradient information for the entire run of the process (in which case, stochastic gradient dynamics would become *ordinary* gradient dynamics).

Theorem 4. *Fix some tolerance level* $\delta > 0$, *let* x^* *be a regular minimizer of* f, *and suppose that Assumption 4 holds. Assume further that* (SGD) *is run with a step-size schedule of the form* $\gamma_n = \gamma/(n+m)^p$ *for some* $p \in (2/(q+2), 1]$ *and large enough* $m, \gamma > 0$. *Then:*

(1) There exist neighborhoods $\mathcal{U}$ *and* $\mathcal{U}_1$ *of* x^* *such that, if* $X_1 \in \mathcal{U}_1$, *the event*

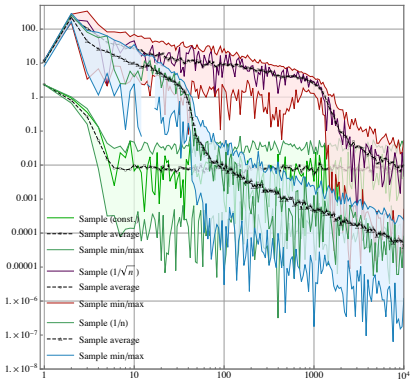
$$\Omega_{\mathcal{U}} = \{X_n \in \mathcal{U} \text{ for all } n = 1, 2, \dots\} \tag{10}$$

occurs with probability at least $1 - \delta$.

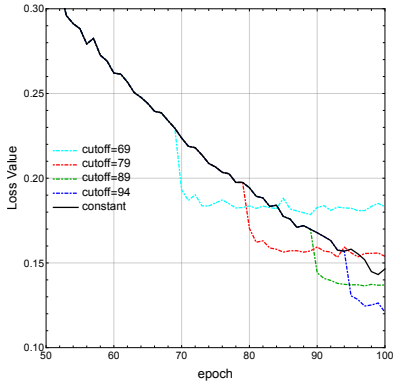
(2) Conditioned on $\Omega_{\mathcal{U}}$, *we have*

$$\mathbb{E}[\|X_n - x^*\|^2 \,|\, \Omega_{\mathcal{U}}] = \mathcal{O}(1/n^p). \tag{11}$$

Remark. Note that Theorem 4 does not presuppose Assumptions 1–3; since the rate analysis is local, the differentiability of f suffices.



(a) Speed of convergence in the Shekel benchmark.



(b) Training ResNet with a cooldown heuristic.

The proof of Theorem 4 relies on showing that *a*) x^* is *stochastically stable*, i.e., with high probability, any initialization that is close enough to x^* remains close enough; and *b*) conditioned on this event, the distance $D_n = (1/2)\|X_n - x^*\|^2$ to a regular local minimizers behaves as an "almost" supermartingale. A major complication that arises here is that this conditioning changes the statistics of the noise, so the martingale property ceases to hold. Overcoming this difficulty requires an intricate probablistic argument that we present in the supplement (where we also provide explicit expressions of the constants in the estimate of Theorem 4).

5. Numerical experiments

As an illustration of our theoretical analysis, we plot in Fig. 2a the convergence rate of (SGD) in the standard Shekel risk benchmark function $f(x) = \sum_{i=1}^{N} \big[\sum_{j=1}^{d}(x_j - a_{ij})^2 + c_i\big]^{-1}$ where $A = (a_{ij})$ is a skew data matrix and $c = (c_1, \dots, c_N)$ is a bias vector of dimension $d = 500$ [19]. For our experiments, we ran $N = 10^3$ instances of (SGD) with a constant, $1/\sqrt{n}$, and $1/n$ step-size schedule, and we plotted the value difference $f(X_n) - f_\infty$ of the sample average (marked black lines) and the min-max spread of the samples for a 95% confidence level region (shaded green, red and blue respectiely for the constant, $1/\sqrt{n}$ and $1/n$ policies respectively). The constant step-size schedule initially performs better, but quickly saturates and is overcome by the $1/n$ schedule; overall, the $1/n$ policy converges faster than the other two by 2 to 4 orders of magnitude.

Coupled with our theoretical results, these tests suggest that a vanishing step-size policy could have significant advantages when used for training machine learning models. The key drawback to this approach is that a rapidly vanishing step-size could cause the algorithm to traverse the loss landscape at a very slow pace and/or get trapped at inferior local minima. However, it also provides a sound theoretical justification for the following "best of both worlds" training heuristic: given a budget of gradient iterations, run SGD with a constant step-size for a fraction of this budget, and then implement a "cooldown" phase with a vanishing step-size for the rest. We demonstrate the benefits of this "cooldown" heuristic in a standard ResNet18 architecture for a classification task over CIFAR10. In particular, in Fig. 2b, we ran (SGD) with a constant step-size for 100 epochs, with checkpoints at different cutoffs; then, at each checkpoint, we launched the "cooldown" period with step-size $1/n$. Fig. 2b demonstrates the improvement due to the cool-off period over the training loss: specifically, it shows that it is always beneficial to run the last training epochs with a vanishing step-size.

6. Concluding remarks

Our aim in this paper was to present a novel trajectory-based analysis of (SGD) showing that, under minimal assumptions, *(i)* all of its limit points are stationary; *(ii)* it avoids strict saddle manifolds with probability 1; and *(iii)* it converges at a fast $\mathcal{O}(1/n)$ rate to regular minimizers. This opens the door to many interesting directions – from constrained/composite problems to adaptive gradient methods. We defer these to the future.

Appendix A. Convergence in continuous time

For completeness, we begin with a proof of the convergence of (GD) under our blanket assumptions:

Proposition A.1 (Gradient flow convergence)**.** *Under Assumptions 1 and 2, every solution $x(t)$ of (GD) converges to $\mathcal{X}^*$.*

Proof of Proposition A.1. To begin, existence and uniqueness of (global) solutions to (GD) follows readily from the Picard–Lindelöf theorem [42] and Assumption 1. With this point settled, and given that the sublevel sets of f are bounded (cf. Assumption 2), the fact that f is non-increasing along the orbits of (GD) shows that $x(t)$ converges to some compact invariant set $\mathcal{K} \subseteq \mathbb{R}^d$.

Suppose now that there exists a sequence of times t_n, $n = 1, 2, \dots$, such that $x(t_n)$ converges to some *non-critical* point $\hat{x} \notin \operatorname{crit}(f) \equiv \mathcal{X}^*$. Letting $c = \|\nabla f(\hat{x})\|^2 > 0$, there exists a neighborhood $\mathcal{U}$ of $\hat{x}$ such that $\|\nabla f(x)\|^2 \geq c/2$ for all $x \in \mathcal{U}$ (again, by Assumption 1). Hence, by passing to a subsequence if necessary, we can assume without loss of generality that $x(t_n) > c/2$ for all n. Furthermore, by Assumption 1 (which implies that $\|\nabla f(x)\| \leq G$ for all $x \in \mathbb{R}^d$) and the definition of (GD), we have

$$\|x(t_n + \tau) - x(t_n)\| \leq \int_{t_n}^{t_n+\tau} \|\nabla f(x(s))\| \, ds \leq G\tau \tag{A.1}$$

for all $\tau > 0$. Therefore, by picking τ sufficiently small, we can assume that $x(t) \in \mathcal{U}$ for all $t \in [t_n, t_n + \tau]$ and all n (recall here that $x(t_n) \in \mathcal{U}$ for all n). Then, by the definition of $\mathcal{U}$, we readily get

$$f(x(t_n + \tau)) - f(x(t_n)) = \int_{t_n}^{t_n+\tau} \frac{d}{dt} f(x(s)) \, ds = -\int_{t_n}^{t_n+\tau} \|\nabla f(x(s))\|^2 \, ds \leq -\frac{c\tau}{2}, \tag{A.2}$$

and hence:

$$\begin{aligned} f(x(t_n + \tau)) - f(x(0)) &= -\int_0^{t_n+\tau} \|\nabla f(x(s))\|^2 \, ds \leq -\sum_{k=1}^{n} \int_{t_k}^{t_k+\tau} \|\nabla f(x(s))\|^2 \, ds \\ &\leq -\sum_{k=1}^{n} \frac{c\tau}{2} = -\frac{nc\tau}{2} \end{aligned} \tag{A.3}$$

i.e., $\lim_{n\to\infty} f(x(t_n+\tau)) = -\infty$, a contradiction. Since $x(t)$ converges to a compact invariant set $\mathcal{K}$, we conclude that $x(t)$ in fact converges to the critical set $\mathcal{X}^*$ of f. ∎

Appendix B. Stability and boundedness of APTs

B.1. **Discrepancies between flows and APTs.** Our first goal in this appendix is to provide a concrete example where asymptotic pseudotrajectories and the underlying continuous-time flow exhibit qualitatively different behaviors in the long run. To that end, consider the autonomous ODE

$$\dot{x}_1(t) = 1 \qquad \dot{x}_2(t) = -\frac{x_2(t)}{1 + x_1(t)} \tag{B.1}$$

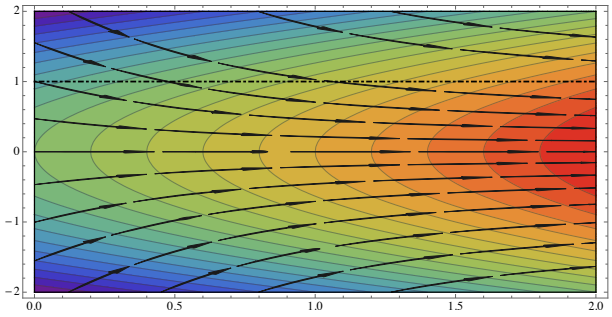


Figure 3: Flowlines vs. asymptotic pseudotrajectories: the dashed black line is an APT of the depicted gradient flow but it stays at a constant height ($x_2 = 1$), even though all flow lines converge to the x_1-axis ($x_2 = 0$).

which is a pseudo-gradient flow of the function $f(x_1, x_2) = x_2^2/2 - x_1$. The general solution of this system with initial condition $(0, b)$ at time $t = 0$ is

$$x(t) = \left(t, \frac{b}{1+t}\right). \tag{B.2}$$

As a result, we have $x_2(t) \to 0$ as $t \to \infty$ from any initial condition (for a graphical illustration, see Fig. 3).

On the other hand, as we show below, the "constant height" curve $X(t) = (t, 1)$ is an asymptotic pseudotrajectory of (B.1). To show this, fix some accuracy threshold $\varepsilon > 0$ and a horizon $T > 0$. Then, with respect to Definition 1, it suffices to show that, for some sufficiently large $t_0 > 0$ and all $h \in [0, T]$, we have

$$|1 - x_2(t_0 + h)| \leq \varepsilon \tag{B.3}$$

for the solution trajectory $x(t) = (x_1(t), x_2(t))$ that passes through the point $(t_0, 1)$ at time t_0.[3]

Substituting in the general solution of (B.1) and backsolving, we readily obtain that this trajectory has

$$x_2(t) = \frac{1+t_0}{1+t}. \tag{B.4}$$

In turn, this implies that the maximal difference between 1 and $x_2(t)$ over a window of size T starting at t_0 is

$$\max_{0\leq h\leq T} |1 - x_2(t_0 + h)| = \frac{1+t_0+T}{1+t_0} - 1 = \frac{T}{1+t_0} \leq \varepsilon \tag{B.5}$$

if t_0 is chosen sufficiently large – specifically, if $t_0 \geq T/\varepsilon - 1$.

Since ε is arbitrary, the above shows that the APT condition (6) holds for all $T > 0$, i.e., $X(t)$ is an APT of (B.1). On the other hand, we have $\lim_{t\to\infty} X_2(t) = 1$, which is different than the limit of *any* solution of (B.1).

B.2. **Boundedness of APTs.** Our aim in the rest of this appendix will be to prove Theorem 1, which, for convenience, we restate below:

Theorem 1. *Suppose that Assumptions 1–4 hold and* (SGD) *is run with a variable step-size sequence of the form $\gamma_n \propto 1/n^p$ for some $p \in (2/(q+2), 1]$. Then, with probability 1, every APT $X(t)$ of* (GD) *that is induced by* (SGD) *has $\sup_{t\geq 0}\|X(t)\| < \infty$.*

[3]That this is so is a consequence of the fact that the trajectories of (B.1) intersect the line $x_2 = 1$ at a vanishing angle as $t \to \infty$. More precisely, if we show the statement in question for t_0, it will also hold for all $\tau \geq t_0$ by virtue of the monotonicity of the exponential function.

To begin, we recall the basic APT property of (SGD):

Proposition 1. *Suppose that Assumptions 1 and 4 hold and (SGD) is employed with a step-size sequence such that $\sum_{n=1}^{\infty}\gamma_n=\infty$ and $\sum_{n=1}^{\infty}\gamma_n^{1+q/2}<\infty$ with $q\geq 2$ as in Assumption 4. Then, with probability 1, X_n is an APT of (GD).*

The proof of Proposition 1 follows by a tandem application of Propositions 4.1 and 4.2 of Benaïm [2], so we omit it; instead, we focus directly on the proof of Theorem 1. To that end, as we explained in the main body of the paper, the first part of our proof consists of showing that (SGD) admits a subsequence converging to $\mathcal{X}^*$, i.e., that $\liminf_{n\to\infty}\operatorname{dist}(X_n,\mathcal{X}^*)=0$:

Lemma B.1. *With assumptions as in Theorem 1, there exists a (possibly random) subsequence X_{n_k} of X_n that converges to $\mathcal{X}^*$; formally, $\liminf_{n\to\infty}\operatorname{dist}(X_n,\mathcal{X}^*)=0$ (a.s.).*

Before proving Lemma B.1, we will require an intermediate result:

Lemma B.2. *Let $\mathcal{C}$ be a closed subset of $\mathbb{R}^d$ such that $\mathcal{X}^*\cap\mathcal{C}=\varnothing$. Then, under Assumption 3, $\inf_{x\in\mathcal{C}}\|\nabla f(x)\|>0$.*

Proof. Arguing by contradiction, assume there exists some sequence $x_n\in\mathcal{C}$ such that $\|\nabla f(x_n)\|\to 0$ as $n\to\infty$. If x_n admits a subsequence converging to some limit point $\hat{x}\in\mathcal{C}$, then, by continuity (recall that f is assumed C^d), we would also have $\|\nabla f(\hat{x})\|=0$. In turn, this would imply $\hat{x}\in\mathcal{X}^*$, contradicting the assumption that $\mathcal{C}$ is closed and disjoint from $\mathcal{X}^*$.

Therefore, to prove our claim, it suffices to examine the case where x_n has no convergent subsequence, i.e., $\liminf_{n\to\infty}\|x_n\|=\infty$. However, this would mean that the gradient sublevel set $M_\varepsilon=\{x\in\mathbb{R}^d:\|\nabla f(x)\|\leq\varepsilon\}$ is unbounded for all $\varepsilon>0$, in contradiction to Assumption 3. We conclude that $\liminf_{n\to\infty}\|\nabla f(x_n)\|>0$ for every sequence x_n in $\mathcal{C}$, i.e., $\liminf_{x\in\mathcal{C}}\|\nabla f(x)\|>0$. ∎

Proof of Lemma B.1. Assume ad absurdum that the event

$$\Omega_0=\{\liminf_{n\to\infty}\operatorname{dist}(X_n,\mathcal{X}^*)>0\} \tag{B.6}$$

occurs with positive probability. By Lemma B.2, if $\liminf_{n\to\infty}\operatorname{dist}(X_n,\mathcal{X}^*)>0$, we must also have $\liminf_{n\to\infty}\|\nabla f(X_n)\|>0$ (since X_n will eventually be contained in a closed set that is disjoint from $\mathcal{X}^*$). Therefore, fixing a realization X_n, $n=1,2,\dots$, of (SGD) such that Ω_0 holds, there exists some (random) positive constant $c>0$ with $\|f(X_n)\|^2\geq c$ for all sufficiently large n; without loss of generality, we may – and will – assume in the sequel that this actually holds for all $n\geq 1$.

In view of all this, by the smoothness assumption for f and the definition of (SGD) we readily get:

$$\begin{aligned}f(X_{n+1})=f(X_n-\gamma_nV_n)&\leq f(X_n)-\gamma_n\langle\nabla f(X_n),V_n\rangle+\frac{L}{2}\gamma_n^2\|V_n\|^2\\&=f(X_n)-\gamma_n\|\nabla f(X_n)\|^2-\gamma_n\langle\nabla f(X_n),Z_n\rangle+\frac{L}{2}\gamma_n^2\|V_n\|^2\\&\leq f(X_n)-\gamma_nc-\gamma_n\xi_n+\gamma_n^2L\|\nabla f(X_n)\|^2+\gamma_n^2L\|Z_n\|^2,\end{aligned} \tag{B.7}$$

where we set $\xi_n=\langle\nabla f(X_n),Z_n\rangle$. Therefore, setting $f_n=f(X_n)$ and telescoping, we obtain

$$f_{n+1}\leq f_1-\tau_n\left[c+\underbrace{\frac{\sum_{k=1}^n\gamma_k\xi_k}{\tau_n}}_{A_n}-L\underbrace{\frac{\sum_{k=1}^n\gamma_k^2\|\nabla f(X_k)\|^2}{\tau_n}}_{B_n}-L\underbrace{\frac{\sum_{k=1}^n\gamma_k^2\|Z_k\|^2}{\tau_n}}_{C_n}\right], \tag{B.8}$$

where $\tau_n = \sum_{k=1}^{n} \gamma_k$ is the "elapsed time" of X_n as defined in Section 3. We will proceed to show that all the summands in the brackets of (B.8) except the first converge to 0; since $c > 0$ and $\tau_n \uparrow \infty$, this will show that $\lim_{n\to\infty} f_n = -\infty$, in direct contradiction to Assumption 2.

We carry out this plan term-by-term below:

(1) For the first term (A_n), note that

$$\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] = \mathbb{E}[\langle \nabla f(X_n), Z_n\rangle \,|\, \mathcal{F}_n] = \langle \nabla f(X_n), \mathbb{E}[Z_n \,|\, \mathcal{F}_n]\rangle = 0 \tag{B.9}$$

by Assumption 4. This means that $\sum_{k=1}^{n} \gamma_k \xi_k$ is a zero-mean martingale, so, by the law of large numbers for martingale difference sequences [15, Theorem 2.18], we have $\gamma_n^{-1} \sum_{k=1}^{n} \gamma_k \xi_k \to 0$ (a.s.) on the event

$$\Omega_1 = \left\{ \sum_{n=1}^{\infty} \frac{\gamma_n^2}{\tau_n^2} \mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_n] < \infty \right\}. \tag{B.10}$$

However, by Assumptions 1 and 4, we have

$$\begin{aligned}
\mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_n] &= \mathbb{E}[\langle \nabla f(X_n), Z_n\rangle^2 \,|\, \mathcal{F}_n] \\
&\leq G^2 \,\mathbb{E}[\|Z_n\|^2 \,|\, \mathcal{F}_n] && \{\text{by Assumption 1}\} \\
&\leq G^2 \,\mathbb{E}[\|Z_n\|^q \,|\, \mathcal{F}_n]^{2/q} && \{\text{by Jensen}\} \\
&\leq G^2 \sigma^2 && \{\text{by Assumption 4}\}
\end{aligned}$$

where, in the second-to-last line, we applied Jensen's inequality to the function $z \mapsto z^{q/2}$ (recall here that $q \geq 2$). Moreover, for all $p \in (0, 1]$, we have $\gamma_n^2/\tau_n^2 = \tilde{\mathcal{O}}(1/n^2)$, so $\sum_{n=1}^{\infty} \gamma_n^2/\tau_n^2 < \infty$. Thus, going back to (B.10), we conclude that $\gamma_n^{-1} \sum_{k=1}^{n} \gamma_k \xi_k \to 0$ with probability 1.

(2) For the second term (B_n), simply note that $\|\nabla f(X_n)\|^2 \leq G^2$, so we have:

$$B_n = \frac{\sum_{k=1}^{n} \gamma_k^2 \|\nabla f(X_k)\|^2}{\tau_n} = \begin{cases} \mathcal{O}(1/n^p) & \text{if } 0 < p < 1/2, \\ \mathcal{O}(\log n/\sqrt{n}) & \text{if } p = 1/2, \\ \mathcal{O}(1/n^{1-p}) & \text{if } 1/2 < p < 1, \\ \mathcal{O}(1/\log n) & \text{if } p = 1. \end{cases} \tag{B.11}$$

Thus, from the above, we conclude that $B_n \to 0$.

(3) For the third term (C_n), we will require a series of estimates. First, with a fair degree of hindsight, let $Q_n = \sum_{k=1}^{n} \gamma_k^{1+q/2} \|Z_k\|^q$. Noting that $\mathbb{E}[\|Z_n\|^q] < \infty$ (a.s.) and $\mathbb{E}[Q_n \,|\, \mathcal{F}_n] = Q_{n-1} + \gamma_n^{1+q/2} \|Z_n\|^q \geq Q_{n-1}$ for all $n = 1, 2, \dots$, we deduce that Q_n is a submartingale. Furthermore, we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{\infty} \gamma_n^{1+q/2} \|Z_n\|^q\right] &\leq \sum_{n=1}^{\infty} \gamma_n^{1+q/2} \,\mathbb{E}[\|Z_n\|^q] \leq \sum_{n=1}^{\infty} \gamma_n^{1+q/2} \sigma^q \\
&= \mathcal{O}\left(\sum_{n=1}^{\infty} n^{-\frac{p(q+2)}{2}}\right) < \infty,
\end{aligned} \tag{B.12}$$

i.e., Q_n is bounded in L^1 (recall that $p > 2/(q+2)$ by assumption). Hence, by Doob's submartingale convergence theorem [15, Theorem 2.1], it follows that Q_n converges (a.s.) to a random variable Q_∞ with $\mathbb{E}[Q_\infty] < \infty$ (and hence $Q_\infty < \infty$ with probability 1 as well).

To proceed, we will need to consider two cases, depending on whether $q = 2$ or $q > 2$. For the latter (which is more difficult), we will require the following variant of Hölder's inequality:

$$\left(\sum_{k=1}^{n} \alpha_k \beta_k\right)^r \leq \left(\sum_{k=1}^{n} \alpha_k^{\frac{\delta r}{r-1}}\right)^{r-1} \sum_{k=1}^{n} \alpha_k^{(1-\delta)r} \beta_k^r, \tag{B.13}$$

valid for all $r > 1$ and all $\delta \in (0, 1)$. Then, applying this inequality with $\alpha_k = \gamma_k^2$, $\beta_k = \|Z_k\|^2$, $r = q/2$ and $\delta = (q-2)/(2q)$, we obtain:

$$\left(\sum_{k=1}^{n} \gamma_k^2 \|Z_k\|^2\right)^{q/2} \leq \left(\sum_{k=1}^{n} \gamma_k\right)^{q/2-1} \sum_{k=1}^{n} \gamma_k^{1+q/2} \|Z_k\|^q = \tau_n^{q/2-1} Q_n, \tag{B.14}$$

and hence:

$$C_n = \frac{\sum_{k=1}^{n} \gamma_k^2 \|Z_k\|^2}{\tau_n} \leq \frac{\tau_n^{1-2/q} Q_n^{2/q}}{\tau_n} = \frac{Q_n^{2/q}}{\tau_n^{2/q}}. \tag{B.15}$$

Since $q > 2$ and Q_n converges (a.s.) to Q_∞, it follows that $\lim_{n\to\infty} C_n = 0$ with probability 1 (since $\lim_{n\to\infty} \tau_n = \infty$ by our assumptions for γ_n). Finally, if $q = 2$, we have $C_n = Q_n/\tau_n$ by definition, so we get $C_n \to 0$ (a.s.) directly.

Putting together all of the above, we get $A_n + B_n + C_n \to 0$ with probability 1, and hence, with probability 1 conditioned on Ω_0 (since $\mathbb{P}(\Omega_0) > 0$). This means that, for sufficiently large n, we have

$$f_{n+1} \leq f_1 - \tau_n(c/2) \tag{B.16}$$

which, together with the fact that $\lim_{n\to\infty} \tau_n = \infty$, implies that $\lim_{n\to\infty} f(X_n) = -\infty$. This contradicts Assumption 2 and completes our proof. ∎

We now move on to the deterministic elements of the proof of Theorem 1. To that end, let

$$a = \max_{x \in \mathcal{X}^*} f(x) \tag{B.17}$$

denote the maximum value of f over its critical set, and let

$$K_\varepsilon = L_{a+\varepsilon} = \{x \in \mathbb{R}^d : f(x) \leq a + \varepsilon\} \tag{B.18}$$

denote the $(a+\varepsilon)$-sublevel set of f. We then have the following "uniform decrease" estimate:

Lemma B.3. *Fix some $\varepsilon > 0$. Under Assumptions 1–3, there exists some $\tau \equiv \tau(\varepsilon)$ such that, for all $x \in \mathbb{R}^d$, we have (i) $f(\Phi_\tau(x)) \leq f(x) - \varepsilon$; or (ii) $\Phi_\tau(x) \in K_\varepsilon$.*

Proof. By Lemma B.2, there exists some positive constant $c > 0$ such that $\|\nabla f(x)\|^2 \geq c > 0$ for all $x \in \mathbb{R}^d \setminus K_\varepsilon$. Then, with $df/dt = -\|\nabla f(x(t))\|^2$, if we let $\tau_x = \inf\{t \geq 0 : \Phi_t(x) \in K_\varepsilon\}$, we get:

$$f(\Phi_t(x)) = f(x) - \int_0^t \|\nabla f(x(s))\|^2 \, ds \leq f(x) - ct \quad \text{for all } t \in [0, \tau_x]. \tag{B.19}$$

Accordingly, letting $\tau = \varepsilon/c$, we may consider the following two case:

(1) If $\tau_x \geq \tau$, applying (B.19) for $t = \tau$ yields $f(\Phi_\tau(x)) \leq f(x) - \varepsilon$.
(2) Otherwise, if $\tau_x < \tau$, we have $f(\Phi_\tau(x)) \leq f(\Phi_{\tau_x}(x)) \leq a + \varepsilon$, implying in particular that $\Phi_\tau(x) \in K_\varepsilon$.

Our claim then follows by combining the two cases above. ∎

Finally, we establish below the required comparison bound between an APT of (GD) and its solution trajectories:

Lemma B.4. *Fix some $\varepsilon, \delta > 0$. Then, with assumptions and notation as in Lemma B.3, there exists some $t_0 \equiv t_0(\delta, \varepsilon)$ such that, for all $t \geq t_0$ and all $h \in [0, \tau]$, we have:*

$$f(X(t+h)) \leq f(\Phi_h(X(t))) + G\delta + \tfrac{1}{2}L\delta^2. \tag{B.20}$$

Proof. By the definition of an APT, there exists some $t_0 \equiv t_0(\delta, \varepsilon)$ such that

$$\sup_{0 \leq h \leq \tau} \|X(t+h) - \Phi_h(X(t))\| \leq \delta \tag{B.21}$$

for all $t \geq t_0$. Hence, for all $t \geq t_0$ and all $h \in [0, \tau]$, we have

$$\begin{aligned}
f(X(t+h)) &= f(\Phi_h(X(t)) + X(t+h) - \Phi_h(X(t))) \\
&\leq f(\Phi_h(X(t))) + \langle \nabla f(\Phi_h(X(t))), X(t+h) - \Phi_h(X(t)) \rangle \\
&\quad + \frac{L}{2}\|X(t+h) - \Phi_h(X(t))\|^2 \\
&\leq f(\Phi_h(X(t))) + G\|X(t+h) - \Phi_h(X(t))\| + \frac{L}{2}\|X(t+h) - \Phi_h(X(t))\|^2 \\
&\leq f(\Phi_h(X(t))) + G\delta + \frac{L}{2}\delta^2,
\end{aligned} \tag{B.22}$$

as claimed. ∎

With all this in hand, we are finally in a position to formally prove Theorem 1.

Proof of Theorem 1. We will prove the stronger statement that, with probability 1, $X(t)$ converges to the sublevel set $L_a = \{x \in \mathbb{R}^d : f(x) \leq a\}$ with a defined as in (B.17). Since the sublevel sets of f are bounded, convergence to L_a suffices.

To prove this claim, fix some $\varepsilon > 0$ and let $X(t)$ be the affine interpolation of the sequence of iterates X_n generated by (SGD). Under the stated assumptions, Proposition 1 guarantees that $X(t)$ is an APT of (GD) with probability 1. Moreover, again with probability 1, Lemma B.1 guarantees the existence of some (possibly random) t_1 such that $X(t_1) \in K_{2\varepsilon}$. To streamline the analysis to come, we will condition our statements on the intersection of these two events (which still occurs with probability 1), and we will argue trajectory-wise.

Moving forward, Lemma B.3 guarantees the existence of some $\tau \equiv \tau(\varepsilon)$ such that $f(\Phi_\tau(x)) \leq f(x) - \varepsilon$ or $\Phi_\tau(x) \in K_\varepsilon$ for all $x \in \mathbb{R}^d$. Fixing this τ and taking $\delta > 0$ such that $G\delta + L\delta^2/2 < \varepsilon$, Lemma B.4 further implies that there exists some t_0 such that (B.20) holds for all $t \geq t_0$ and all $h \in [0, \tau]$. Note also that, without loss of generality, we can assume that $t_1 > t_0$; otherwise, if this is not the case, it suffices to wait for the first instance n such that $X_n \in K_{2\varepsilon}$ and $\tau_n \geq t_0$ (by Lemma B.1, this occurs with probability 1).

Combining all of the above, we have *(i)* $X(t_1) \in K_{2\varepsilon}$; and *(ii)* $f(X(t+h)) \leq f(\Phi_h(X(t))) + \varepsilon$ for all $t \geq t_1$ and all $h \in [0, \tau]$. Since $f(\Phi_t(x)) \leq f(x)$ for all $t \geq 0$, this further implies that

$$f(X(t+h)) \leq f(X(t)) + \varepsilon \tag{B.23}$$

for all $h \in [0, \tau]$. We thus get

$$f(X(t)) \leq f(X(t_1)) + \varepsilon \leq a + 3\varepsilon \tag{B.24}$$

for all $t \in [t_1, t_1 + \tau]$. Moreover, since $X(t_1) \in K_{2\varepsilon}$, Lemma B.3 also gives $\Phi_\tau(X(t_1)) \in K_\varepsilon$ because the two conditions of the lemma coincide if $x \in K_{2\varepsilon}$. As a result, we finally obtain

$$f(X(t_1 + \tau)) \leq f(\Phi_\tau(X(t_1))) + \varepsilon \leq a + \varepsilon + \varepsilon = a + 2\varepsilon, \tag{B.25}$$

i.e., $X(t_1 + \tau) \in K_{2\varepsilon}$.

From the above, we conclude that *(i)* $X(t) \in K_{3\varepsilon}$ for all $t \in [t_1, t_1 + \tau]$; and, in particular, *(ii)* $X(t_1 + \tau) \in K_{2\varepsilon}$. Proceeding inductively, we get $X(t) \in K_{3\varepsilon}$ for all $t \in [t_1 + (k-1)\tau, t_1 + k\tau]$,

$k = 1, 2, \dots$, i.e., $X(t) \in K_{3\varepsilon}$ for all $t \geq t_1$. Since $\varepsilon > 0$ is arbitrary, this means that $X(t)$ converges to $K_0 \equiv L_a$ as claimed. ∎

Appendix C. Avoidance analysis

As we stated in the main body of the paper, the proof of Theorem 3 will require two different threads of arguments: *a*) a series of probabilistic estimates to show that a certain class of stochastic processes avoids zero; and *b*) the construction of a suitable (average) Lyapunov function that grows exponentially along the unstable directions of a strict saddle manifold.

C.1. **Probabilistic estimates.** The probabilistic estimates that we will need date back to Pemantle [35] and concern a class of stochastic processes defined as follows: let Y_n, $n = 1, 2, \dots$, be a sequence of $\mathcal{F}_n$-measurable random variables, let $E_n = \sum_{k=1}^{n} Y_k$, and assume that

$$\mathbb{E}[E_{n+1}^2 - E_n^2 \,|\, \mathcal{F}_n] \geq C/n^{2p} \quad \text{for some } C > 0 \text{ and all } n = 1, 2, \dots \tag{C.1}$$

In the above, E_n will play the role of a "distance measure" from $\mathcal{S}$. Informally, the requirement (C.1) posits that E_n increases in "root mean square" by $\Theta(\gamma_n)$ where $\gamma_n \propto 1/n^p$ is the step-size of (SGD); constructing such a process will be the topic of the geometric constructions of the next section. For now, we state without proof a number of conditions guaranteeing that the process E_n cannot converge to 0:

Lemma C.1 ($0 < p \leq 1/2$; 4, Lemma 4.2)**.** *Suppose that* (C.1) *holds for some* $p \in (0, 1/2]$. *Then,* $\mathbb{P}(\lim_{n\to\infty} E_n = 0) = 0$.

Lemma C.2 ($1/2 < p \leq 1$: 36, Lemma 5.5)**.** *Suppose that* (C.1) *holds for some* $p \in (1/2, 1]$. *Assume further that there exist constants* $a, b > 0$ *such that, for all* $n = 1, 2, \dots$, *we have:*

(1) $|Y_n| \leq a/n^p$ *with probability* 1.

(2) $\mathbb{1}_{\{E_n > b/n^p\}} \mathbb{E}[Y_{n+1} \,|\, \mathcal{F}_n] \geq 0$ *with probability* 1.

Then, $\mathbb{P}(\lim_{n\to\infty} E_n = 0) = 0$.

A first version of Lemma C.2 was originally proven by Pemantle [35] for the special case $p = 1$ but the proof techniques are similar for all $1/2 < p \leq 1$; for a more general estimate (which we will not need here), see Benaïm [2, Lemma 9.6].

C.2. **Center manifold theory and geometric constructions.** We now proceed with the construction of a suitable Lyapunov function that will allow us to apply Lemmas C.1 and C.2. This construction follows Benaïm and Hirsch [3] and Benaïm [2] and relies crucially on center manifold theory; for a general introduction to the topic, we refer the reader to Lee [25], Shub [41], and Robinson [40].

To begin, let $\mathcal{S}$ be a strict saddle manifold as defined in Section 4.2. Then, for all $x^* \in \mathcal{S}$, we define the *center*, *stable* and *unstable* directions of x^* to be respectively the eigenspaces of $H(x^*) = \nabla^2 f(x^*)$ corresponding to zero, positive and negative eigenvalues thereof, i.e.,

$$\mathcal{E}^c_{x^*} = \{v \in \mathbb{R}^d : H(x^*)v = 0\} = \ker H(x^*), \qquad \text{[central directions]} \tag{C.2a}$$

$$\mathcal{E}^s_{x^*} = \{v \in \mathbb{R}^d : H(x^*)v = \lambda v \text{ for some } \lambda > 0\} \qquad \text{[stable directions]} \tag{C.2b}$$

$$\mathcal{E}^u_{x^*} = \{v \in \mathbb{R}^d : H(x^*)v = \lambda v \text{ for some } \lambda < 0\} \qquad \text{[unstable directions]} \tag{C.2c}$$

The reason for this terminology is that $H(x) = \operatorname{Jac}(\nabla f(x))$, so these subspaces correspond to directions that are respectively neutral (or *slow*), attracting, and repelling under (GD). More precisely, by the center manifold theorem [40, 41], there exists a neighborhood $\mathcal{U}$ of $\mathcal{S}$ and a submanifold $\mathcal{M}$ of $\mathbb{R}^d$, called the *center stable manifold* of $\mathcal{S}$, and satisfying the

following: *a*) $\mathcal{M}$ is *locally invariant* under Φ, i.e., there exists some positive $t_0 > 0$ such that $\Phi_t(\mathcal{U} \cup \mathcal{M}) \subseteq \mathcal{M}$ for all $t \geq t_0$; and *b*) $\mathbb{R}^d = T_{x^*}\mathcal{M} \oplus \mathcal{E}^u_{x^*}$ for all $x^* \in \mathcal{S}$, where $T_{x^*}\mathcal{M}$ denotes the tangent space to $\mathcal{M}$ at x^*. In view of this: *a*) perturbations along central directions are tangent to $\mathcal{M}$ and are thus expected to evolve "along" $\mathcal{M}$ under (GD); *b*) stable perturbations along $\mathcal{E}^s_{x^*}$ will converge along $\mathcal{M}$ to $\mathcal{S}$ under (GD); and *c*) unstable perturbations along $\mathcal{E}^u_{x^*}$ are transverse to $\mathcal{M}$ and may escape.

A key property of $\mathcal{M}$ is that any globally bounded orbit of (GD) which is contained in a sufficiently small neighborhood of $x^* \in \mathcal{S}$ must be entirely contained in $\mathcal{M}$ [41]. Moreover, by the non-minimality assumption for $\mathcal{S}$, it follows that $d_u \equiv \dim \mathcal{E}^u_{x^*} \geq 1$, so the dimension of $\mathcal{M}$ is at most $d-1$. This suggests that perturbations along any direction that is transverse to $\mathcal{M}$ will be repelled under (GD); we make this statement precise in the lemma below.

Lemma C.3. *Let $\Psi_t(x) = \nabla_x \Phi_t(x^*)$ denote the infinitesimal generator of the flow of (GD). Then:*

(1) The unstable subspaces $\mathcal{E}^u_{x^*}$ are invariant under (GD); specifically, $\Psi_t(x^*)\mathcal{E}^u_{x^*} = \mathcal{E}^u_{x^*}$ for all $t \geq 0$ and all $x^* \in \mathcal{S}$.

(2) There exists a positive constant $c > 0$ such that, for all $x^* \in \mathcal{S}$, $w \in \mathcal{E}^u_{x^*}$ and $t \geq 0$, we have

$$\|\Psi_t(x^*)w\| \geq e^{ct}\|w\|. \tag{C.3}$$

Remark 1. In the above (and what follows), we write AW for the image of a vector space W under a linear operator A. Specifically, if $A\colon V \to V'$ is a linear operator between two vector spaces V and V', and if $W \leq V$ is a subpace of V, we let $AW \equiv \mathrm{im}_A(W) = \{Aw : w \in W\}$. We also treat linear operators and matrices interchangeably.

Remark 2. The proof of Lemma C.3 (and, in fact, all of our analysis in this section) does not require the uniformity condition $\min \lambda_+(H(x^*)) \geq c_+$ for the Hessian's positive eigenvalues (if such eigenvalues exist). We only make it to simplify the presentation and avoid cases where the dimension of $\mathcal{E}^s_{x^*}$ may change; in that case, it would be sufficient to work with a subset of $\mathcal{S}$ over which this does not occur.

In words, Lemma C.3 states that *a*) the unstable directions along $\mathcal{S}$ are consistent with the flow of (GD); and *b*) perturbations along unstable directions are repelled from $\mathcal{S}$ at a geometric rate. The proof is as follows:

Proof of Lemma C.3. Recall first that, for all $t \geq 0$ and all $x \in \mathbb{R}^d$, we have $\Psi_t(x) = \nabla_x \Phi_t(x) = \exp(t\,\mathrm{Jac}(-\nabla f(x))) = \exp(-tH(x))$. Therefore, since $\mathcal{S}$ consists entirely of stationary points of (GD), we readily get

$$\begin{aligned}
\Psi_t(x^*)\mathcal{E}^u_{x^*} &= e^{-tH(x^*)}\mathcal{E}^u_{x^*} \\
&= \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} H(x^*)^k \mathcal{E}^u_{x^*} = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} \mathcal{E}^u_{x^*} && \{\text{because } H(x^*)\mathcal{E}^u_{x^*} = \mathcal{E}^u_{x^*}\} \\
&= e^{-t}\mathcal{E}^u_{x^*} = \mathcal{E}^u_{x^*}, && \{\text{C.4}\}
\end{aligned}$$

so our first claim follows.

For our second claim, let $\{u_i : i = 1, \dots, d\}$ be an orthnormal set of eigenvectors of $H(x^*)$.[4] Moreover, let $\lambda_i \equiv \lambda_i(x^*) < 0$ be the eigenvalue of $H(x^*)$ corresponding to u_i, and assume without loss of generality that the indexing labels $i = 1, \dots, d$ have been chosen in ascending eigenvalue order, i.e., $\lambda_1 \leq \cdots \leq \lambda_d$. It then follows that $\{u_i : i = 1, \dots, d_u \equiv \dim \mathcal{E}^u_{x^*}\}$

[4]That such a set exists follows from the fact that $H(x^*)$ is symmetric.

is an orthonormal basis of $\mathcal{E}^u_{x^*}$ consisting entirely of eigenvectors of $H(x^*)$. Thus, writing $w = \sum_i w_i u_i$ for a given vector $w \in \mathcal{E}^u_{x^*}$, we have:

$$\Psi_t(x^*)w = e^{-tH(x^*)}w = \sum_{i=1}^{d_u} w_i e^{-tH(x^*)}u_i = \sum_{i=1}^{d_u} w_i e^{-t\lambda_i}u_i, \tag{C.5}$$

where, in the last step, we used the fact that u_i is an eigenvector of $H(x^*)$ with eigenvalue λ_i (and hence, also of $e^{-tH(x^*)}$ with eigenvalue $e^{-t\lambda_i}$). Therefore, by orthonormality, we obtain:

$$\|\Psi_t(x^*)w\|^2 = \sum_{i=1}^{d_u} e^{-2t\lambda_i} w_i^2 \geq e^{2c_- t}\|w\|^2, \tag{C.6}$$

where $c_- > 0$ is defined in Section 4.3. ∎

To proceed, we will need to define a suitable "projector" from neighborhoods of $\mathcal{S}$ to $\mathcal{M}$. To carry out this construction, consider the vector bundle

$$\mathcal{E}^u_{\mathcal{S}} \equiv \{(x^*, w) : x^* \in \mathcal{S}, w \in \mathcal{E}^u_{x^*}\} \tag{C.7}$$

of the unstable directions of (GD) over $\mathcal{S}$. Since each $\mathcal{E}^u_{x^*}$ is a subspace of $\mathbb{R}^d$, we can view $\mathcal{E}^u_{\mathcal{S}}$ as a map from $\mathcal{S}$ to the Grassmannian $\mathbf{Gr}(d_u, d)$ of d_u-dimensional spaces of $\mathbb{R}^d$. By the Whitney embedding theorem [25], $\mathbf{Gr}(d_u, d)$ can be embedded as a $d_u \times (d - d_u)$-dimensional submanifold of $\mathbb{R}^{2d_u(d-d_u)}$; as such, $\mathcal{E}^u_{\mathcal{S}}$ may be seen as a map $\mathcal{S} \to \mathbb{R}^{2d_u(d-d_u)}$ with values in $\mathbf{Gr}(d_u, d) \hookrightarrow \mathbb{R}^{2d_u(d-d_u)}$. Since $\mathcal{S}$ is closed (as a connected component of $\mathcal{X}^*$), the Tietze extension theorem [1] further implies that this map admits a continuous extension $\pi\colon \mathbb{R}^d \to \mathbb{R}^{2d_u(d-d_u)}$ to all of $\mathbb{R}^d$. By mollifying this map with an approximate identity supported on $\mathcal{S}$, we can further assume that this extension is smooth in a neighborhood of $\mathcal{S}$. Moreover, by standard results in differential topology [16, Chap. 4], there exists a smooth retraction of a neighborhood of $\mathbf{Gr}(d_u, d)$ onto $\mathbf{Gr}(d_u, d)$ in $\mathbb{R}^{2d_u(d-d_u)}$. Hence, by composing π with this retraction, we finally obtain a smooth vector bundle

$$\mathcal{E}^u_{\mathcal{U}} \equiv \{(x, w) : x \in \mathcal{U}, w \in \mathcal{E}^u_x\} \tag{C.8}$$

which, by construction, coincides with $\mathcal{E}^u_{\mathcal{S}}$ over $\mathcal{S}$ (explaining the slight abuse of notation).

By taking a smaller neighborhood if necessary, we may assume that $\mathcal{U}$ is compact and coincides with the one in the definition of $\mathcal{M}$, i.e., $\Phi_t(\mathcal{U} \cap \mathcal{M}) \subseteq \mathcal{M}$ for small enough t. We may now construct a "projector" from a (potentially smaller) neighborhood of $\mathcal{M}$ to $\mathcal{M}$ as follows: First, consider the simple vector addition mapping $Q\colon \mathcal{E}^u_{\mathcal{U}} \to \mathbb{R}^d \equiv \mathbb{R}^d$ sending $(x, w) \in \mathcal{E}^u_{\mathcal{U}} \mapsto x + w \in \mathbb{R}^d$. Clearly, the zero section $(x, 0)$ of $\mathcal{E}^u_{\mathcal{U}}$ is mapped diffeomorphically to $\mathcal{U}$ so, by the inverse function theorem [25], it follows that Q is a local diffeomorphism. Thus, letting $\mathcal{U}'$ be a neighborhood of $\mathcal{M}$ over which Q is a diffeomorphism, and letting $\mathcal{U}_0 = Q(\mathcal{U}')$, we get a map $\Pi\colon \mathcal{U}_0 \to \mathcal{M}$ such that

$$\Pi(y) = x \iff Q(x, w) = x + w = y \tag{C.9}$$

The reason for this sophisticated construction (as opposed to e.g., taking a Euclidean projection from $\mathcal{U}_0$ to $\mathcal{M}$) is that Π respects the unstable directions of $\mathcal{S}$ under (GD). More precisely, we have:

Lemma C.4. *For $x \in \mathcal{U}$, let $P_x\colon T_x\mathcal{M} \oplus \mathcal{E}^u_x \to T_x\mathcal{M}$ denote the projection*

$$\begin{array}{ccc} z + w & \mapsto P_x(z + w) = & z \\ \cap & & \cap \\ T_x\mathcal{M}\oplus\mathcal{E}^u_x & & T_x\mathcal{M} \end{array} \tag{C.10}$$

Then, for all $x \in \mathcal{U}_0 \cap \mathcal{M}$, we have $\mathrm{Jac}(\Pi(x)) = P_x$.

Proof. Let $y(t)$, $t \in (-1,1)$ be a smooth curve on $\mathcal{U}_0$ going through $x = y(0) \in \mathcal{M}$ at time $t = 0$, and let $x(t) = \Pi(y(t))$ so $y(t) = x(t) + \psi(t)$ for some smooth $\psi(t) \in \mathcal{E}^u_{x(t)}$. By differentiating, we get $\dot{y}(0) = \dot{x}(0) + \dot{\psi}(0)$; since $x(t) \in \mathcal{M}$ and $\psi(t) \in \mathcal{E}^u_{x(t)}$ for all t, we readily get $\dot{x}(0) \in T_{x(0)}\mathcal{M}$ and $\dot{\psi}(0) \in \mathcal{E}^u_{x(0)}$. Letting $z = \dot{x}(0)$ and $w = \dot{\psi}(0)$, this shows that the pushforward of $\dot{y}(0) = z + w$ under Π at x is $\mathrm{D}\Pi_x(z+w) \equiv \mathrm{Jac}(\Pi(x))(z+w) = z = P_x(z+w)$. With $y(t)$ arbitrary, our claim follows. ∎

We are finally in a position to define a "potential function" on $\mathcal{U}_0$ as

$$V(y) = \|\Pi(y) - y\| \tag{C.11}$$

i.e., as the (normed) distance of $y \in \mathcal{U}_0$ from its vector projection $\Pi(y)$ on $\mathcal{M}$ along the unstable directions of (GD). By construction, we have

$$V(y) \geq 0 \quad \text{with equality if and only if } y \in \mathcal{M} \cap \mathcal{U}_0. \tag{C.12}$$

Coupling (C.12) with Lemmas C.3 and C.4, we see that f satisfies the requirements of Benaïm [2, Proposition 9.5], which, when adapted to our setting, provides the following:

Proposition C.1 (2)**.** *There exists a compact neighborhood $\mathcal{U}_\mathcal{S}$ of $\mathcal{S}$, a positive constant $\beta > 0$, and a time horizon $\tau > 0$ such that the energy function*

$$E(x) = \int_0^\tau V(\Phi_{-t}(x))\, dt \qquad x \in \mathcal{U}_\mathcal{S}, \tag{C.13}$$

enjoys the following properties:

(1) For all $x \in \mathcal{U}_\mathcal{S}$, E has a Lipschitz continuous and positively homogeneous right derivative $\nabla^+ E(x)$;[5] in addition, E is continuously differentiable on $\mathcal{U}_\mathcal{S} \setminus \mathcal{M}$.

(2) For all $x \in \mathcal{U}_\mathcal{S}$, we have

$$\nabla^+ E(x)[\nabla f(x)] \leq -\beta E(x). \tag{C.14}$$

In particular, for all $x \in \mathcal{U}_\mathcal{S} \setminus \mathcal{M}$, we have:

$$\langle \nabla E(x), \nabla f(x) \rangle \leq -\beta E(x) \tag{C.15}$$

(3) There exists a constant $\alpha > 0$ such that, for all $x \in \mathcal{U}_\mathcal{S}$ and all sufficiently small $v \in \mathbb{R}^d$, we have

$$E(x+v) \geq E(x) + \nabla^+ E(x)[v] - \frac{\alpha}{2}\|v\|^2. \tag{C.16}$$

(4) There exists a constant $\beta > 0$ such that, for all $v \in \mathbb{R}^d$, we have:

$$\|\nabla E(x)\| \geq \beta \qquad \text{for all } x \in \mathcal{U}_\mathcal{S} \setminus \mathcal{M}, \tag{C.17a}$$

and

$$\nabla^+ E(x)[v] \geq \beta \|P_x(v) - v\| \qquad \text{for all } x \in \mathcal{U}_\mathcal{S} \cap \mathcal{M}. \tag{C.17b}$$

Proposition C.1 follows from Benaïm [2, Proposition 9.5], so we do not present a proof. More important for our purposes are the following immediate consequences thereof:

(1) By (C.15), the energy $E(x(t))$ of a solution orbit $x(t)$ of (GD) will grow at a (locally) geometric rate if $x(t)$ doesn't already lie in the center stable manifold $\mathcal{M}$ of $\mathcal{S}$. This means that asymptotic pseudotrajectories of (GD) that do not lie on $\mathcal{M}$ for arbitrarily long windows of time will also escape $\mathcal{M}$ (and hence $\mathcal{S}$).

[5] Recall here that a function ϕ has a right derivative when the limit $\nabla^+\phi(x)[v] \equiv \lim_{t\to 0^+}[\phi(x+tv) - \phi(x)]/t$ exists for all $v \in \mathbb{R}^d$.

(2) The bound (C.16) provides the basis for a discrete-time version of the above argument: as long as X_n is sufficiently close to $\mathcal{S}$, the energy before and after a stochastic gradient step will be linked as

$$E(X_{n+1}) \geq E(X_n) + \beta\gamma_n E(X_n) - \gamma_n\psi_n - \frac{\alpha\gamma_n^2}{2}\|V_n\|^2, \tag{C.18}$$

where ψ_n is an additive noise term which is non-antagonistic in expectation. This means that, on average, the iterates $E_n \equiv E(X_n)$ will grow at a (locally) geometric rate, so X_n cannot remain in the vicinity of $\mathcal{S}$ for very long periods.

To make the above precise, we will need to invoke the probabilistic estimates stated in Appendix C.1. We do so in the following section.

C.3. **Avoidance of saddle-point manifolds.** For convenience, we begin by restating our main avoidance result below:

Theorem 3. *Suppose that* (SGD) *is run with a variable step-size sequence of the form* $\gamma_n \propto 1/n^p$ *for some* $p \in (0,1]$. *If Assumptions 1–5 hold (with* $q = \infty$ *for Assumption 4), and* $\mathcal{S}$ *is a strict saddle manifold of* f, *we have* $\mathbb{P}(X_n \to \mathcal{S} \text{ as } n \to \infty) = 0$.

Proof. Our proof follows the arguments of Benaïm and Hirsch [3], suitably adapted to our setting. To begin, let $\mathcal{U}_\mathcal{S}$ be the compact neighborhood of $\mathcal{S}$ identified in Proposition C.1 and assume without loss of generality that $X_1 \in \mathcal{U}_\mathcal{S}$. We may then define the exit time from $\mathcal{U}_\mathcal{S}$ as

$$T_\mathcal{S} = \inf\{n \geq 1 : n \notin \mathcal{U}_\mathcal{S}\}. \tag{C.19}$$

We will prove our claim by showing that $T_\mathcal{S} < \infty$ with probability 1.

To that end, consider the process

$$Y_{n+1} = \begin{cases} E(X_{n+1}) - E(X_n) & \text{if } n \leq T_\mathcal{S}, \\ \gamma_n & \text{otherwise,} \end{cases} \tag{C.20}$$

with $E(X_0) \equiv 0$ by convention. Heuristically, Y_n measures the change in energy of X_n as long as it remains in $\mathcal{U}_\mathcal{S}$; subsequently, for book-keeping purposes, it is incremented by a token amount of γ_n per iteration once X_n exits $\mathcal{U}_\mathcal{S}$. To make this idea more formal, let

$$E_n = \sum_{k=1}^{n} Y_k \tag{C.21}$$

so $E_n = E(X_n)$ if $X_n \in \mathcal{U}_\mathcal{S}$ while $E_n = \Theta(\tau_n)$ after X_n exits $\mathcal{U}_\mathcal{S}$.

Assume now that $X_n \in \mathcal{U}_\mathcal{S}$ for all n (i.e., $T_\mathcal{S} = \infty$). By Theorem 2, every limit point $\hat{x}$ of X_n must be contained in $\mathcal{S}$, so, by (C.12), we must have $\lim_{n\to\infty} E_n = 0$. Hence, to establish our claim, it suffices to show that $\mathbb{P}(E_n \to 0) = 0$. We will do this by showing that Y_n defined as in (C.20) satisfies the requirements of Lemmas C.1 and C.2.

We begin with the conditions required by Lemma C.2 for the case $1/2 < p \leq 1$:

(1) For the condition $|Y_n| = \mathcal{O}(1/n^p)$ of Lemma C.2, the claim is tautological if $n > T_\mathcal{S}$. Otherwise, if $n \leq T_\mathcal{S}$, note that

$$\|X_{n+1} - X_n\| = \gamma_n\|V_n\| \leq \gamma_n[\|\nabla f(X_n)\| + \|Z_n\|] \leq \gamma_n(G + \sigma) \tag{C.22}$$

by Assumptions 1 and 4 (recall here that we are taking $q = \infty$ in Assumption 4). Since $X_n \in \mathcal{U}_\mathcal{S}$ as long as $n \leq T_\mathcal{S}$, and given that E is continuously differentiable on $\mathcal{U}_\mathcal{S}$ (and hence Lipschitz continuous therein), we also have:

$$|E(X_{n+1}) - E(X_n)| = \mathcal{O}(\|X_{n+1} - X_n\|) = \mathcal{O}(\gamma_n) = \mathcal{O}(1/n^p), \tag{C.23}$$

as claimed.

(2) For the condition $\mathbb{1}_{\{E_n>b/n^p\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]\geq 0$, note first that if $n>T_{\mathcal{S}}$, then $Y_n=\gamma_n$, so

$$\mathbb{1}_{\{n>T_{\mathcal{S}}\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]\geq \mathbb{1}_{\{n>T_{\mathcal{S}}\}}\,\gamma_n>0. \tag{C.24}$$

Otherwise, if $n\leq T_{\mathcal{S}}$, we have $X_n\in\mathcal{U}_{\mathcal{S}}$, so Proposition C.1 yields

$$Y_{n+1}=E(X_{n+1})-E(X_n)\geq \beta\gamma_n E(X_n)-\gamma_n\psi_n-2\alpha\gamma_n^2(G^2+\sigma^2), \tag{C.25}$$

where we set

$$\psi_n=\nabla^+E(X_n)[Z_n] \tag{C.26}$$

and used the estimate $\|V_n\|^2=\|\nabla f(X_n)+Z_n\|^2\leq[\|\nabla f(X_n)\|^2+\|Z_n\|^2]\leq 2(G^2+\sigma^2)$ (compare also with (C.18) and the surrounding discussion). By the conditional Jensen inequality and the definition of $\nabla^+E(x)$, we have

$$\mathbb{E}[\psi_n\,|\,\mathcal{F}_n]=\mathbb{E}[\nabla^+E(X_n)[Z_n]\,|\,\mathcal{F}_n]\geq\nabla^+E(X_n)[\mathbb{E}[Z_n\,|\,\mathcal{F}_n]]=0. \tag{C.27}$$

which, in turn, implies that

$$\mathbb{1}_{\{n\leq T_{\mathcal{S}}\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]\geq\gamma_n\,\mathbb{1}_{\{n\leq T_{\mathcal{S}}\}}\big[\beta E(X_n)-2\alpha(G^2+\sigma^2)\gamma_n\big]. \tag{C.28}$$

Hence, taking $b>0$ such that $b\beta/n^p=2\alpha(G^2+\sigma^2)\gamma_n$, and recalling that $E_n=E(X_n)$ if $n\leq T_{\mathcal{S}}$, we get

$$\begin{aligned}\mathbb{1}_{\{E_n>b/n^p\}}\,\mathbb{1}_{\{n\leq T_{\mathcal{S}}\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]&\geq\mathbb{1}_{\{E_n>b/n^p\wedge n\leq T_{\mathcal{S}}\}}\big[\beta E_n-2\alpha(G^2+\sigma^2)\gamma_n\big]\\&\geq\mathbb{1}_{\{E_n>b/n^p\wedge n\leq T_{\mathcal{S}}\}}\big[b\beta/n^p-2\alpha(G^2+\sigma^2)\gamma_n\big]\\&\geq 0.\end{aligned} \tag{C.29}$$

Thus, combining the above, we conclude that the specific conditions required to apply Lemma C.2 are satisfied.

We are left to establish the general condition (C.1) which is required to apply both Lemmas C.1 and C.2; the proof is the same for all $p\in(0,1]$, so we no longer assume $1/2<p\leq 1$ below. To begin, note that

$$\begin{aligned}\mathbb{E}[E_{n+1}^2-E_n^2\,|\,\mathcal{F}_n]&=\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]+2E_n\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]\\&=\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]+2E_n\,\mathbb{1}_{\{E_n\leq b/n^p\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]\\&\qquad+2E_n\,\mathbb{1}_{\{E_n>b/n^p\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]\\&\geq\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]+2E_n\,\mathbb{1}_{\{E_n\leq b/n^p\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n],\end{aligned} \tag{C.30}$$

where, in the last line, we used the inequalities proved in the previous paragraph, namely (C.24) and (C.29). To proceed, recall that $Y_n=\gamma_n>0$ if $n>T_{\mathcal{S}}$, so, by (C.28) we get

$$\begin{aligned}\mathbb{E}[E_{n+1}^2-E_n^2\,|\,\mathcal{F}_n]&\geq\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]\\&\qquad+2E_n\,\mathbb{1}_{\{E_n\leq b/n^p\}}\,\mathbb{1}_{\{n\leq T_{\mathcal{S}}\}}\,\mathbb{E}[Y_{n+1}\,|\,\mathcal{F}_n]\\&\geq\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]\\&\qquad+2\gamma_nE_n\,\mathbb{1}_{\{E_n\leq b/n^p\}}\,\mathbb{1}_{\{n\leq T_{\mathcal{S}}\}}\big[\beta E_n-2\alpha(G^2+\sigma^2)\gamma_n\big]\\&\geq\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]-2\gamma_n\cdot b/n^p\cdot 2\alpha(G^2+\sigma^2)\gamma_n\\&=\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]-4\alpha^2\beta^{-1}(G^2+\sigma^2)^2\gamma_n^3.\end{aligned} \tag{C.31}$$

In view of the above, to establish (C.1), it suffices to show that $\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]\geq B\gamma_n^2$ for some $B>0$ and sufficiently large n. In this regard, Jensen's inequality gives

$$\mathbb{E}[Y_{n+1}^2\,|\,\mathcal{F}_n]\geq\mathbb{E}[Y_{n+1}^+\,|\,\mathcal{F}_n]^2 \tag{C.32}$$

so it suffices to show that $\mathbb{E}[Y_{n+1}^+ \,|\, \mathcal{F}_n] = \Omega(\gamma_n)$. This is trivial if $n > T_{\mathcal{S}}$, so we are left to treat the case $n \le T_{\mathcal{S}}$. For this case, (C.25) gives

$$\mathbb{1}_{\{n\le T_{\mathcal{S}}\}}\,\mathbb{E}[Y_{n+1}^+ \,|\, \mathcal{F}_n] \ge \mathbb{1}_{\{n\le T_{\mathcal{S}}\}}\,\gamma_n\,\mathbb{E}[\psi_n^- \,|\, \mathcal{F}_n] - 2\,\mathbb{1}_{\{n\le 2T_{\mathcal{S}}\}}\,\alpha(G^2+\sigma^2)\gamma_n^2 \tag{C.33}$$

meaning that we need to focus on the expectation $\mathbb{E}[\psi_n^- \,|\, \mathcal{F}_n]$.

We consider two further cases (this is where Assumption 4 kicks in and plays a crucial role). First, if $X_n \notin \mathcal{M}$, Proposition C.1 and Assumption 5 applied to $v = -\nabla E(X_n)/\|\nabla E(X_n)\|$ give

$$\begin{aligned}\mathbb{1}_{\{n\le T_{\mathcal{S}}\wedge X_n\notin\mathcal{M}\}}\,\mathbb{E}[\psi_n^- \,|\, \mathcal{F}_n] &= \mathbb{1}_{\{n\le T_{\mathcal{S}}\wedge X_n\notin\mathcal{M}\}}\,\mathbb{E}[\langle -\nabla E(X_n), Z_n\rangle^+ \,|\, \mathcal{F}_n]\\ &\ge \mathbb{1}_{\{n\le T_{\mathcal{S}}\wedge X_n\notin\mathcal{M}\}}\cdot c\|\nabla E(X_n)\|\\ &\ge \beta c\,\mathbb{1}_{\{n\le T_{\mathcal{S}}\wedge X_n\notin\mathcal{M}\}}\,.\end{aligned} \tag{C.34}$$

Otherwise, if $X_n \in \mathcal{M}$ (which, heuristically, should only happen with probability 0), choose a unit normal vector u_n such that

$$\langle u_n, z\rangle = 0 \quad \text{for all } z \in T_{X_n}\mathcal{M}. \tag{C.35}$$

Since the projector P_{X_n} defined in (C.10) takes values in $T_{X_n}\mathcal{M}$, we will have $\langle u_n, P_{X_n}(Z_n)\rangle = 0$, and hence:

$$\langle u_n, Z_n\rangle = \langle u_n, Z_n - P_{X_n}(Z_n)\rangle. \tag{C.36}$$

Therefore, by Proposition C.1, we get the chain of inequalities:

$$\begin{aligned}\mathbb{E}[[\nabla^+E(X_n)[Z_n]]^- \,|\, \mathcal{F}_n] &\ge \beta\,\mathbb{E}[\|P_{X_n}(Z_n) - Z_n\| \,|\, \mathcal{F}_n] && \{\text{by Proposition C.1}\}\\ &\ge \beta\,\mathbb{E}[\langle u_n, Z_n - P_{X_n}(Z_n)\rangle^+ \,|\, \mathcal{F}_n] && \{\text{by Cauchy–Schwarz}\}\\ &= \beta\,\mathbb{E}[\langle u_n, Z_n\rangle^+ \,|\, \mathcal{F}_n] && \{\text{by (C.36)}\}\\ &\ge \beta c && \{\text{by Assumption 5}\}\end{aligned}$$

valid on the event $\{n \le T_{\mathcal{S}} \wedge X_n \in \mathcal{M}\}$.

Putting together all of the above, we finally get

$$\mathbb{1}_{\{n\le T_{\mathcal{S}}\}}\,\mathbb{E}[\psi_n^- \,|\, \mathcal{F}_n] \ge \mathbb{1}_{\{n\le T_{\mathcal{S}}\}}\,\beta c \tag{C.37}$$

and hence, by (C.33):

$$\mathbb{E}[Y_{n+1}^+ \,|\, \mathcal{F}_n] \ge \beta c\gamma_n - 2\alpha(G^2+\sigma^2)\gamma_n^2 = \Omega(\gamma_n) \tag{C.38}$$

on the event $\{n \le T_{\mathcal{S}}\}$. This completes our proof. ∎

Appendix D. Rates of convergence

Our aim in this appendix is to establish the rate of convergence of (SGD) to local minima that are regular in the sense of Hurwicz, i.e., $H(x^*) \succ 0$. For convenience, we restate the relevant result below:

Theorem 4. *Fix some tolerance level $\delta > 0$, let x^* be a regular minimizer of f, and suppose that Assumption 4 holds. Assume further that (SGD) is run with a step-size schedule of the form $\gamma_n = \gamma/(n+m)^p$ for some $p \in (2/(q+2), 1]$ and large enough $m, \gamma > 0$. Then:*

(1) There exist neighborhoods $\mathcal{U}$ and $\mathcal{U}_1$ of x^* such that, if $X_1 \in \mathcal{U}_1$, the event

$$\Omega_{\mathcal{U}} = \{X_n \in \mathcal{U} \text{ for all } n = 1, 2, \dots\} \tag{10}$$

occurs with probability at least $1-\delta$.

(2) Conditioned on $\Omega_{\mathcal{U}}$, we have

$$\mathbb{E}[\|X_n - x^*\|^2 \,|\, \Omega_{\mathcal{U}}] = \mathcal{O}(1/n^p). \tag{11}$$

Auxiliary results. The proof of Theorem 4 requires several ancillary results, which we state and prove below. The first is a lemma on numerical sequences, usually attributed to Chung [9]:

Lemma D.1 (9, Lemma 1). *Let a_n, $n = 1, 2, \dots$, be a non-negative sequence such that*

$$a_{n+1} \leq \left[1 - \frac{P}{(n+m)^p}\right] a_n + \frac{R}{(n+m)^{p+r}} \tag{D.1}$$

where $p \in (0, 1]$, $r > 0$ and $P, R > 0$. Then:

(1) If $p < 1$, we have

$$a_n \leq \frac{R}{P}\frac{1}{n^r} + o\left(\frac{1}{n^r}\right). \tag{D.2a}$$

(2) If instead $p = 1$ and $P > r$, we have

$$a_n \leq \frac{R}{P - r}\frac{1}{n} + o\left(\frac{1}{n}\right). \tag{D.2b}$$

The next ingredient of the proof of Theorem 4 provides a handle on the local behavior of f near a regular minimizer:

Lemma D.2. *Let x^* be a regular minimum of f. Then, there exists a convex compact neighborhood $\mathcal{K}$ of x^* and constants $\alpha, \beta > 0$ (possibly depending on $\mathcal{K}$) such that*

$$\alpha \|x - x^*\|^2 \leq \langle \nabla f(x), x - x^* \rangle \leq \beta \|x - x^*\|^2 \quad \text{for all } x \in \mathcal{K}. \tag{D.3}$$

Proof. Let $\mathcal{K}$ be a sufficiently small convex compact neighborhood of x^* such that $H(x) \succ 0$ for all $x \in \mathcal{K}$ (that such a neighborhood exists is a consequence of the regularity of x^* and the smoothness of f). Then, by compactness, there exist constants α, β such that $\alpha I \preccurlyeq H(x) \preccurlyeq \beta I$ for all $x \in \mathcal{K}$. Moreover, for all $x \in \mathcal{K}$, we have

$$\nabla f(x) = (x - x^*)^\top \int_0^1 H(x^* + t(x - x^*))\, dt, \tag{D.4}$$

where we used the fact that $\nabla f(x^*) = 0$ (since x^* is a minimizer of f). Hence, multiplying both sides by $x - x^*$, the mean value theorem for integrals yields:

$$\begin{aligned} \langle \nabla f(x), x - x^* \rangle &= \int_0^1 (x - x^*)^\top H(x^* + t(x - x^*))(x - x^*)\, dt \\ &= (x - x^*)^\top H(x')(x - x^*) \end{aligned} \tag{D.5}$$

for some $x' \in [x^*, x]$. Since $\alpha I \preccurlyeq H(x') \preccurlyeq \beta I$, our claim follows. ∎

Thanks to Lemma D.2, we obtain the following recursive estimate for (SGD):

Proposition D.1. *Let x^* be a regular minimum of f and let $\mathcal{K}$ and α be as in Lemma D.2. Assume moreover that $X_n \in \mathcal{K}$ for some $n \geq 1$ and let*

$$D_n = \frac{1}{2}\|X_n - x^*\|^2. \tag{D.6}$$

We then have:

$$D_{n+1} \leq (1 - 2\alpha\gamma_n) D_n + \gamma_n \xi_n + \tfrac{1}{2}\gamma_n^2 \|V_n\|^2, \tag{D.7}$$

where $\xi_n = -\langle Z_n, X_n - x^* \rangle$ is a martingale difference sequence.

Proof. Recall first that $X_{n+1} = X_n - \gamma_n(\nabla f(X_n) + Z_n)$ where Z_n is the gradient error at X_n. Then, by the definition of D_n, we have:

$$\begin{aligned} D_{n+1} = \tfrac{1}{2}\|X_{n+1} - x^*\|^2 &= \tfrac{1}{2}\|X_n - x^* - \gamma_n V_n\|^2 \\ &= \tfrac{1}{2}\|X_n - x^*\|^2 - \gamma_n\langle V_n, X_n - x^*\rangle + \tfrac{1}{2}\gamma_n^2\|V_n\|^2 \\ &= D_n - \gamma_n\langle \nabla f(X_n), X_n - x^*\rangle - \gamma_n\langle Z_n, X_n - x^*\rangle + \tfrac{1}{2}\gamma_n^2\|V_n\|^2 \\ &\leq D_n - \alpha\gamma_n\|X_n - x^*\|^2 + \gamma_n\xi_n + \tfrac{1}{2}\gamma_n^2\|V_n\|^2 \\ &= (1 - 2\alpha\gamma_n)D_n + \gamma_n\xi_n + \tfrac{1}{2}\gamma_n^2\|V_n\|^2 \end{aligned} \tag{D.8}$$

where the second-to-last line follows from Lemma D.2. Since $\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] = \langle \mathbb{E}[Z_n \,|\, \mathcal{F}_n], X_n - x^*\rangle = 0$, our claim follows (recall here that, by definition, Z_n is not $\mathcal{F}_n$-measurable but X_n is). ∎

With these basic results at our disposal, the proof of Theorem 4 will roughly follow the technical trajectory outlined below:

(1) By Proposition D.1, D_n grows at most by $\gamma_n\xi_n + \frac{1}{2}\gamma_n^2\|V_n\|^2$ at each step. This quantity can be big for any given n but we will show that, with high probability (and, in particular, with probability at least $1-\delta$), the aggregation of these errors remains controllably small. This will be the most technical and involved part of our argument.

(2) Using the above, we will show that, with probability at least $1-\delta$, D_n cannot grow more than a token quantity ε. As a result, if the initial distance to x^* is not too big, X_n will remain in a neighborhood thereof for all time.

(3) For the final part of the theorem, we will condition on this event to map (D.7) to a recursion of the form (D.1), and we will subsequently employ Lemma D.1 to obtain the stated result. The main problem here is that, after conditioning, the noise in (D.7) is no longer zero-mean, so we will need to adapt our analysis to the new noise distribution.

We make all this precise below. For convenience, we focus on the case $p > 1/2$; the case $p \in (2/(q+2), 1/2]$ follows by modifying the arguments that follow with the Hölder estimates we introduced in the proof of Lemma B.1.

Controlling the error terms. We begin by encoding the error terms in (D.7) as

$$M_n = \sum_{k=1}^{n} \gamma_k\xi_k \tag{D.9}$$

and

$$S_n = \frac{1}{2}\sum_{k=1}^{n} \gamma_k^2\|V_k\|^2 \tag{D.10}$$

Since $\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] = 0$, we have $\mathbb{E}[M_n \,|\, \mathcal{F}_n] = M_{n-1}$, so M_n is a zero-mean martingale; likewise, $\mathbb{E}[S_n \,|\, \mathcal{F}_n] \geq S_{n-1}$, so S_n is a submartingale. Interestingly, even though M_n is more "neutral" as an error (because ξ_n is zero-mean), it is more difficult to control because the variance of its increments is

$$\mathbb{E}[\|\gamma_n\xi_n\|^2 \,|\, \mathcal{F}_n] = \gamma_n^2\,\mathbb{E}[\langle Z_n, X_n - x^*\rangle^2 \,|\, \mathcal{F}_n] \tag{D.11}$$

and this last quantity can become arbitrarily big if X_n does not remain in the vicinity of x^* (which is what we are trying to prove). Because of this, we need to take a less direct, step-by-step approach to bound the total error increments *conditioned* on the event that X_n

remains close to x^*. Our approach builds on a range of ideas and techniques due to Hsieh et al. [17, 18] and Mertikopoulos and Zhou [30].

We begin by introducing the "cumulative mean square" error

$$R_n = M_n^2 + S_n. \tag{D.12}$$

By construction, we have

$$\begin{aligned} R_n &= (M_{n-1} + \gamma_n \xi_n)^2 + S_{n-1} + \tfrac{1}{2}\gamma_n^2 \|V_n\|^2 \\ &= R_{n-1} + 2M_{n-1}\gamma_n \xi_n + \gamma_n^2 \xi_n^2 + \tfrac{1}{2}\gamma_n^2 \|V_n\|^2 \end{aligned} \tag{D.13}$$

and hence, after taking expectations:

$$\mathbb{E}[R_n \,|\, \mathcal{F}_n] = R_{n-1} + 2M_{n-1}\gamma_n \,\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] + \gamma_n^2 \,\mathbb{E}[\xi_n^2 + \tfrac{1}{2}\|V_n\|^2 \,|\, \mathcal{F}_n] \geq R_{n-1} \tag{D.14}$$

i.e., R_n is a submartingale. To condition it further, let $\mathcal{U}$ be a neighborhood of x^*, let $\varepsilon > 0$, and define the events

$$\Omega_n \equiv \Omega_n(\mathcal{U}) = \{X_n \in \mathcal{U} \text{ for all } k = 1, 2, \dots, n\} \tag{D.15}$$

and

$$E_n \equiv E_n(\varepsilon) = \{R_k \leq \varepsilon \text{ for all } k = 1, 2, \dots, n\}. \tag{D.16}$$

By definition, we also have $\Omega_0 = E_0 = \Omega$ (because the set-building index set for k is empty in this case, and every statement is true for the elements of the empty set). These events will play a crucial role in the sequel as indicators of whether X_n has escaped the vicinity of x^* or not.

To proceed, we will instantiate $\mathcal{U}$ and ε in the definition of Ω and E respectively as follows. First, for (D.15), we will choose a neighborhood $\mathcal{U}$ contained in the convex compact neighborhood $\mathcal{K}$ of x^* (whose existence is guaranteed by Lemma D.2); in particular, this implies that (D.3) holds for all $x \in \mathcal{U}$. Moreover, with a fair degree of hindsight, we will also choose $\varepsilon > 0$ such that

$$\{x \in \mathbb{R}^d : \|x - x^*\|^2 \leq 4\varepsilon + 2\sqrt{\varepsilon}\} \subseteq \mathcal{U}. \tag{D.17}$$

and we will assume that X_1 is initialized in a neighborhood $\mathcal{U}_1 \subseteq \mathcal{U}$ such that

$$\mathcal{U}_1 \subseteq \{x \in \mathbb{R}^d : \|x - x^*\|^2 \leq 2\varepsilon\} \tag{D.18}$$

These will be the neighborhoods $\mathcal{U}$ and $\mathcal{U}_1$ whose existence is postulated by Theorem 4. Then, with all this in hand, we have:

Lemma D.3. *Let x^* be a regular minimizer of f as above and assume that Assumption 4 holds. Then, for all $n = 1, 2, \dots$, we have:*

(1) $\Omega_{n+1} \subseteq \Omega_n$ and $E_{n+1} \subseteq E_n$.

(2) $E_{n-1} \subseteq \Omega_n$.

(3) Consider the "large noise" event

$$\begin{aligned} \tilde{E}_n &\equiv E_{n-1} \setminus E_n = E_{n-1} \cap \{R_n > \varepsilon\} \\ &= \{R_k \leq \varepsilon \text{ for all } k = 1, 2, \dots, n-1 \text{ and } R_n > \varepsilon\}, \end{aligned} \tag{D.19}$$

and let $\tilde{R}_n = R_n \mathbb{1}_{E_{n-1}}$ denote the cumulative error subject to the noise being "small" until time n. Then:

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_{n-1}] + [G^2 + (1 + r_{\mathcal{U}}^2)\sigma^2]\gamma_n^2 - \varepsilon\, \mathbb{P}(\tilde{E}_{n-1}), \tag{D.20}$$

where $r_{\mathcal{U}} = \sup_{x \in \mathcal{U}} \|x - x^*\|$ and, by convention, we write $\tilde{E}_0 = \varnothing$ and $\tilde{R}_0 = 0$.

Remark. In the above (and what follows), the notation $\mathbb{1}_A$ is used to indicate the logical indicator of an event $A \subseteq \Omega$, i.e., $\mathbb{1}_A(\omega) = 1$ if $\omega \in A$ and $\mathbb{1}_A(\omega) = 0$ otherwise.

Proof. The first claim is obvious. For the second, we proceed inductively:

(1) For the base case $n = 1$, we have $\Omega_1 = \{X_1 \in \mathcal{U}\} \supseteq \{X_1 \in \mathcal{U}_1\} = \Omega$ because X_1 is initialized in $\mathcal{U}_1 \subseteq \mathcal{U}$. Since $E_0 = \Omega$, our claim follows.

(2) For the inductive step, assume that $E_{n-1} \subseteq \Omega_n$ for some $n \geq 1$. To show that $E_n \subseteq \Omega_{n+1}$, fix a realization in E_n so $R_k \leq \varepsilon$ for all $k = 1, 2, \dots, n$. Since $E_n \subseteq E_{n-1}$, the inductive hypothesis posits that Ω_n also occurs, i.e., $X_k \in \mathcal{U}$ for all $k = 1, 2, \dots, n$; hence, it suffices to show that $X_{n+1} \in \mathcal{U}$.

To that end, given that $X_k \in \mathcal{U} \subseteq \mathcal{K}$ for all $k = 1, 2, \dots n$, the distance estimate (D.7) readily gives

$$D_{k+1} \leq D_k + \gamma_k \xi_k + \tfrac{1}{2}\gamma_k^2 \|V_k\|^2 \quad \text{for all } k = 1, 2, \dots n. \tag{D.21}$$

Therefore, after telescoping, we obtain

$$D_{n+1} \leq D_1 + M_n + S_n \leq D_1 + \sqrt{R_n} + R_n \leq \varepsilon + \sqrt{\varepsilon} + \varepsilon = 2\varepsilon + \sqrt{\varepsilon} \tag{D.22}$$

by the inductive hypothesis. We conclude that $\|X_{n+1} - x^*\|^2 = 2D_{n+1} \leq 4\varepsilon + 2\sqrt{\varepsilon}$, so $X_{n+1} \in \mathcal{U}$ and the induction is complete.

For our third claim, we decompose $\tilde{R}_n$ as

$$\begin{aligned}
\tilde{R}_n = R_n \mathbb{1}_{E_{n-1}} &= R_{n-1} \mathbb{1}_{E_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{E_{n-1}} \\
&= R_{n-1} \mathbb{1}_{E_{n-2}} - R_{n-1} \mathbb{1}_{\tilde{E}_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{E_{n-1}}, \\
&= \tilde{R}_{n-1} + (R_n - R_{n-1}) \mathbb{1}_{E_{n-1}} - R_{n-1} \mathbb{1}_{\tilde{E}_{n-1}},
\end{aligned} \tag{D.23}$$

where we used the fact that $E_{n-1} = E_{n-2} \setminus \tilde{E}_{n-1}$ so $\mathbb{1}_{E_{n-1}} = \mathbb{1}_{E_{n-2}} - \mathbb{1}_{\tilde{E}_{n-1}}$ (recall here that $E_{n-1} \subseteq E_{n-2}$). Now, to proceed, (D.13) yields

$$R_n - R_{n-1} = 2M_{n-1}\gamma_n \xi_n + \gamma_n^2 \xi_n^2 + \tfrac{1}{2}\gamma_n^2 \|V_n\|^2 \tag{D.24}$$

so

$$\mathbb{E}[(R_n - R_{n-1}) \mathbb{1}_{E_{n-1}}] = 2\gamma_n \mathbb{E}[M_{n-1}\xi_n \mathbb{1}_{E_{n-1}}] \tag{D.25a}$$

$$+ \gamma_n^2 \mathbb{E}[\xi_n^2 \mathbb{1}_{E_{n-1}}] \tag{D.25b}$$

$$+ \tfrac{1}{2}\gamma_n^2 \mathbb{E}[\|V_n\|^2 \mathbb{1}_{E_{n-1}}] \tag{D.25c}$$

However, since E_{n-1} and M_{n-1} are both $\mathcal{F}_n$-measurable, we have the following estimates:

(1) For the noise term in (D.25a), the second part of Proposition D.1 gives:

$$\mathbb{E}[M_{n-1}\xi_n \mathbb{1}_{E_{n-1}}] = \mathbb{E}[M_{n-1} \mathbb{1}_{E_{n-1}} \mathbb{E}[\xi_n \,|\, \mathcal{F}_n]] = 0. \tag{D.26}$$

(2) The term (D.25b) is where the conditioning on E_{n-1} plays the most important role because it allows us to control the distance $\|X_n - x^*\|$. Specifically, we have:

$$\begin{aligned}
\mathbb{E}[\xi_n^2 \mathbb{1}_{E_{n-1}}] &= \mathbb{E}[\mathbb{1}_{E_{n-1}} \mathbb{E}[\langle Z_n, X_n - x^* \rangle^2 \,|\, \mathcal{F}_n]] \\
&\leq \mathbb{E}[\mathbb{1}_{E_{n-1}} \|X_n - x^*\|^2 \mathbb{E}[\|Z_n\|^2 \,|\, \mathcal{F}_n]] && \{\text{by Cauchy–Schwarz}\} \\
&\leq \mathbb{E}[\mathbb{1}_{\Omega_n} \|X_n - x^*\|^2 \mathbb{E}[\|Z_n\|^2 \,|\, \mathcal{F}_n]] && \{\text{because } E_{n-1} \subseteq \Omega_n\} \\
&\leq r_{\mathcal{U}}^2 \sigma^2. && \{\text{by Assumption 4}\}
\end{aligned}$$

(3) Finally, for the term (D.25c), we have:

$$\mathbb{E}[\|V_n\|^2 \mathbb{1}_{E_{n-1}}] \leq \mathbb{E}[\|V_n\|^2] \leq 2\,\mathbb{E}[\|\nabla f(X_n)\|^2 + \|Z_n\|^2] \leq 2(G^2 + \sigma^2), \tag{D.27}$$

with the last step following from Assumptions 1 and 4.

Thus, putting together all of the above, we obtain:

$$\mathbb{E}[(R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}}] \leq [G^2 + (1 + r_{\mathcal{U}}^2)\sigma^2]\gamma_n^2 \tag{D.28}$$

Going back to (D.23), we have $R_{n-1} > \varepsilon$ if $\tilde{E}_{n-1}$ occurs, so the last term becomes

$$\mathbb{E}[R_{n-1}\,\mathbb{1}_{\tilde{E}_{n-1}}] \geq \varepsilon\,\mathbb{E}[\mathbb{1}_{\tilde{E}_{n-1}}] = \varepsilon\,\mathbb{P}(\tilde{E}_{n-1}). \tag{D.29}$$

Our claim then follows by combining Eqs. (D.23), (D.27) and (D.29). ∎

Controlling the probability of escape. Lemma D.3 is the technical key to show that X_n remains close to x^* with high probability; we formalize this in a final intermediate result below.

Proposition D.2. *Fix some tolerance level $\delta > 0$. If Assumption 4 holds and (SGD) is run with a step-size schedule of the form $\gamma_n = \gamma/(n+m)^p$ for some sufficiently large $m > 0$, we have*

$$\mathbb{P}(E_n) \geq 1 - \delta \quad \textit{for all } n = 1, 2, \dots \tag{D.30}$$

Proof. We begin by bounding the probability of the "large noise" event $\tilde{E}_n = E_{n-1} \setminus E_n$ as follows:

$$\begin{aligned}
\mathbb{P}(\tilde{E}_n) &= \mathbb{P}(E_{n-1} \setminus E_n) = \mathbb{P}(E_{n-1} \cap \{R_n > \varepsilon\}) \\
&= \mathbb{E}[\mathbb{1}_{E_{n-1}} \times \mathbb{1}_{\{R_n > \varepsilon\}}] \\
&\leq \mathbb{E}[\mathbb{1}_{E_{n-1}} \times (R_n/\varepsilon)] \\
&= \mathbb{E}[\tilde{R}_n]/\varepsilon
\end{aligned} \tag{D.31}$$

where, in the second-to-last line, we used the fact that $R_n \geq 0$ (so $\mathbb{1}_{\{R_n > \varepsilon\}} \leq R_n/\varepsilon$). Now, by telescoping (D.20), we get

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_0] + R_* \sum_{k=1}^{n} \gamma_k^2 - \varepsilon \sum_{k=1}^{n} \mathbb{P}(\tilde{E}_{k-1}) \tag{D.32}$$

where we set $R_* = G^2 + (1 + r_{\mathcal{U}}^2)\sigma^2$. Hence, combining (D.31) and (D.32), we obtain the estimate

$$\sum_{k=1}^{n} \mathbb{P}(\tilde{E}_k) \leq \frac{R_*}{\varepsilon} \sum_{k=1}^{n} \gamma_k^2 \leq \frac{R_* \Gamma}{\varepsilon}, \tag{D.33}$$

where we set $\Gamma = \sum_{n=1}^{\infty} \gamma_n^2 = \gamma^2 \sum_{n=1}^{\infty} (n+m)^{-2p}$ and we used the fact that $\tilde{R}_0 = 0$ and $\tilde{E}_0 = \varnothing$ (by convention).

By choosing m sufficiently large, we can ensure that $R_* \Gamma/\varepsilon < \delta$; moreover, since the events $\tilde{E}_k$ are disjoint for all $k = 1, 2, \dots$, we get

$$\mathbb{P}\left(\bigcup_{k=1}^{n} \tilde{E}_k\right) = \sum_{k=1}^{n} \mathbb{P}(\tilde{E}_k) \leq \delta \tag{D.34}$$

and hence:

$$\mathbb{P}(E_n) = \mathbb{P}\left(\bigcap_{k=1}^{n} \tilde{E}_k^{\mathsf{c}}\right) \geq 1 - \delta, \tag{D.35}$$

as claimed. ∎

Putting everything together. We are finally in a position to combine all of the ingredients for the proof of Theorem 4.

Proof of Theorem 4. To begin, define $\mathcal{U}$ and $\mathcal{U}_1$ as in Lemma D.3. Then, by construction, we have:

$$\Omega_{\mathcal{U}} \equiv \{X_n \in \mathcal{U} \text{ for all } n = 1, 2, \dots\} = \bigcap_{n=1}^{\infty} \Omega_n. \tag{D.36}$$

Since the sequence Ω_n is decreasing and $\Omega_n \supseteq E_{n-1}$ (by the second part of Lemma D.3), Proposition D.2 yields

$$\mathbb{P}(\Omega_{\mathcal{U}}) = \inf_n \mathbb{P}(\Omega_n) \geq \inf_n \mathbb{P}(E_{n-1}) \geq 1 - \delta, \tag{D.37}$$

provided that m is chosen large enough. This proves the first part of the theorem, i.e., to the effect that X_n remains close to x^* with probability at least $1 - \delta$.

For the second part of the theorem, Proposition D.1 readily gives

$$D_{n+1} \mathbb{1}_{\Omega_n} \leq (1 - 2\alpha\gamma_n) D_n \mathbb{1}_{\Omega_n} + [\gamma_n \xi_n + \tfrac{1}{2}\gamma_n^2 \|V_n\|^2] \mathbb{1}_{\Omega_n}. \tag{D.38}$$

Now, for any given γ, we can choose m sufficiently large so that $\inf_n(1 - 2\alpha\gamma_n) > 0$ and $\mathbb{P}(\Omega_{\mathcal{U}}) \geq 1 - \delta$ (the latter by Proposition D.2). Moreover, working as in the proof of Lemma D.3, we get

$$\begin{aligned} \mathbb{E}\big[(\gamma_n \xi_n + \tfrac{1}{2}\|V_n\|^2) \mathbb{1}_{\Omega_n}\big] &= \mathbb{E}\big[\mathbb{1}_{\Omega_n} \mathbb{E}\big[\gamma_n \xi_n + \tfrac{1}{2}\gamma_n^2 \|V_n\|^2 \,\big|\, \mathcal{F}_n\big]\big] \\ &\leq \gamma_n^2 (G^2 + \sigma^2). \end{aligned} \tag{D.39}$$

Then, letting $\bar{D}_n = \mathbb{E}[D_n \mathbb{1}_{\Omega_n}] \geq 0$ and recalling that $\Omega_{n+1} \subseteq \Omega_n$ (so $\mathbb{1}_{\Omega_{n+1}} \leq \mathbb{1}_{\Omega_n}$), the two estimates above yield

$$\begin{aligned} \bar{D}_{n+1} \leq \mathbb{E}[D_{n+1} \mathbb{1}_{\Omega_n}] &\leq (1 - 2\alpha\gamma_n)\bar{D}_n + (G^2 + \sigma^2)\gamma_n^2 \\ &\leq \left[1 - \frac{2\alpha\gamma}{(n+m)^p}\right] \bar{D}_n + \frac{(G^2 + \sigma^2)\gamma^2}{(n+m)^{2p}}. \end{aligned} \tag{D.40}$$

Thus, by Lemma D.1, we obtain the bounds:

$$\bar{D}_n \leq \frac{G^2 + \sigma^2}{2\alpha} \frac{\gamma}{n^p} + o\left(\frac{1}{n^p}\right) \qquad \text{if } p < 1, \tag{D.41a}$$

and

$$\bar{D}_n \leq \frac{G^2 + \sigma^2}{2\alpha\gamma - 1} \frac{\gamma^2}{n} + o\left(\frac{1}{n^p}\right) \qquad \text{if } p = 1, \tag{D.41b}$$

provided that $2\alpha\gamma > 1$ for the latter. The claim of the theorem then follows by noting that

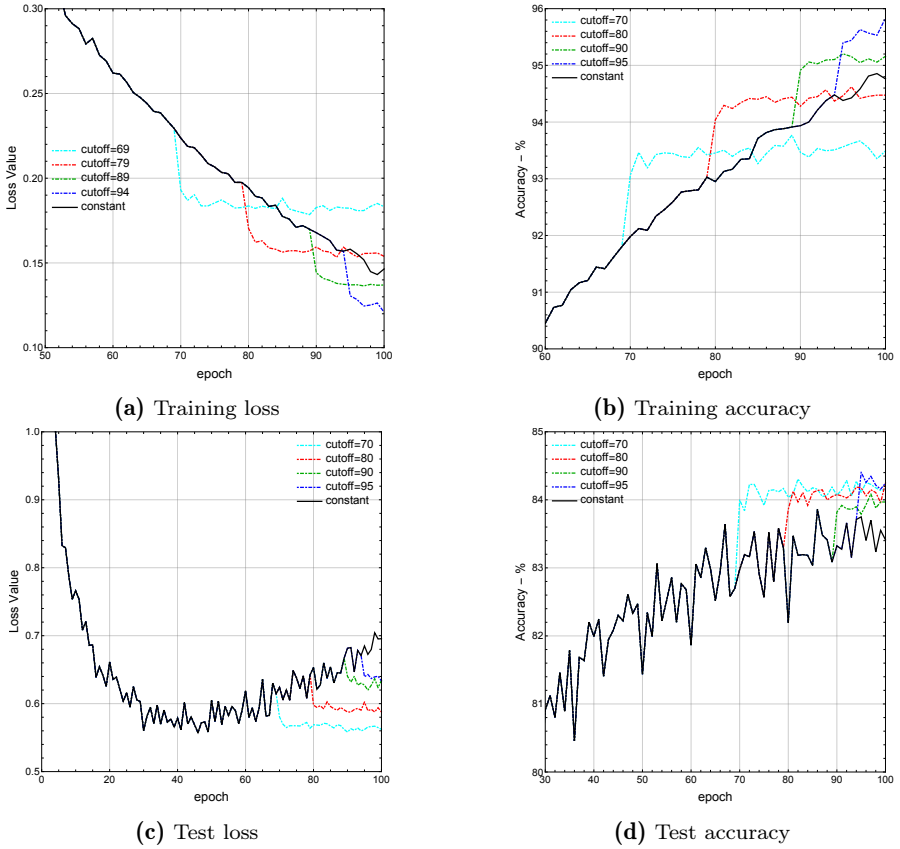
$$\mathbb{E}[\|X_n - x^*\|^2 \,|\, \Omega_{\mathcal{U}}] \leq \frac{\mathbb{E}[\|X_n - x^*\|^2 \mathbb{1}_{\Omega_{\mathcal{U}}}]}{\mathbb{P}(\Omega_{\mathcal{U}})} \leq \frac{2}{1 - \delta} \bar{D}_n \tag{D.42}$$

and applying Eqs. (D.41a) and (D.41b). ∎

Appendix E. Numerical experiments

In this appendix, we present some more details on our ResNet training setup and some additional numerical results. We used the *python/pytorch* implementation of Resnet18 from the *torchvision* package and, for consistency, we downloaded the CIFAR10 dataset from the same package. For training/evaluation purposes, we used the the standard training/test split of 50000/10000 examples, with training and test batches of size 120.

The purposes of our experiments is to demonstrate the possible benefits of the "cooldown" heuristic that is derived from our convergence analysis in Section 4. To that end, we initially



(a) Training loss **(b)** Training accuracy

(c) Test loss **(d)** Test accuracy

Figure 4: Results for training ResNet18 model for classification over CIFAR10 dataset, with cooldown heuristic. Constant step-size SGD is run for 100 epochs and cooldown phase starts at epochs 70, 80, 90 and 95, with diminishing step-size policy of $1/n$.

trained the model with constant step-size (SGD) whose step-size is picked through grid-search over the set $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. We then took checkpoints of the model at certain epochs and launched the cooldown heuristic from such points with a $1/n$ step-size policy. It is important to emphasize that the iteration counter n starts at the first iteration of cooldown phase so that we have a "continuous" sequence of step-sizes across epochs. In Fig. 4, we provide the complementary plots for the setting described in Section 5, again exhibiting a clear benefit (especially in test accuracy and loss) when using the cool-down heuristic for the last part of the experiment's runtime budget.

References

[1] Mark Anthony Armstrong. *Basic Topology*. Springer, 1983.

[2] Michel Benaïm. Dynamics of stochastic approximation algorithms. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, editors, *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer Berlin Heidelberg, 1999.

[3] Michel Benaïm and Morris W. Hirsch. Dynamics of Morse-Smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6):1005–1030, December 1995.

[4] Michel Benaïm and Morris W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

[5] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

[6] Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[7] Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press and Hindustan Book Agency, 2008.

[8] Odile Brandière and Marie Duflo. Les algorithmes stochastiques contournent-ils les pièges ? *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 32(3):395–427, 1996.

[9] Kuo-Liang Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25 (3):463–483, 1954.

[10] Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[11] Lampros Flokas, Emmanouil Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[12] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — Online stochastic gradient for tensor decomposition. In *COLT '15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

[13] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[14] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

[15] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980.

[16] Morris W. Hirsch. *Differential Topology*. Springer-Verlag, Berlin, 1976.

[17] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6936–6946, 2019.

[18] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. https://arxiv.org/abs/2003.10162, 2020.

[19] Momin Jamil and Xin-She Yang. A literature survey of benchmark functions for global optimization problems. https://arxiv.org/abs/1308.4008v1, 2013.

[20] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017. URL http://proceedings.mlr.press/v70/jin17a.html.

[21] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[22] Harold J. Kushner and G. G. Yin. *Stochastic approximation algorithms and applications*. Springer-Verlag, New York, NY, 1997.

[23] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *COLT '16: Proceedings of the 29th Annual Conference on Learning Theory*, 2016.

[24] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176 (1):311–337, February 2019.

[25] John M. Lee. *Introduction to Smooth Manifolds*. Number 218 in Graduate Texts in Mathematics. Springer-Verlag, New York, NY, 2003.

[26] Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, PP, 12 2019. doi: 10.1109/TNNLS.2019.2952219.

[27] Hao Li, Zheng Xu, Gavin Taylor, Christoph Suder, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[28] Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Autom. Control*, 22(4):551–575, August 1977.

[29] Lennart Ljung. *System Identification Theory for the User*. Prentice Hall, Englewood Cliffs, NJ, 1986.

[30] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, January 2019.

[31] John Willard Milnor. *Topology from the Differentiable Viewpoint*. Princeton University Press, Princeton, NJ, 1965.

[32] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

[33] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *ITCS '17: Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*, 2017.

[34] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[35] Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic aproximations. *Annals of Probability*, 18(2):698–712, April 1990.

[36] Robin Pemantle. Vertex-reinforced random walk. *Probability Theory and Related Fields*, 92:117–136, 1992.

[37] Boris Teodorovich Polyak. *Introduction to Optimization*. Optimization Software, New York, NY, USA, 1987.

[38] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

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

[40] R. Clark (Rex) Robinson. *An Introduction to Dynamical Systems: Continuous and Discrete*. American Mathematical Society, Providence, RI, 2 edition, 2012.

[41] Michael Shub. *Global Stability of Dynamical Systems*. Springer-Verlag, Berlin, 1987.

[42] Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[43] Stefan Vlaski and Ali H. Sayed. Second-order guarantees of stochastic gradient descent in non-convex optimization. https://arxiv.org/abs/1908.07023, 2019.