Journal of Machine Learning Research 00 (0000) 1-59 Submitted 7/19; Published 00/00

Convergence rates for the stochastic gradient descent method for non-convex objective functions

Benjamin Fehrman BENJAMIN.FEHRMAN@MATHS.OX.AC.UK
Mathematical Institute, University of Oxford
Oxford OX2 6GG, United Kingdom

Benjamin Gess BENJAMIN.GESS@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany
Fakultät für Mathematik, Universität Bielefeld
33615 Bielefeld, Germany

Arnulf Jentzen ARNULF.JENTZEN@SAM.MATH.ETHZ.CH
Seminar for Applied Mathematics, Department of Mathematics,
ETH Zurich, 8092 Zurich, Switzerland

Editor:

Abstract

We prove the convergence to minima and estimates on the rate of convergence for the stochastic gradient descent method in the case of not necessarily locally convex nor contracting objective functions. In particular, the analysis relies on a quantitative use of mini-batches to control the loss of iterates to non-attracted regions. The applicability of the results to simple objective functions arising in machine learning is shown.
Keywords: stochastic gradient descent, mini-batch algorithm, machine learning, non-convex optimization

1. Introduction

Stochastic gradient descent algorithms (SGD), going back to Robbins and Monro (1951), are the most common way to train neural networks. However, despite their relevance to machine learning and much recent interest, estimates on their rate of convergence have only been obtained under assumptions that are often not satisfied or not known to be satisfied by objective functions arising in machine learning[1]. Indeed, citing from Vidal et al. (2017), "While SGD has been rigorously analyzed only for convex loss functions [...], in deep learning the loss is a non-convex function of the network parameters, hence there are no guarantees that SGD finds the global minimizer." In the present work, we prove the local convergence of SGD with rates to minima of the objective function, while avoiding convexity or contractivity assumptions. By randomly sampling the initialization, convergence to global minima with rates is deduced. We demonstrate the relevance of these results through their application to the training of (simple) neural networks.

Stochastic gradient descent methods are used to numerically minimize functions $f\colon \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(\theta) = \mathbb{E}\left[F(\theta, X)\right], \tag{1}$$

[1]For comments on recent progress on the landscape of objective functions in overparametrized settings see Section 1.1 below.

for some product measurable function $F\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ and some random variable $X\colon \Omega \to \mathbb{R}^m$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The analysis of SGD has attracted considerable attention in the literature (see, for example, Bach (2014); Bach and Moulines (2013); Bottou et al. (2018); Dereich and Mueller-Gronbach (2015); Jentzen et al. (2018); Moulines and Bach (2011); Tang and Monteleoni (2015) and the references therein). The convergence of SGD with rates was shown, for example, in Dereich and Mueller-Gronbach (2015); Jentzen et al. (2018) under the assumption that the objective function f satisfies the following contraction property: There is an $L > 0$ and a zero θ^* of $\nabla_\theta f$ such that, for every $\theta \in \mathbb{R}^d$,

$$(-\nabla_\theta f(\theta), \theta - \theta^*) \leq -L\|\theta - \theta^*\|^2. \tag{2}$$

Property (2) implies the uniqueness of the zero θ^* of $\nabla_\theta f$ and thus the uniqueness of local minima of f. This is in stark contrast to the actual objective functions that arise in the training of neural networks, which are expected to show rich sets of local minima and saddle points/plateaus. Consequently, it is vital for the application to machine learning that we avoid such global contraction assumptions. In addition, for example due to the positive homogeneity of the ReLU function, the objective functions typically satisfy certain symmetries, implying that global (and local) minima are not isolated points nor unique, but form (possibly non-compact) manifolds. Indeed, this is demonstrated for simple neural networks in Section 7 below. We are therefore led to the task of analyzing the convergence properties of SGD in local neighborhoods of the manifolds of minima[2]. Indeed, under Assumptions 1 and 2 below, we prove that there exist local basins of attraction for SGD in the sense that SGD beginning in these sets converges with high probability to a minima of the objective function (cf. Theorem 3 below). Then, by random sampling of the initialization, the convergence of SGD beginning in such local neighborhoods can be upgraded to global convergence (cf. Corollary 4 below).

Our results apply to objective functions which are neither locally contracting nor locally convex and, in particular, we avoid assumptions like the contraction property (2). Precisely, we will consider objective functions $f\colon \mathbb{R}^d \to \mathbb{R}$ whose set of global minima is somewhere locally smooth and whose Hessian is somewhere maximally nondegenerate on this smooth subset:

Assumption 1 *Let $d \in \mathbb{N}$, let $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, let $f\colon \mathbb{R}^d \to \mathbb{R}$, and let $\mathcal{M} \subseteq \mathbb{R}^d$ be defined by*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}. \tag{3}$$

We assume that there exists an open subset $U \subseteq \mathbb{R}^d$ such that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^2-submanifold of $\mathbb{R}^d$, such that the restriction $f|_U\colon U \to \mathbb{R}$ is three times continuously differentiable, and such that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$ for every $\theta \in (\mathcal{M} \cap U)$.

We are interested in objective functions of the form (1), which are defined by a jointly measurable function $F\colon \mathbb{R}^d \times S \to \mathbb{R}$, defined on a measurable space $(S, \mathcal{S})$, and a random variable $X\colon \Omega \to S$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The random variable X will play the role of the noise driving the SGD algorithm. We assume that this collection satisfies the following assumption.

[2] We emphasize that this is disjoint from the recent works Bottou et al. (2018); Li and Orabona (2018); Ward et al. (2018) where the global convergence of the *gradient* of the objective function to zero has been shown for SGD and AdaGrad. This does not imply the local convergence to minima, since the gradient also vanishes in saddles/plateaus.

Assumption 2 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, and let $X\colon \Omega \to S$ be measurable. We assume that for every $x \in S$ the map $\theta \in \mathbb{R}^d \mapsto F(\theta, x)$ is locally Lipschitz continuous and that, for every compact set $\mathfrak{C} \subseteq \mathbb{R}^d$,*

$$\sup_{\theta \in \mathfrak{C}} \mathbb{E}\left[|F(\theta, X)|^2 + |\nabla_\theta F(\theta, X)|^2\right] < \infty.$$

In the following theorem, under Assumptions 1 and 2, we identify a basin of attraction to the local manifold of minima for SGD, and prove that SGD beginning in this basin of attraction converges with high probability to the local manifold of minima with an explicit estimate on the rate of convergence.

Theorem 3 *(cf. Theorem 25, Remark 26, Corollary 28 below) Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, let $\{X_{n,m}\colon \Omega \to \mathbb{R}\}_{n,m\in\mathbb{N}}$ be i.i.d. random variables, and let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1})]$. Assume that f satisfies Assumption 1 and assume that F and $X_{1,1}$ satisfy Assumption 2. Let $\rho \in (2/3, 1)$ and, for every $r \in (0, \infty)$, $M \in \mathbb{N}$, and $x \in \mathbb{R}^d$, let $\{\Theta_n = \Theta_n(\rho, r, M, x)\}_{n\in\mathbb{N}_0}$ be defined by $\Theta_0 = x$ and*

$$\Theta_n = \Theta_{n-1} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}, X_{n,m})\right]. \tag{4}$$

Then, for every $x_0 \in (\mathcal{M} \cap U)$, there exists an open set $V \subseteq \mathbb{R}^d$ containing x_0, $\mathfrak{r} \in (0, \infty)$, and $c \in (0, \infty)$ such that, for every $x \in V$, $r \in (0, \mathfrak{r})$, $\varepsilon \in (0, 1)$, and $n, M \in \mathbb{N}$,

$$\mathbb{P}\left[\left(f(\Theta_n) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\right) \geq \varepsilon\right] \leq c\left(\frac{1}{\varepsilon^2 n^\rho} + \frac{n^{1-\rho}}{M^{\frac{1}{2}}} + r\right). \tag{5}$$

The terms on the righthand side of (5) should be interpreted in the following sense. The first term estimates the convergence of SGD to $\mathcal{M} \cap U$ on the event that the trajectory remains in the basin of attraction V up to time n. The second term estimates the probability that SGD leaves the basin of attraction. In particular, since the exponent $1 - \rho > 0$, for large running times n this term is controlled quantitatively by the choice of the mini-batch size M. The final term accounts for the fact that if r is large then SGD can overshoot the local manifold of minima and thereby land outside the basin of attraction. The role of r is controlled more precisely in Theorem 25 and Remark 26 below, where it is shown that r can roughly be chosen on the order of the diameter of the basin of attraction V.

The following corollary generalizes Theorem 3 to the case of multiple independent copies of SGD with initial data sampled uniformly at random from a non-empty, bounded open set A. We emphasize that the sampling of the initial data is assumed to be independent of the noise driving the independent copies of SGD.

Corollary 4 *(cf. Corollary 28 below) Let $d \in \mathbb{N}$, let $A \subseteq \mathbb{R}^d$ be a bounded non-empty open set, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, let $\{X_{n,m,k}\colon \Omega \to \mathbb{R}\}_{n,m,k\in\mathbb{N}}$ be i.i.d. random variables, and let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1,1})]$. Assume that f satisfies Assumption 1, assume that F and $X_{1,1,1}$ satisfy Assumption 2, and assume that $(\mathcal{M} \cap U \cap A)$ is non-empty. Let $\rho \in (2/3, 1)$, let $\{\Theta_{0,k}\}_{k\in\mathbb{N}}$ be i.i.d random variables that are independent of $\{X_{n,m,k}\}_{n,m,k\in\mathbb{N}}$ and that are uniformly distributed on A, and for every $r \in (0, \infty)$ and $M \in \mathbb{N}$ let $\{\Theta_{n,k} = \Theta_{n,k}(\rho, r, M, A)\}_{n,k\in\mathbb{N}}$ be defined by*

$$\Theta_{n,k} = \Theta_{n-1,k} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,k}, X_{n,m,k})\right]. \tag{6}$$

Then there exists an open set $V \subseteq \mathbb{R}^d$ containing $(\mathcal{M} \cap U \cap A)$, $\mathfrak{r} \in (0, \infty)$, and $c \in (0, \infty)$ such that, for every $r \in (0, \mathfrak{r})$, $\varepsilon \in (0,1)$, and $n, M, K \in \mathbb{N}$,

$$\mathbb{P}\left[\left(\min_{k\in\{1,2,\dots,K\}} f(\Theta_{n,k}) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\right) \geq \varepsilon\right] \leq \left(\frac{|A \setminus V|}{|A|} + c\left(\frac{1}{\varepsilon^2 n^\rho} + \frac{n^{1-\rho}}{M^{\frac{1}{2}}} + r\right)\right)^K. \tag{7}$$

We emphasize that Corollary 4 is only nontrivial due to the assumption that $\mathcal{M} \cap U \cap A$ is non-empty. Indeed, if the initialization is chosen outside of a basin of attraction for $\mathcal{M} \cap U$, then the algorithm cannot in general converge. The simple networks presented in Section 7 demonstrate this fact, where in their case the lack of convergence is due to the vanishing gradient of the rectifier function. This observation corresponds to the pronounced relevance of initialization in practice, see, for example, (Li et al., 2018, p. 8). The large, flat, nearly convex basins of attraction observed therein correspond to large sets V in our context, that is, to small ratios $\frac{|A\setminus V|}{|A|}$ in (7).

It remains to identify the optimal choice of parameter $\{\Theta_{n,k}\}_{k\in\{1,2,\dots,K\}}$ that attains the minimum appearing on the lefthand side of (7). For this, we introduce a second mini-batch approximation due to the fact that, in practice, the objective function cannot be computed or cannot be efficiently computed. Theorem 5 below introduces this mini-batch approximation, and Corollary 6 below estimates computational efficiency of the algorithm.

Theorem 5 *(cf. Theorem 30 below) In the setting of Corollary 4, for every $\mathfrak{M} \in \mathbb{N}$ there exists a random variable $\Theta_n \colon \Omega \to \mathbb{R}^d$ which satisfies*

$$\frac{1}{\mathfrak{M}}\sum_{k=1}^{\mathfrak{M}} F(\Theta_n, X_{n+1,1,k}) = \min_{k\in\{1,2\dots,K\}}\left[\frac{1}{\mathfrak{M}}\sum_{k=1}^{\mathfrak{M}} F(\Theta_{n,k}, X_{n+1,1,k})\right], \tag{8}$$

(cf. Lemma 29 below) and there exists an open set $V \subseteq \mathbb{R}^d$ containing $(\mathcal{M} \cap U \cap A)$, $\mathfrak{r} \in (0, \infty)$, and $c \in (0, \infty)$ such that, for every $r \in (0, \mathfrak{r})$, $\varepsilon \in (0,1)$, and $n, M, \mathfrak{M}, K \in \mathbb{N}$,

$$\mathbb{P}\left[\left(f(\Theta_n) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\right) \geq \varepsilon\right] \leq \frac{cK}{\varepsilon^2\mathfrak{M}} + \left(\frac{|A \setminus V|}{|V|} + c\left(\frac{1}{\varepsilon^2 n^\rho} + \frac{n^{1-\rho}}{M^{1/2}} + r\right)\right)^K. \tag{9}$$

Corollary 6 *(cf. Corollary 31 below) In the setting of Theorem 5, there exist $\{c_i \in (0,\infty)\}_{i\in\{1,2,3,4\}}$ such that, for every $\varepsilon, \eta \in (0,1)$, for $n(\varepsilon), M(\varepsilon), K(\eta), \mathfrak{M}(\varepsilon,\eta) \in \mathbb{N}$ defined by*

$$n(\varepsilon) = c_1\varepsilon^{-2/\rho}, \;\; M(\varepsilon) = c_2\varepsilon^{-4/\rho+4}, \;\; \mathfrak{M}(\varepsilon,\eta) = c_3\varepsilon^{-2}\eta^{-1}\left|\log(\eta)\right|, \;\; \text{and} \;\; K = c_4\left|\log(\eta)\right|, \tag{10}$$

we have that

$$\mathbb{P}\Big(\Big[f(\Theta_{n(\varepsilon)} - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \eta. \tag{11}$$

The conclusion of Corollary 6 proves that, for every $\varepsilon, \eta \in (0,1)$, to guarantee the conclusion of (11) it is sufficient to make a number of computations on the order of

$$\varepsilon^{-6/\rho+4}\left|\log(\eta)\right| + \varepsilon^{-2}\eta^{-1}\left|\log(\eta)\right|. \tag{12}$$

The first term of (12) estimates the cost of simulating K-independent copies of mini-batch SGD and the final term estimates the cost of the mini-batch approximation introduced in Theorem 5. In particular, we observe that as $\rho \to 1$ the sufficient mini-batch size for the SGD algorithm $M \simeq \varepsilon^{-4/\rho+4}$ is effectively an order one quantity.

We next comment on some of the difficulties arising in the proofs of the statements above. In a non-globally stable setting, i.e. when (2) is not satisfied, several obstacles in

the proof of local convergence to minima and the estimation of the rate for SGD appear. In particular, even pretending a local minimum to be isolated and such that (2) holds in a neighborhood V of the minimum, the global analysis put forward in Jentzen et al. (2018) is not immediately localizable, since deterministic bounded sets are not invariant under the dynamics of SGD. On the contrary, with probability one each realization of SGD will eventually leave the local basin of attraction V, outside of which no control on the dynamics can be expected. Furthermore, as pointed out above, (local) minima are not expected to appear in an isolated manner, but as (local) manifolds. This needs to be accounted for in the mathematical analysis, giving rise to a quantitative analysis inspired by the center manifold theorem.

The probability that SGD leaves a basin of attraction is estimated by effectively splitting the dynamics of SGD in directions normal and directions tangential to the local manifold of minima. In Proposition 20, we quantify the convergence in normal directions by proving an optimal rate of convergence for SGD conditioned on the event that it remains in the basin of attraction. In Proposition 21 below, we estimate the tangential movement of SGD by estimating the maximal excursions of SGD on the event that it remains in the basin of attraction. In Proposition 24 below, we argue inductively that the normal convergence and maximal excursion estimate imply with high probability that SGD remains in the basin of attraction for large times. Theorem 25 then proves that SGD beginning in the basin of attraction converges to the global minimum with high probability.

1.1 Literature

The SGD algorithm has attained considerable interest in the literature, and a complete account on the existing results would go beyond the scope of this article. Therefore, we restrict to works that seem most relevant to the current results and we refer to Bercu and Fort (2013); Bottou et al. (2018); Ruder (2016) and the references therein for overview articles on SGD type optimization algorithms.

The case of a convex loss function is well-understood under mild further assumptions. For example, rates of convergence of the order $O(1/\sqrt{n})$ for SGD have been established in Bottou et al. (2018); Zhang (2004). In the case of a strongly convex objective function these can be improved to $O(1/n)$, see Hazan et al. (2007); Nemirovski et al. (2009); Nesterov (2013), Tang and Monteleoni (2015) and Rakhlin et al. (2011).

The case of a non-convex objective function is considerably less well understood. In this case we have to distinguish two classes of results: The first class proves the convergence to zero (with or without rates) for the gradient of the objective function, thus implying the convergence to a critical point. The second class of results proves the (local or global) convergence of the values of the loss function to their global minimum. Obviously, the second class of results are stronger and not implied by the first class, since the latter do not exclude the convergence to saddle points or local minima. The results of this work fall into the second class of results.

Rather complete results are known concerning the minimization of the gradient of a non-convex loss function. For an introduction and overview we refer to Bottou et al. (2018). The convergence of the gradient to zero with rates was shown in Lei et al. (2019) assuming a Hölder-regularity condition on the gradient. This generalizes previous work Ghadimi et al. (2013) which required a second moment boundedness condition, which in turn was generalized by the works Ghadimi and Lan (2013) and Reddi et al. (2016).

In the convex case, the convergence of SGD to the global minimum with rates was obtained in Ghadimi et al. (2013). However, in the non-convex case, only partial results are known. For example, Ge et al. (2015) introduced the strict saddle property, which effectively excludes the occurrence of plateaus. Under this assumption rates of convergence

to local minima could be shown, later improved by Jin et al. (2017) to dimension-free estimates. In contrast, in the present work, the local convergence to a global minimum is shown without excluding the presence of plateaus. Several other assumptions replacing (strong) convexity have been considered, for example, the error bounds condition in Luo and Tseng (1993), essential strong convexity in Liu et al. (2013), weak strong convexity in Necoara et al. (2015), the restricted secant inequality in Zhang and Yin (2013), and the quadratic growth condition in Anitescu (2000). In these works, linear convergence rates to a global minimum are shown. In the notable contribution Karimi et al. (2016) it is shown that all of these conditions imply the Polyak-Lojasiewicz (PL) inequality, introduced in Lojasiewicz (1963) and Polyak (1963), under which linear convergence of the SGD to a global minimum is proven in Karimi et al. (2016), thus generalizing these previous works. Recently, further progress was made in Lei et al. (2019) where a boundedness assumption on the gradient of the objective function, required in Karimi et al. (2016), was relaxed. We note that, while the PL condition does not require convexity, nor the uniqueness of global minimizers, it does exclude the existence of local minima. That is, assuming the PL condition, every local minimum is a global minimum. Therefore, it is not implied by the assumptions made in the current work.

In a line of recent developments structural properties of objective functions arising in machine learning have been better understood in the overparametrized setting, that is, in the regime of number of degrees of freedom d being much larger than the size of the training set. More precisely, in this overparametrized regime, the approximate convexity of the mean-field limit (cf. Chizat and Bach (2018)) has been used to prove the convergence of gradient descent to a global minimum (cf. Du et al. (2018b); Rotskoff and Vanden-Eijnden (2018)) for shallow networks. Taking the related viewpoint of gradient kernel methods, similar effects have been unveiled in Jacot et al. (2018), applying also to deep networks. Subsequently, extensions to deep networks have been recently given in Du et al. (2018a).

In the literature a large number of variants and modifications of the SGD algorithm have been introduced. For example, Bach (2014) and Bach and Moulines (2013, 2011) analyze the averaged stochastic gradient descent algorithm in convex but non-strictly convex situations. A discussion of the respective choice of the learning rate can be found in Xu (2013). For an analysis of the choice of the learning rate in SGD we refer to Darken et al. (1992). The SGD with a constant learning rate has been considered in Dieuleveut et al. (2017) and Bottou et al. (2018), and adaptive learning rates in Schaul et al. (2012). Second order stochastic gradient descent and a discussion of its efficiency are available in Bottou (2010), Bottou and Bousquet (2011). A principle comparison of online versus batch learning algorithms can be found in Bottou and LeCun (2004). Natural gradient descent has been introduced in Amari (1998) and analyzed in Rattray et al. (1998) and Inoue et al. (2003). Relations to Hessian-Free Optimization, Krylov Subspace Descent, and TONGA have been pointed out in Pascanu and Bengio (2014). The relevance of the initialization for SGD with momentum has been numerically analyzed in Sutskever et al. (2013). See also Qian (1999) and Zhang (2004) for further details on gradient descent with momentum. The convergence of adaptive gradient methods to first-order stationary points with rates has been shown in Zhou et al. (2018a). SGD algorithms based on nested variance reduction have been introduced and studied in Zhou et al. (2018b).

We further point out that in the non-convex setting modifications of the SGD algorithm have been explored in order to avoid the un-favorable behavior of SGD in case of local minima and plateaus. For example, Raginsky et al. (2017), Xu et al. (2017) and Zhang et al. (2017) consider stochastic gradient Langevin dynamics and prove rates of convergence to the corresponding stationary distribution. This modification is based on the introduction of an additional random perturbation to SGD, which may cause SGD to leave unfavorable regions

like plateaus. Instead, in the present work—as it is often done in practice—resampling of the initial datum is used to avoid such unfavorable regions (see Corollary 4).

Details on the implementation of the SGD and its variants, with focus on particular applications can be found in the following contributions: A general software framework allowing parallelization of SGD has been presented in Dean et al. (2012). Details on the implementation of the SGD in speech recognition can be found in Deng et al. (2013). The use of SGD in text generation by means of recurrent neural networks has been considered in Graves (2013). Deep recurrent neural networks in speech recognition have been trained via SGD in Graves et al. (2013) and in Hinton et al. (2012) also variants of the SGD like momentum SGD have been used. The efficiency of gradient descent in the context of dimension reduction via autoencoders has been discussed in Hinton and Salakhutdinov (2006). Application of the SGD to image classification can be found in Krizhevsky et al. (2012). A comparison of several types of gradient descent algorithms in handwritten character recognition has been presented in LeCun et al. (1998).

1.2 Structure of the work

The paper is organized as follows. In Section 2, we present some geometric preliminaries that are used to identify a basin of attraction for SGD. In Section 3, for objective functions which satisfy Assumption 1, we prove a local exponential rate of convergence for a deterministic gradient descent algorithm in continuous time. In Section 4, for objective functions that satisfy Assumption 1, we establish a local exponential rate of convergence for a deterministic gradient descent algorithm in discrete time. These results are meant to explain the role of our assumptions in a simplified setting. In Section 5, in the setting of Theorem 3, we analyze the convergence of SGD to the local manifold of minima. In Section 6, we prove that the estimates of Section 5 can be improved under the additional assumption that $\mathcal{M} \cap U$ is compact. And in Section 7, we prove that Assumptions 1 and 2 are satisfied by simple loss functions arising in machine learning applications.

2. Geometric preliminaries

In this section, for an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ that satisfies Assumption 1, we will characterize the local geometry of the local manifold of minima $\mathcal{M} \cap U$. The analysis will rely on on the notion of a local projection to $\mathcal{M} \cap U$.

Proposition 7 *Let $d \in \mathbb{N}$, let $\mathfrak{d} \in \{1, \dots, d-1\}$, and let $\mathcal{M} \cap U \subseteq \mathbb{R}^d$ be a non-empty $\mathfrak{d}$-dimensional C^2-submanifold of $\mathbb{R}^d$. Then for every $x_0 \in (\mathcal{M} \cap U)$ there exists an open neighborhood $V = V(x_0) \subseteq \mathbb{R}^d$ that satisfies the following three properties.*

(i) V is a neighborhood of x_0: we have $x_0 \in V$

(ii) Projections exist in V: there exists a unique function $p\colon V \to (\mathcal{M} \cap U)$ such that, for every $x \in V$,

$$|x - p(x)| = \inf\{|x - y| : y \in (\mathcal{M} \cap U)\} \tag{13}$$

(iii) The projection map is locally C^1-smooth: the map $p\colon V \to (\mathcal{M} \cap U)$ is once continuously differentiable

Proof [Proof of Proposition 7] The proof is an immediate consequence of (Foote, 1984, Lemma) and the C^2-regularity of $\mathcal{M} \cap U$. ∎

In the following definition, we define the projection map on a global neighborhood of $\mathcal{M} \cap U$. The existence of the projection map is an immediate consequence of Proposition 7.

Definition 8 *Let $d \in \mathbb{N}$, let $\mathfrak{d} \in \{1, \ldots, d-1\}$, and let $\mathcal{M} \cap U \subseteq \mathbb{R}^d$ be a non-empty $\mathfrak{d}$-dimensional C^2-submanifold of $\mathbb{R}^d$.*

(i) For every $x \in (\mathcal{M} \cap U)$ let $\mathrm{Proj}(x) \subseteq \mathcal{B}(\mathbb{R}^d)$ be defined by

$$\mathrm{Proj}(x) = \{V \subseteq \mathbb{R}^d \colon V \text{ satisfies the conclusion of Proposition 7 with } x_0 = x\}. \tag{14}$$

(ii) Let $p \colon \cup_{x \in (\mathcal{M} \cap U)} \big(\cup_{V \in \mathrm{Proj}(x)} V \big) \to (\mathcal{M} \cap U)$ be the unique function which satisfies for every $x \in \cup_{x \in (\mathcal{M} \cap U)} \big(\cup_{V \in \mathrm{Proj}(x)} V \big)$ that

$$|x - p(x)| = \inf\{|x - y| : y \in (\mathcal{M} \cap U)\}. \tag{15}$$

The following proposition characterizes the tangent and normal spaces of $\mathcal{M} \cap U$ in terms of the Hessian matrix of the objective function. The tangent space $T_x(\mathcal{M} \cap U)$ of $M \cap U$ at $x \in \mathcal{M} \cap U$ is the null space of the Hessian, and the normal space $(T_x(\mathcal{M} \cap U))^{\perp}$ of $\mathcal{M} \cap U$ at $x \in \mathcal{M} \cap U$ is the space on which the Hessian is positive definite.

Proposition 9 *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 1. Then for every $x \in (\mathcal{M} \cap U)$ there exist a $(d - \mathfrak{d})$-dimensional subvectorspace $P_x \subseteq \mathbb{R}^d$ and a $\mathfrak{d}$-dimensional subvectorspace $N_x \subseteq \mathbb{R}^d$ that satisfy the following properties.*

(i) We have that

$$\big(\mathrm{Hess}\, f\big)(x)(P_x) = P_x. \tag{16}$$

(ii) For every $v \in P_x \backslash \{0\}$,

$$\big(\big[\big(\mathrm{Hess}\, f\big)(x)\big]v\big) \cdot v > 0. \tag{17}$$

(iii) We have

$$\big(\mathrm{Hess}\, f\big)(x)|_{N_x} = 0. \tag{18}$$

(iv) We have that

$$N_x = T_x(\mathcal{M} \cap U). \tag{19}$$

(v) We have that

$$P_x = \big(T_x(\mathcal{M} \cap U)\big)^{\perp}. \tag{20}$$

Proof [Proof of Proposition 9] Let $x \in (\mathcal{M} \cap U)$. Since $\mathrm{rank}((\mathrm{Hess}\, f)(\theta)) = d - \mathfrak{d}$, the symmetry of the Hessian implies that there exist subspaces $N_x, P_x \subseteq \mathbb{R}^d$ such that $\mathbb{R}^d = P_x \oplus N_x$, that $\dim(P_x) = d - \mathfrak{d}$, that

$$\big(\mathrm{Hess}\, f\big)(x)(P_x) \subseteq P_x \text{ with } \big(\mathrm{Hess}\, f\big)(x)|_{P_x} \text{ strictly positive definite on } P_x, \tag{21}$$

that $\dim(N_x) = \mathfrak{d}$, and that

$$\big(\mathrm{Hess}\, f\big)(x)|_{N_x} = 0. \tag{22}$$

Let $\varepsilon \in (0, 1)$ and suppose that $\gamma \colon (-\varepsilon, \varepsilon) \to \mathcal{M} \cap U$ is a smooth curve which satisfies $\gamma(0) = x$. Since $\nabla f|_{\mathcal{M} \cap U} = 0$, it follows from the chain rule that

$$\frac{d}{dt} \nabla f(\gamma(t))\bigg|_{t=0} = \big(\mathrm{Hess}\, f\big)(x) \cdot \dot{\gamma}(0) = 0. \tag{23}$$

It follows that $T_x(\mathcal{M} \cap U) \subseteq N_x$ and therefore, since $\dim(T_x(\mathcal{M} \cap U)) = \mathfrak{d}$, it holds that $T_x(\mathcal{M} \cap U) = N_x$. Since $\mathbb{R}^d = T_x(\mathcal{M} \cap U) \oplus \big(T_x(\mathcal{M} \cap U)\big)^{\perp}$, it holds that $P_x = \big(T_x(\mathcal{M} \cap U)\big)^{\perp}$, which completes the proof of Proposition 9. ∎

In the following lemma, for a point $x \in \mathbb{R}^d$ such that the projection $p(x) \in (\mathcal{M} \cap U)$ is well-defined, we prove that the difference $x - p(x) \in \mathbb{R}^d$ lies in the space normal to $\mathcal{M} \cap U$ at $p(x)$.

Lemma 10 *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 1. Then for every $x_0 \in (\mathcal{M} \cap U)$, for every $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8), we have for every $x \in V$ that*

$$x - p(x) \in T_{p(x)} (\mathcal{M} \cap U)^{\perp}. \tag{24}$$

Proof [Proof of Lemma 10] Let $x_0 \in (\mathcal{M} \cap U)$, let $V \in \mathrm{Proj}(x_0)$, and let $p : V \to (\mathcal{M} \cap U)$ denote the projection map. Let $x \in V$. If $x \in (\mathcal{M} \cap U)$, the claim is immediate since then $x - p(x) = 0$. If $x \notin \mathcal{M} \cap U$, for some $\varepsilon \in (0, 1)$ suppose that $\gamma\colon (-\varepsilon, \varepsilon) \to \mathcal{M} \cap U$ is a smooth path which satisfies $\gamma(0) = p(x)$. It holds that

$$\frac{d}{dt} |x - \gamma(t)|^2 \Big|_{t=0} = -2\dot{\gamma}(0) \cdot (x - p(x)) = 0. \tag{25}$$

Therefore, since the curve γ was arbitrary, it holds that $x - p(x) \in T_{p(x)} (\mathcal{M} \cap U)^{\perp}$, which completes the proof of Lemma 10. ■

In the following lemma, we derive a formula for the derivative of the distance function to the manifold in a neighborhood of $\mathcal{M} \cap U$.

Lemma 11 *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 1 and let $\mathbf{d}(\cdot, \mathcal{M} \cap U) : \mathbb{R}^d \to \mathbb{R}$ be defined by*

$$\mathbf{d}(x, \mathcal{M} \cap U) = \inf \{|x - y| : y \in (\mathcal{M} \cap U)\}. \tag{26}$$

Then for every $x_0 \in (\mathcal{M} \cap U)$, for every $V \in \mathrm{Proj}(x_0)$ (cf. Definitition 8), we have for every $x \in (V \setminus (\mathcal{M} \cap U))$ that

$$(\nabla \mathbf{d})(x, \mathcal{M} \cap U) = \frac{x - p(x)}{|x - p(x)|}. \tag{27}$$

Proof [Proof of Lemma 11] Let $x_0 \in (\mathcal{M} \cap U)$ and let $V \in \mathrm{Proj}(x_0)$. It follows from Proposition 7 that

$$x \in V \mapsto |x - p(x)|^2 = \mathbf{d}(x, \mathcal{M} \cap U)^2 \text{ is } \mathrm{C}^1. \tag{28}$$

The chain rule implies for every $i \in \{1, \ldots, d\}$ that

$$\frac{\partial}{\partial x_i} \mathbf{d}(x, \mathcal{M} \cap U)^2 = \frac{\partial}{\partial x_i} |x - p(x)|^2 = 2(x - p(x)) \cdot e_i - 2(x - p(x)) \cdot \frac{\partial}{\partial x_i} p(x). \tag{29}$$

Since $\frac{\partial}{\partial x_i} p(x) \in N_{p(x)}$ and since $x - p(x) \in P_{p(x)}$ it follows from Lemma 10 that

$$(x - p(x)) \cdot \frac{\partial}{\partial x_i} p(x) = 0. \tag{30}$$

Since, for every $x \in V \setminus \mathcal{M} \cap U$,

$$\nabla \mathbf{d}(x, \mathcal{M} \cap U)^2 = 2\mathbf{d}(x, \mathcal{M} \cap U) \nabla \mathbf{d}(x, \mathcal{M} \cap U) = 2(x - p(x)), \tag{31}$$

we have for every $x \in V \setminus \mathcal{M} \cap U$ that

$$\nabla \mathbf{d}(x, \mathcal{M} \cap U) = \frac{x - p(x)}{|x - p(x)|}, \tag{32}$$

which completes the proof of Lemma 11. ■

We will now quantify what are essentially local tubular neighborhoods of the local manifold $\mathcal{M} \cap U$. The following definition will play an important role throughout the paper.

Definition 12 *Let $d \in \mathbb{N}$, let $\mathfrak{d} \in \{1, \ldots, d-1\}$, and let $\mathcal{M} \cap U \subseteq \mathbb{R}^d$ be a non-empty $\mathfrak{d}$-dimensional C^2-submanifold of $\mathbb{R}^d$. For every $x \in (\mathcal{M} \cap U)$ and $R, \delta \in (0, \infty)$ let $V_{R,\delta}(x) \subseteq \mathbb{R}^d$ be defined by*

$$V_{R,\delta}(x) = \{y + v \colon y \in (\overline{B}_R(x) \cap \mathcal{M} \cap U) \text{ and } v \in \big(T_y(\mathcal{M} \cap U)\big)^{\perp} \text{ with } |v| < \delta\}. \tag{33}$$

A useful feature of the sets defined in Definition 12 is that the parameter $R \in (0, \infty)$ can be used to quantify distance in directions tangential to the manifold $\mathcal{M} \cap U$, and the parameter $\delta \in (0, \infty)$ can be used to quantify distance in directions normal to the manifold. The following proposition is essentially the tubular neighborhood theorem, and it gives a useful characterization of the sets $V_{R,\delta}(x)$.

Proposition 13 *Let $d \in \mathbb{N}$, let $\mathfrak{d} \in \{1, \ldots, d-1\}$, and let $\mathcal{M} \cap U \subseteq \mathbb{R}^d$ be a non-empty $\mathfrak{d}$-dimensional C^2-submanifold of $\mathbb{R}^d$. Then for every $x_0 \in (\mathcal{M} \cap U)$, for every $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8), there exist $R_0, \delta_0 \in (0, \infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$,*

(i) $\overline{V}_{R,\delta}(x_0) \subseteq V$ (cf. Definition 12),

(ii) we have that

$$V_{R,\delta}(x_0) = \{x \in \mathbb{R}^d \colon \mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta\}, \tag{34}$$

(iii) and, for every $x \in (\overline{B}_R(x_0) \cap \mathcal{M} \cap U)$ and $v \in \big(T_x(\mathcal{M} \cap U)\big)^{\perp}$ with $|v| < \delta$,

$$p(x + v) = x. \tag{35}$$

Proof [Proof of Proposition 13] Let $x_0 \in (\mathcal{M} \cap U)$. For every $R, \delta \in (0, \infty)$ let $\tilde{V}_{R,\delta}(x_0) \subseteq \mathbb{R}^d$ be defined by

$$\tilde{V}_{R,\delta}(x_0) = \{x \in \mathbb{R}^d \colon \mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta\}. \tag{36}$$

Let $V \in \mathrm{Proj}(x_0)$. Since $U, V \subseteq \mathbb{R}^d$ are open, there exist $R_0, \delta_0 \in (0, \infty)$ such that for every $R \in (0, R_0]$ it holds that

$$\overline{B}_R(x_0) \cap \mathcal{M} \subseteq \mathcal{M} \cap U, \tag{37}$$

and for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ that

$$V_{R,\delta}(x_0) \subseteq V \text{ and } \tilde{V}_{R,\delta}(x_0) \subseteq V. \tag{38}$$

Following (Foote, 1984, Lemma), the normal bundle $T\,(\mathcal{M} \cap U)^{\perp} \subseteq \mathbb{R}^{2d}$ satisfies that

$$T\,(\mathcal{M} \cap U)^{\perp} = \left\{(x, v) \in \mathbb{R}^d \times \mathbb{R}^d \colon x \in (\mathcal{M} \cap U) \text{ and } v \in T_x\,(\mathcal{M} \cap U)^{\perp}\right\}. \tag{39}$$

Since $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional C^2-submanifold, it follows that $T(\mathcal{M} \cap U)^{\perp} \subseteq \mathbb{R}^{2d}$ is a d-dimensional C^2-submanifold. Furthermore, the map $\Psi \colon T(\mathcal{M} \cap U)^{\perp} \to \mathbb{R}^d$ which satisfies for every $(x, v) \in T\,(\mathcal{M} \cap U)^{\perp}$ that $\Psi(x, v) = x + v$ satisfies for every $x \in (\mathcal{M} \cap U)$ that

$$D_{(x,0)}\Psi \colon T_{(x,0)}\big(T\,(\mathcal{M} \cap U)^{\perp}\big) \to T_x\mathbb{R}^d \text{ is nonsingular.} \tag{40}$$

It follows from the inverse function theorem that there exists $\delta_1 \in (0, (\delta_0 \wedge {}^{R_0}\!/_4))$ such that, for every $R \in (0, {}^{R_0}\!/_2]$, $\delta \in (0, \delta_1]$,

$$\Psi \colon \{(x, v) \in (TM)^{\perp} : x \in \overline{B}_{R+2\delta_1}(x_0) \text{ and } |v| < \delta\} \to V_{R+2\delta_1,\delta}(x_0) \text{ is injective.} \tag{41}$$

Let $R \in (0, {}^{R_0}/_2]$, $\delta \in (0, \delta_1]$. We will first prove that $\tilde{V}_{R,\delta}(x_0) \subseteq V_{R,\delta}(x_0)$. Let $x \in \tilde{V}_{R,\delta}(x_0)$. If $x \in \overline{B}_R(x_0) \cap \mathcal{M} \cap U$ then it holds by definition that $x \in V_{R,\delta}(x_0)$. If $x \notin \overline{B}_R(x_0) \cap \mathcal{M} \cap U$, since $x \in \tilde{V}_{R,\delta}(x_0)$ implies that $\mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U)$ and since the choice of $R_0 \in (0, \infty)$ implies that

$$\overline{B}_R(x_0) \cap \mathcal{M} \cap U = \overline{B}_R(x_0) \cap \mathcal{M} \text{ is a closed subset of } \mathbb{R}^d, \tag{42}$$

we have $p(x) \in \overline{B}_R(x_0) \cap \mathcal{M} \cap U$. Since $\mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) = |x - p(x)| < \delta$ and since

$$x = p(x) + |x - p(x)| \, \frac{x - p(x)}{|x - p(x)|}, \tag{43}$$

for $\frac{x-p(x)}{|x-p(x)|} \in T_x \left(\mathcal{M} \cap U\right)^{\perp}$ by Lemma 10, we have that $x \in V_{R,\delta}(x_0)$. This completes the proof that $\tilde{V}_{R,\delta}(x_0) \subseteq V_{R,\delta}(x_0)$. It remains to prove that $V_{R,\delta}(x_0) \subseteq \tilde{V}_{R,\delta}(x_0)$. Let $x \in V_{R,\delta}(x_0)$. It is necessary to show that $\mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta$. The definition of $V_{R,\delta}(x_0)$ implies that there exist $\tilde{x} \in (\overline{B}_R(x_0) \cap \mathcal{M} \cap U)$ and $\tilde{v} \in T_{\tilde{x}} \left(\mathcal{M} \cap U\right)^{\perp}$ with $|\tilde{v}| < \delta$ which satisfy that $x = \tilde{x} + \tilde{v}$. We will prove that $p(x) = \tilde{x}$. By contradiction, suppose that $p(x) \neq \tilde{x}$. This implies that

$$|x - p(x)| < |x - \tilde{x}| = |\tilde{v}| < \delta. \tag{44}$$

It follows from the triangle inequality that

$$|p(x) - \tilde{x}| \leq |p(x) - x| + |x - \tilde{x}| < 2\delta \leq 2\delta_1, \tag{45}$$

which proves that

$$x = \tilde{x} + \tilde{v} = p(x) + (x - p(x)), \tag{46}$$

for $x - p(x) \in T_{p(x)} \left(\mathcal{M} \cap U\right)^{\perp}$ by Lemma 10 with $|x - p(x)| < \delta$. Since $\tilde{x} \in (\overline{B}_R(x_0) \cap \mathcal{M} \cap U)$, it follows from (45) that $p(x) \in (\overline{B}_{R+2\delta_1}(x_0) \cap \mathcal{M} \cap U)$. Since $R \in (0, {}^{R_0}/_2]$ and since $\delta \in (0, \delta_1]$, equation (46) contradicts (41), which states that Ψ is injective on the set

$$\{(x, v) \in (TM)^{\perp} : x \in B_{R+2\delta_1}(x_0) \text{ and } |v| < \delta\}. \tag{47}$$

We conclude that $p(x) = \tilde{x}$, which implies that

$$\mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) = |x - p(x)| = |\tilde{v}| < \delta. \tag{48}$$

We conclude that $V_{R,\delta}(x_0) \subseteq \tilde{V}_{R,\delta}(x_0)$, which completes the proof that $\tilde{V}_{R,\delta}(x_0) = V_{R,\delta}(x_0)$. The final claim follows from a repetition of the arguments leading to (45) and (46). This completes the proof of of Proposition 13. ■

The following two lemmas contain the primary use of the nondegeneracy assumption, which states for every $\theta \in (\mathcal{M} \cap U)$ that

$$\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d} = \operatorname{codim}(\mathcal{M} \cap U). \tag{49}$$

The first of these proves that ∇f can be split into a component that is approximately normal to the local manifold of minima $\mathcal{M} \cap U$, and into a component that is approximately tangential to $\mathcal{M} \cap U$.

Lemma 14 *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 1. Then for every $x_0 \in (\mathcal{M} \cap U)$ there exist $R_0, \delta_0, c \in (0, \infty)$ and $V \in \operatorname{Proj}(x_0)$ (cf. Definition 8) such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ we have that (cf. Definition 12)*

$$\overline{V}_{R,\delta}(x_0) \subseteq V, \tag{50}$$

and for every $x \in V_{R,\delta}(x_0)$ there exists $\varepsilon_x \in \mathbb{R}^d$ which satisfies $|\varepsilon_x| \leq c\mathbf{d}(x, \mathcal{M} \cap U)^2$ such that

$$\nabla f(x) = \big(\operatorname{Hess} f\big)(p(x)) \cdot (x - p(x)) + \varepsilon_x. \tag{51}$$

Proof [Proof of Lemma 14] Let $x_0 \in (\mathcal{M} \cap U)$ and $R \in (0, \infty)$. Since $U \subseteq \mathbb{R}^d$ is an open set, there exists $V \in \text{Proj}(x_0)$ which satisfies that $V \subseteq U$. Since V is open, fix $R_0, \delta_0 \in (0, \infty)$ such that, for every $R \in (0, R_0]$ and $\delta \in (0, \delta_0]$,

$$\overline{V}_{R,\delta}(x_0) \subseteq V. \tag{52}$$

Due to the compactness of $\overline{V}_{R,\delta}(x_0)$ and the regularity of f, there exists $c \in (0, \infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$,

$$\|f\|_{\mathrm{C}^3(V_{R,\delta}(x_0))} = \sup_{0 \leq k \leq 3} \left\| \nabla^k f \right\|_{L^\infty(V_{R,\delta_0}(x_0); \mathbb{R}^{(d^k)})} \leq c. \tag{53}$$

Let $x \in V_{R,\delta}(x_0)$. By integration, since $\nabla f|_{\mathcal{M} \cap U} = 0$,

$$\begin{aligned} \nabla f(x) &= \int_0^1 \big(\text{Hess}\, f\big)(p(x) + s(x - p(x))) \cdot (x - p(x)) \, \mathrm{d}s \\ &= \big(\text{Hess}\, f\big)(p(x)) \cdot (x - p(x)) \\ &\quad + \int_0^1 \Big(\big(\text{Hess}\, f\big)(p(x) + s(x - p(x))) - \big(\text{Hess}\, f\big)(p(x))\Big) \cdot (x - p(x)) \, \mathrm{d}s. \end{aligned} \tag{54}$$

It follows from (53), the local regularity of f, and the definition of the projection that there exists $c \in (0, \infty)$ which satisfies

$$\begin{aligned} \left| \int_0^1 \Big(\big(\text{Hess}\, f\big)(p(x) + s(x - p(x))) - \big(\text{Hess}\, f\big)(p(x))\Big) \cdot (x - p(x)) \, \mathrm{d}s \right| &\leq c\mathbf{d}(x, \mathcal{M} \cap U)^2 \int_0^1 s \, \mathrm{d}s \\ &\leq c\mathbf{d}(x, \mathcal{M} \cap U)^2. \end{aligned} \tag{55}$$

Let $\varepsilon_x \in \mathbb{R}^d$ be defined by

$$\varepsilon_x = \int_0^1 \Big(\big(\text{Hess}\, f\big)(p(x) + s(x - p(x))) - \big(\text{Hess}\, f\big)(p(x))\Big) \cdot (x - p(x)) \, \mathrm{d}s. \tag{56}$$

Equation (54) and estimate (55) complete the proof of Lemma 14. ∎

Lemma 15 *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 1. Then for every $x_0 \in (\mathcal{M} \cap U)$ there exist $R_0, \delta_0, \mathfrak{r}, \in (0, \infty)$, $\lambda \in (0, \infty)$, and $V \in \text{Proj}(x_0)$ (cf. Definition 8) satisfying the following four properties.*

(i) We have that

$$\lambda \leq \max_{x \in \mathcal{M} \cap U \cap \overline{B}_R(x_0)} \left| \big(\text{Hess}\, f\big)(x) \right|. \tag{57}$$

(ii) For every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $x \in V_{R,\delta}(x_0)$,

$$\overline{V}_{R,\delta}(x_0) \subseteq V. \tag{58}$$

(iii) For every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $x \in V_{R,\delta}(x_0)$,

$$\begin{aligned} \mathbf{d}\left(x - r\big(\text{Hess}\, f\big)(p(x)) \cdot (x - p(x)), \mathcal{M} \cap U\right) &\leq \left|(x - p(x)) - r\big(\text{Hess}\, f\big)(p(x)) \cdot (x - p(x))\right| \\ &\leq (1 - \lambda r)\, \mathbf{d}(x, \mathcal{M} \cap U). \end{aligned} \tag{59}$$

(iv) For every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $x \in V_{R,\delta}(x_0)$,

$$\big(\big(\operatorname{Hess} f\big)(p(x)) \cdot (x - p(x))\big) \cdot (x - p(x)) \geq \lambda \mathbf{d}(x, \mathcal{M} \cap U)^2. \tag{60}$$

Proof [Proof of Lemma 15] Let $x_0 \in (\mathcal{M} \cap U)$. Since $U \subseteq \mathbb{R}^d$ is an open subset, there exists $V \in \operatorname{Proj}(x_0)$ which satisfies that $V \subseteq U$. Fix $R_0, \delta_0 \in (0, \infty)$ such that every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ we have that (cf. Definition 12)

$$\overline{V}_{R,\delta}(x_0) \subseteq V. \tag{61}$$

Due to the compactness of $\overline{V}_{R_0,\delta_0}(x_0)$ and the regularity of f, there exists $c \in (0, \infty)$ which satisfies for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ that

$$\|f\|_{C^3(V_{R,\delta}(x_0))} \leq c. \tag{62}$$

Let $x \in V_{R,\delta}(x_0)$. For the first claim, using (62), fix $\mathfrak{r} \in (0, \infty)$ which satisfies that

$$\mathfrak{r}\left(\max_{x \in V_{R_0,\delta_0}(x_0)} \left|\big(\operatorname{Hess} f\big)(p(x))\right|\right) \leq 1. \tag{63}$$

Let $r \in (0, \mathfrak{r}]$. The definition of the distance to $\mathcal{M} \cap U$ implies that

$$\begin{aligned} &\mathbf{d}\left(x - r\big(\operatorname{Hess} f\big)(p(x)) \cdot (x - p(x)), \mathcal{M} \cap U\right) \\ &\leq \left|(x - p(x)) - r\big(\operatorname{Hess} f\big)(p(x)) \cdot (x - p(x))\right|. \end{aligned} \tag{64}$$

Since the nondegeneracy assumption states that

$$\operatorname{rank}((\operatorname{Hess} f)(p(x))) = d - \mathfrak{d} = \operatorname{codim}(\mathcal{M} \cap U), \tag{65}$$

Lemma 10 below and (62) prove that there exists for $\lambda \in (0, \infty)$ which satisfies that

$$\lambda \leq \max_{x \in \mathcal{M} \cap U \cap \overline{B}_R(x_0)} \left|\big(\operatorname{Hess} f\big)(p(x))\right|, \tag{66}$$

for which we have that

$$\left|(x - p(x)) - r\big(\operatorname{Hess} f\big)(p(x)) \cdot (x - p(x))\right| \leq (1 - r\lambda)\,|x - p(x)| = (1 - r\lambda)\mathbf{d}(x, \mathcal{M} \cap U), \tag{67}$$

where the choice of $\mathfrak{r}$ and (66) guarantee that $(1 - r\lambda) \geq 0$. In combination, estimates (64), (66), and (67) complete the proof of the first claim. The proof of the second claim is similar. For every $x \in V_{R,\delta}(x_0)$, the nondegeneracy assumption, Lemma 10, and (62) prove that there exists $\lambda \in (0, \infty)$ which satisfies (66) such that

$$\big(\big(\operatorname{Hess} f\big)(p(x)) \cdot (x - p(x))\big) \cdot (x - p(x)) \geq \lambda\,|x - p(x)|^2 = \lambda \mathbf{d}(x, \mathcal{M} \cap U)^2, \tag{68}$$

which completes the proof of Lemma 15. ■

3. Continuous deterministic gradient descent

In this section, for an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ which satisfies Assumption 1, we will analyze the local convergence to the local manifold of minima $\mathcal{M} \cap U$ of the deterministic gradient descent algorithm in continuous time

$$\frac{d}{dt}\theta_t = -\nabla f(\theta_t). \tag{69}$$

In Proposition 16 below, we prove the existence of a basin of attraction for (69) to the local manifold of minima $\mathcal{M} \cap U$. Provided the initial data is taken in this basin of attraction, the solution (69) converges with an exponential rate.

Proposition 16 *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 1. Then for every $x_0 \in (\mathcal{M} \cap U)$ there exist $R_0, \delta_0, \lambda \in (0,\infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, and $\theta_0 \in V_{R/2,\delta}(x_0)$ (cf. Definition 12), for $\{\theta_t\}_{t\in[0,\infty)}$ the solution of*

$$\frac{d}{dt}\theta_t = -\nabla f(\theta_t), \tag{70}$$

we have, for every $t \in [0,\infty)$,

$$\mathbf{d}(\theta_t, \mathcal{M}\cap U) \leq \exp(-\lambda t)\mathbf{d}(\theta_0, \mathcal{M}\cap U). \tag{71}$$

Proof [Proof of Proposition 16] Let $x_0 \in (\mathcal{M}\cap U)$. Since $U \subseteq \mathbb{R}^d$ is an open set, fix $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. In view or Proposition 13, fix $R_0, \delta_0 \in (0,\infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0,\delta_0]$ the set $V_{R,\delta}(x_0)$ (cf. Definition 12) satisfies that $\overline{V}_{R,\delta}(x_0) \subseteq V$ and that

$$V_{R,\delta}(x_0) = \{x \in \mathbb{R}^d \colon \mathbf{d}(x, \mathcal{M}\cap U) = \mathbf{d}(x, \overline{B}_R(x_0)\cap\mathcal{M}\cap U) < \delta\}. \tag{72}$$

In particular, the compactness of $\overline{V}_{R_0,\delta_0}(x_0)$ and the regularity of f imply that there exists $c \in (0,\infty)$ which satisfies that

$$\|f\|_{\mathrm{C}^3(V_{R_0,\delta_0}(x_0))} \leq c. \tag{73}$$

Let $R \in (0,R_0]$, $\delta \in (0,\delta_0]$. Let $\theta_0 \in V_{R/2,\delta}(x_0)$, let $\theta_t \in \mathbb{R}^d$, $t \in [0,\infty)$, satisfy for every $t \in (0,\infty)$ that

$$\frac{d}{dt}\theta_t = -\nabla f(\theta_t), \tag{74}$$

and let $\tau \in (0,\infty)$ denote the exit time

$$\tau = \inf\{\, t \geq 0 \mid \theta_t \notin V_{R,\delta}(x_0)\,\}. \tag{75}$$

Lemma 11 and the chain rule prove that

$$\begin{cases} \dfrac{d}{dt}\mathbf{d}(\theta_t, \mathcal{M}\cap U) = -\nabla f(\theta_t)\cdot\nabla\mathbf{d}(\theta_t,\mathcal{M}\cap U) = -\nabla f(\theta_t)\cdot\dfrac{\theta_t - p(\theta_t)}{|\theta_t - p(\theta_t)|} & \text{in } (0,\tau), \\ \qquad \dfrac{d}{dt}p(\theta_t) = -Dp(\theta_t)\cdot\nabla f(\theta_t) & \text{in } (0,\tau), \end{cases} \tag{76}$$

where the local regularity of f and the stopping time τ guarantee the well-posedness of this equation. Let $t \in (0,\tau)$. It follows from Lemma 14 and Lemma 15 that there exist $\lambda, c_1 \in (0,\infty)$ which satisfy that

$$\nabla f(\theta_t)\cdot\frac{\theta_t - p(\theta_t)}{|\theta_t - p(\theta_t)|} \geq \lambda\mathbf{d}(\theta_t,\mathcal{M}\cap U) - c_1\mathbf{d}(\theta_t,\mathcal{M}\cap U)^2. \tag{77}$$

Proposition 7, (73), and $\nabla f|_{\mathcal{M}\cap U} = 0$ prove that there exists $c_2 \in (0,\infty)$ which satisfies that

$$|Dp(\theta_t)\cdot\nabla f(\theta_t)| \leq c_2\mathbf{d}(\theta_t,\mathcal{M}\cap U). \tag{78}$$

Returning to (76), it follows from (77) and (78) that

$$\begin{cases} \dfrac{d}{dt}\mathbf{d}(\theta_t,\mathcal{M}\cap U) \leq -\lambda\mathbf{d}(\theta_t,\mathcal{M}\cap U) + c_1\mathbf{d}(\theta_t,\mathcal{M}\cap U)^2 & \text{in } (0,\tau), \\ \qquad \left|\dfrac{d}{dt}p(\theta_t)\right| \leq c_2\mathbf{d}(\theta_t,\mathcal{M}\cap U) & \text{in } (0,\tau). \end{cases} \tag{79}$$

Let $\delta_1 \in (0, \delta_0]$ satisfy that

$$c_1 \delta_1 \leq \lambda/2. \tag{80}$$

Let $\delta \in (0, \delta_1]$. For every $t \in (0, \tau)$ it follows from (79) and (80) that

$$\frac{d}{dt}\mathbf{d}(\theta_t, \mathcal{M} \cap U) \leq -\frac{\lambda}{2}\mathbf{d}(\theta_t, \mathcal{M} \cap U). \tag{81}$$

Therefore, for every $\delta \in (0, \delta_1]$, $t \in [0, \tau)$ it holds that

$$\mathbf{d}(\theta_t, \mathcal{M} \cap U) \leq \mathbf{d}(\theta_0, \mathcal{M} \cap U) \exp(-\lambda t/2) \leq \delta_1 \exp(-\lambda t/2). \tag{82}$$

For every $t \in [0, \tau)$, it follows from (79) and (82) that

$$\max_{0 \leq t \leq \tau} |p(\theta_t) - p(\theta_0)| \leq c_2 \int_0^\tau \delta_1 \exp\left(-\frac{\lambda t}{2}\right) \mathrm{d}t = \frac{2c_2\delta_1}{\lambda}\left(1 - \exp\left(-\frac{\lambda \tau}{2}\right)\right) \leq \frac{2c_2\delta_1}{\lambda}. \tag{83}$$

Fix $\delta_2 \in (0, \delta_1]$ which satisfies that

$$\frac{2c_2\delta_2}{\lambda} < \frac{R}{2}. \tag{84}$$

Let $\delta \in (0, \delta_2]$. In combination (82), (83), $\theta_0 \in V_{R/2,\delta}(x_0)$, and the triangle inequality prove that $\theta_t \in V_{R,\delta}(x_0)$ for every $t \in (0, \infty)$. This is to say that $\tau = \infty$. Since $\theta_0 \in V_{R/2,\delta}(x_0)$ was arbitrary, this completes the proof of Proposition 16. ∎

4. Discrete deterministic gradient descent

In this section, for an objective function $f \colon \mathbb{R}^d \to \mathbb{R}$ which satisfies Assumption 1, we will analyze the convergence of the deterministic gradient descent algorithm $\{\theta_n\}_{n \in \mathbb{N}_0}$ defined for every $n \in \mathbb{N}$ by

$$\theta_n = \theta_{n-1} - \frac{r}{n^\rho}\nabla f(\theta_{n-1}), \tag{85}$$

for some learning rate $\rho \in (0, 1)$ and $r \in (0, \infty)$. The proof is similar to the case of the deterministic gradient descent algorithm in continuous time. However, in the discrete setting, care must be taken to choose $r \in (0, \infty)$ sufficiently small. Since, if r is too large, for small values of n the jump $-\frac{r}{n^\rho}\nabla f$ may be an overcorrection that causes the solution to overshoot the local manifold of minima and to leave the basin of attraction.

Proposition 17 *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 1. Then for every $x_0 \in (\mathcal{M} \cap U)$ there exists $R_0, \delta_0, \mathfrak{r}, c \in (0, \infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $\rho \in (0, 1)$, and $\theta_0 \in V_{R/2,\delta}(x_0)$ (cf. Definition 12), for $\{\theta_n\}_{n \in \mathbb{N}_0}$ defined by*

$$\theta_n = \theta_{n-1} - \frac{r}{n^\rho}\nabla f(\theta_{n-1}), \tag{86}$$

we have, for every $n \in \mathbb{N}_0$,

$$\mathbf{d}(\theta_n, \mathcal{M} \cap U) \leq \exp(-cn^{1-\rho})\mathbf{d}(x_0, \mathcal{M} \cap U). \tag{87}$$

Proof [Proof of Proposition 17] Let $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (0, 1)$. Since $U \subseteq \mathbb{R}^d$ is open, fix $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. In view or Proposition 13, fix

$R_0, \delta_0 \in (0,\infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ the set $V_{R,\delta}(x_0)$ (cf. Definition 12) satisfies that $\overline{V}_{R,\delta}(x_0) \subseteq V$ and that

$$V_{R,\delta}(x_0) = \{x \in \mathbb{R}^d \colon \mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta\}. \tag{88}$$

The regularity of f and the compactness of $\overline{V}_{R_0,\delta_0}(x_0)$ prove that there exists $c \in (0,\infty)$ which satisfies that

$$\|f\|_{\mathrm{C}^3(V_{R_0,\delta_0}(x_0))} \leq c. \tag{89}$$

Fix $\mathfrak{r} \in (0,\infty)$ which satisfies the conclusion of Lemma 15 for the set $V_{R_0,\delta_0}(x_0)$. Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$. Let $\theta_0 \in V_{R/2,\delta}(x_0)$, let $\theta_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, satisfy that

$$\theta_n = \theta_{n-1} - \frac{r}{n^\rho} \nabla f(\theta_{n-1}), \tag{90}$$

and let $\tau \in \mathbb{N}$ be the exit time which satisfies that

$$\tau = \inf\{\, n \in \mathbb{N} \mid \theta_n \notin V_{R,\delta}(x_0) \,\}. \tag{91}$$

Since for every $n \in \{1, \ldots, \tau\}$ the projection of θ_{n-1} is well-defined, we have that

$$\mathbf{d}(\theta_n, \mathcal{M} \cap U) \leq |\theta_n - p(\theta_{n-1})| = \left|\theta_{n-1} - p(\theta_{n-1}) - \frac{r}{n^\rho} \nabla f(\theta_{n-1})\right|. \tag{92}$$

Lemma 14 proves that there exists $c \in (0,\infty)$ such that for every $n \in \{1, \ldots, \tau\}$ there exists $\varepsilon_n \in \mathbb{R}^d$ which satisfies that

$$|\varepsilon_n| \leq c \mathbf{d}(\theta_{n-1}, \mathcal{M} \cap U)^2, \tag{93}$$

such that

$$\nabla f(\theta_{n-1}) = \big(\operatorname{Hess} f\big)(p(\theta_{n-1})) \cdot (x - p(x)) + \varepsilon_n. \tag{94}$$

The triangle inequality, (92), (93), and (94) prove that there exists $c_1 \in (0,\infty)$ such that for every $n \in \{1, \ldots, \tau\}$ it holds that

$$\begin{aligned} \mathbf{d}(\theta_n, \mathcal{M} \cap U) \leq{}& \left|\theta_{n-1} - p(\theta_{n-1}) - \frac{r}{n^\rho}\big(\operatorname{Hess} f\big)(p(\theta_{n-1})) \cdot (\theta_{n-1} - p(\theta_{n-1}))\right| \\ &+ \frac{c_1 r}{n^\rho} \mathbf{d}(\theta_{n-1}, \mathcal{M} \cap U)^2. \end{aligned} \tag{95}$$

Finally, the choice of $\mathfrak{r} \in (0,\infty)$, Lemma 15, and (95) prove that there exists $\lambda \in (0,\infty)$ such that for every $n \in \{1, \ldots, \tau\}$ it holds that

$$\mathbf{d}(\theta_n, \mathcal{M} \cap U) \leq \left(1 - \frac{r\lambda}{n^\rho}\right) \mathbf{d}(\theta_{n-1}, \mathcal{M} \cap U) + \frac{c_1 r}{n^\rho} \mathbf{d}(\theta_{n-1}, \mathcal{M} \cap U)^2, \tag{96}$$

where the choice of $\mathfrak{r} \in (0,\infty)$ guarantees that $(1 - r\lambda) \geq 0$. Fix $\delta_1 \in (0, \delta_0]$ which satisfies that

$$c_1 \delta_1 \leq \frac{\lambda}{2}. \tag{97}$$

Let $\delta \in (0, \delta_1]$. It follows from (96) and (97) that for every $n \in \{1, \ldots, \tau\}$ it holds that

$$\mathbf{d}(\theta_n, \mathcal{M} \cap U) \leq \left(1 - \frac{r\lambda}{2n^\rho}\right) \mathbf{d}(\theta_{n-1}, \mathcal{M} \cap U). \tag{98}$$

After iterating this inequality, we have for every $n \in \{1, \ldots, \tau\}$ that

$$\mathbf{d}(\theta_n, \mathcal{M} \cap U) \leq \prod_{k=1}^{n} \left(1 - \frac{r\lambda}{2k^\rho}\right) \mathbf{d}(\theta_0, \mathcal{M} \cap U). \tag{99}$$

Since there exists $c \in (0,\infty)$ which satisfies for every $n \in \mathbb{N}$ that

$$\log\left(\prod_{k=1}^{n}\left(1-\frac{r\lambda}{2k^{\rho}}\right)\right) = \sum_{k=1}^{n}\log\left(1-\frac{r\lambda}{2k^{\rho}}\right) \leq -c\sum_{k=1}^{n}\frac{r\lambda}{2k^{\rho}} \leq -c\frac{r\lambda}{2}n^{1-\rho}, \tag{100}$$

it follows from (99) that there exists $c_2 \in (0,\infty)$ which satisfies for every $n \in \{1,\ldots,\tau\}$ that

$$\mathbf{d}(\theta_n, \mathcal{M}\cap U) \leq \exp\left(-c_2 n^{1-\rho}\right)\mathbf{d}(\theta_0, \mathcal{M}\cap U). \tag{101}$$

It remains only to show that, provided $\delta \in (0,\delta_1]$ is chosen sufficiently small, we have that $\tau = \infty$. It follows from (89), (101), and $\nabla f|_{\mathcal{M}\cap U} = 0$ that there exists $c \in (0,\infty)$ which satisfies that

$$|\theta_n - \theta_{n-1}| = \frac{r}{n^{\rho}}\left|\nabla f(\theta_{n-1})\right| \leq \frac{c}{n^{\rho}}\mathbf{d}(\theta_{n-1}, \mathcal{M}\cap U) \leq cn^{-\rho}\exp\left(-c_2 n^{1-\rho}\right)\mathbf{d}(\theta_0, \mathcal{M}\cap U). \tag{102}$$

The triangle inequality therefore implies that there exists $c_3 \in (0,\infty)$ such that for every $n \in \{1,\ldots,\tau\}$ it holds that

$$|\theta_n - \theta_0| \leq c\mathbf{d}(\theta_0, \mathcal{M}\cap U)\sum_{k=1}^{\infty} ck^{-\rho}\exp\left(-c_2 k^{1-\rho}\right) = c_3\mathbf{d}(\theta_0, \mathcal{M}\cap U) < \infty. \tag{103}$$

Fix $\delta_2 \in (0,\delta_1]$ which satisfies that

$$c_3\delta_2 < \frac{R}{2} - 2\delta_2. \tag{104}$$

Let $\delta \in (0,\delta_2]$. The choice of $\delta_2 \in (0,\delta_1]$, (103), and the triangle inequality prove for every $n \in \{1,\ldots,\tau\}$ that

$$|\theta_n - x_0| \leq |\theta_n - \theta_0| + |\theta_0 - x_0| < c_3\delta_2 + \frac{R}{2} + \delta_2 < R - \delta_2. \tag{105}$$

In combination (101) and (105) prove for every $n \in \{1,\ldots,\tau\}$ that

$$\mathbf{d}(\theta_n, \mathcal{M}\cap U) < \delta_2 \text{ and } |\theta_n - x_0| \leq R - \delta_2. \tag{106}$$

The triangle inequality therefore implies for every $n \in \{1,\ldots,\tau\}$ that

$$\mathbf{d}(\theta_n, \mathcal{M}\cap U) = \mathbf{d}(\theta_n, \overline{B}_R(x_0)\cap\mathcal{M}\cap U). \tag{107}$$

It follows from Proposition 13, the choice of $R_0, \delta_0 \in (0,\infty)$, and $\theta_0 \in V_{R/2,\delta}(x_0)$ that for every $n \in \mathbb{N}$ it holds that $\theta_n \in V_{R,\delta}(x_0)$. This is to say that $\tau = \infty$, which completes the proof of Proposition 17. ■

Remark 18 *The conclusion of Proposition* 17 *can be extended to the case of* $\rho = 1$ *using the same techniques. In this case, in the setting of Proposition* 17*, we would obtain that*

$$\mathbf{d}(\theta_n, \mathcal{M}\cap U) \leq \exp(-c\log(n))\mathbf{d}(x_0, \mathcal{M}\cap U). \tag{108}$$

The logarithm appears in estimate (100)*, and the remainder of the proof is identical.*

5. Stochastic gradient descent

In this section, for a jointly measurable function F and for noise $X_{1,1}$ that satisfy Assumption 2, for i.i.d. random variables $\{X_{n,m}\}_{n,m\in\mathbb{N}}$, and for the objective function $f(\cdot) = \mathbb{E}[F(\cdot, X_{1,1})]$ that satisfies Assumption 1, we prove the convergence to the manifold of minima of the stochastic gradient descent algorithm

$$\Theta_n = \Theta_{n-1} - \frac{r}{Mn^{\rho}} \sum_{m=1}^{M} \nabla_\theta F(\Theta_{n-1}, X_{n,m}), \tag{109}$$

for mini-batch size $M \in \mathbb{N}$, learning rate $\rho \in (2/3, 1)$, and $r \in (0, \infty)$. The role of the mini-batch size $M \in \mathbb{N}$ is to reduce the variance of the random gradient

$$\frac{1}{M} \sum_{m=1}^{M} \nabla_\theta F(\Theta_{n-1}, X_{n,m}). \tag{110}$$

The variance reduction is quantified by the following well-known lemma, where the function G plays the role of $\nabla_\theta F$.

Lemma 19 *Let $d_1, d_2 \in \mathbb{N}$, let $U \subseteq \mathbb{R}^{d_1}$ be a non-empty open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $G\colon \mathbb{R}^{d_1} \times S \to \mathbb{R}^{d_2}$ be a measurable function, and let $\{X_m\colon \Omega \to S\}_{m\in\mathbb{N}}$ be i.i.d. random variables. Assume that G and X_1 satisfy Assumption 2. Then for every non-empty compact set $\mathfrak{C} \subseteq U$ there exists $c \in (0, \infty)$ such that, for every $M \in \mathbb{N}$,*

$$\sup_{\theta\in\mathfrak{C}} \left(\mathbb{E}\left[\left| \left[\frac{1}{M} \sum_{m=1}^{M} G(\theta, X_m) \right] - \mathbb{E}\big[G(\theta, X_1)\big] \right|^2 \right] \right) \leq \frac{c}{M}. \tag{111}$$

Proof [Proof of Lemma 19] Let $\mathfrak{C} \subseteq U$ be a compact set. For every $\theta \in \mathfrak{C}$, $M \in \mathbb{N}$,

$$\begin{aligned} &\mathbb{E}\Big[\Big| \frac{1}{M} \sum_{m=1}^{M} G(\theta, X_m) - \mathbb{E}\big[G(\theta, X_1)\big] \Big|^2 \Big] \\ &= \frac{1}{M^2} \sum_{i,j=1}^{M} \mathbb{E}\Big[\big(G(\theta, X_i) - \mathbb{E}\big[G(\theta, X_1)\big]\big) \big(G(\theta, X_j) - \mathbb{E}\big[G(\theta, X_1)\big]\big) \Big]. \end{aligned} \tag{112}$$

Since the $\{X_m\}_{m\in\mathbb{N}}$ are i.i.d. and since $\{G(\theta, X_{1,1})\}_{\theta\in\mathbb{R}^{d_1}}$ is locally bounded in $L^2(\Omega; \mathbb{R}^{d_2})$, there exists $c \in (0, \infty)$ such that, for every $M \in \mathbb{N}$,

$$\begin{aligned} &\sup_{\theta\in\mathfrak{C}} \Big(\mathbb{E}\Big[\Big| \frac{1}{M} \sum_{m=1}^{M} G(\theta, X_m) - \mathbb{E}\big[G(\theta, X_1)\big] \Big|^2 \Big] \Big) \\ &= \sup_{\theta\in\mathfrak{C}} \Big(\frac{1}{M^2} \sum_{m=1}^{M} \mathbb{E}\Big[\Big| G(\theta, X_m) - \mathbb{E}\big[G(\theta, X_1)\big] \Big|^2 \Big] \Big) \\ &= \frac{1}{M} \sup_{\theta\in\mathfrak{C}} \Big(\mathbb{E}\Big[\Big| G(\theta, X_1) - \mathbb{E}\big[G(\theta, X_1)\big] \Big|^2 \Big] \Big) \\ &\leq \frac{c}{M}. \end{aligned} \tag{113}$$

This completes the proof of Lemma 19. ∎

In the following proposition, we establish the convergence of SGD in directions normal to the local manifold of minima on the event that SGD begins in and remains in a basin of attraction.

Proposition 20 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, let $\{X_{n,m}\colon \Omega \to \mathbb{R}\}_{n,m\in\mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1}$ satisfy Assumption 2. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1})]$ and assume that f satisfies Assumption 1. For every $M \in \mathbb{N}$, $\rho \in (2/3, 1)$, $r \in (0, \infty)$, $\theta \in \mathbb{R}^d$ let $\{\Theta_{n,\theta} = \Theta_{n,\theta}(M, \rho, r)\}_{n\in\mathbb{N}_0}$ be defined by $\Theta_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,\theta}, X_{n,m})\right], \tag{114}$$

and for every $R, \delta \in (0, \infty)$, $x_0 \in (\mathcal{M} \cap U)$, and $n \in \mathbb{N}$ let $A_n = A_n(M, r, \rho, \theta, R, \delta, x_0) \in \mathcal{F}$ be defined by

$$A_n = \Big\{\forall\, m \in \{0, \ldots, n\}\ \Theta_{m,\theta} \in V_{R,\delta}(x_0) \ \ \textit{(cf. Definition 12)}\Big\}. \tag{115}$$

Then for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (2/3, 1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0, \infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$, and $\theta \in V_{R,\delta}(x_0)$ (cf. Definition 12),

$$\left(\mathbb{E}\left[(\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \wedge 1)^2\, \mathbf{1}_{A_{n-1}}\right]\right)^{\frac{1}{2}} \leq c n^{-\frac{\rho}{2}}. \tag{116}$$

Proof [Proof of Proposition 20] Let $x_0 \in (\mathcal{M} \cap U)$. Since $U \subseteq \mathbb{R}^d$ is open, fix $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. Fix $R_0, \delta_0 \in (0, \infty)$ which satisfy the conclusion of Proposition 13 for this set V. Finally, fix $\mathfrak{r} \in (0, \infty)$ which satisfies the conclusion of Lemma 15. Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$. To simplify the notation, and by a small abuse of notation, let $\nabla_\theta F^{M,n}\colon \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be the functions which satisfy for every $(\theta, \omega) \in \mathbb{R}^d \times \Omega$ that

$$\nabla_\theta F^{M,n}(\theta) = \nabla_\theta F^{M,n}(\theta, \omega) = \frac{1}{M}\sum_{m=1}^{M} (\nabla_\theta F)(\theta, X_{n,m}(\omega)). \tag{117}$$

Let $\theta \in V_{R,\delta}(x_0)$, let $\Theta_{0,\theta}\colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\Theta_{0,\theta}(\omega) = \theta$, and for every $n \in \mathbb{N}$ let $\Theta_{n,\theta}\colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho}\nabla_\theta F^{M,n}(\Theta_{n-1,\theta}). \tag{118}$$

We will analyze the solution $\Theta_{n,\theta}$ of (118) on the event A_{n-1}. We observe that

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta}) + \frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right). \tag{119}$$

Since the event A_{n-1} implies that $\Theta_{n-1,\theta} \in V_{R,\delta}(x_0) \subseteq V$, the projection of $\Theta_{n-1,\theta}$ is well-defined and it holds by definition of the distance to $\mathcal{M} \cap U$ that

$$\begin{aligned}
&\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U)^2 \\
&\leq |\Theta_{n,\theta} - p(\Theta_{n-1,\theta})|^2 \\
&\leq \left|\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}) - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta})\right|^2 \\
&\quad + 2\left(\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}) - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta})\right)\cdot \frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right) \\
&\quad + \left|\frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right)\right|^2.
\end{aligned} \tag{120}$$

The three terms on the righthand side of (120) will be treated separately. For the first term on the righthand side of (120), the choice of $\mathfrak{r} \in (0, \infty)$, Lemma 14, and Lemma 15 prove,

following identically the proof leading from (92) to (96), that there exist $\lambda, c \in (0,\infty)$ such that

$$\begin{aligned}&\left|\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}) - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta})\right| \\ &\leq \left(1 - \frac{r\lambda}{n^\rho}\right)\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U) + c\frac{r}{n^\rho}\mathbf{d}(\Theta_{n-1,\theta},\mathcal{M}\cap U)^2.\end{aligned} \tag{121}$$

Therefore, there exist $\lambda, c \in (0,\infty)$ which satisfy that

$$\begin{aligned}\left|\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}) - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta})\right|^2 \leq &\left(1 - \frac{r\lambda}{n^\rho}\right)^2 \mathbf{d}(\Theta_{n-1,\theta},\mathcal{M}\cap U)^2 \\ &+ c\left(1 - \frac{r\lambda}{n^\rho}\right)\frac{r}{n^\rho}\mathbf{d}(\Theta_{n-1,\theta},\mathcal{M}\cap U)^3 \\ &+ c\frac{r^2}{n^{2\rho}}\mathbf{d}(\Theta_{n-1,\theta},\mathcal{M}\cap U)^4.\end{aligned} \tag{122}$$

The remaining two terms of (120) and the righthand side of (122) will be handled after taking the expectation on the event $A_{n-1} \subseteq \Omega$ which satisfies that

$$A_{n-1} = \big\{\omega \in \Omega\colon \Theta_{m,\theta} \in V_{R,\delta}(x_0)\ \forall\, m \in \{0,\dots,n-1\}\big\}. \tag{123}$$

After returning to (120), it follows from (122) that there exists $c \in (0,\infty)$ which satisfies that

$$\begin{aligned}&\mathbb{E}\left[\mathbf{d}(\Theta_{n,\theta},\mathcal{M}\cap U)^2\mathbf{1}_{A_{n-1}}\right] \\ &\leq \left(1 - \frac{r\lambda}{n^\rho}\right)^2 \mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta},\mathcal{M}\cap U)^2\mathbf{1}_{A_{n-1}}\right] \\ &\quad + c\left(1 - \frac{r\lambda}{n^\rho}\right)\frac{r}{n^\rho}\mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta},\mathcal{M}\cap U)^3\mathbf{1}_{A_{n-1}}\right] + c\frac{r^2}{n^{2\rho}}\mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta},\mathcal{M}\cap U)^4\mathbf{1}_{A_{n-1}}\right] \\ &\quad + 2\mathbb{E}\left[\left(\Theta_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta}) - p(\Theta_{n-1,\theta})\right)\cdot\frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right)\mathbf{1}_{A_{n-1}}\right] \\ &\quad + \mathbb{E}\left[\left|\frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right)\right|^2\mathbf{1}_{A_{n-1}}\right].\end{aligned} \tag{124}$$

For every $m \in \mathbb{R}$ let $\mathcal{F}_m \subseteq \mathcal{F}$ be the sigma algebra defined by

$$\mathcal{F}_m = \sigma\big(\{X_{1,k}\}_{k=1}^M, \dots, \{X_{m,k}\}_{k=1}^M\big). \tag{125}$$

For the penultimate term of (124), since $\mathbf{1}_{A_{n-1}}$ is $\mathcal{F}_{n-1}$-measurable, properties of the conditional expectation imply that

$$\begin{aligned}&\mathbb{E}\Big[\Big(\Theta_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta}) - p(\Theta_{n-1,\theta})\Big)\frac{r}{n^\rho}\Big(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\Big)\mathbf{1}_{A_{n-1}}\Big] = \\ &\mathbb{E}\Big[\mathbb{E}\Big[\Big(\Theta_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta}) - p(\Theta_{n-1,\theta})\Big)\frac{r}{n^\rho}\Big(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\Big)\mathbf{1}_{A_{n-1}}|\mathcal{F}_{n-1}\Big]\Big].\end{aligned} \tag{126}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\Big[\Big(\Theta_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta}) - p(\Theta_{n-1,\theta})\Big)\frac{r}{n^\rho}\Big(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\Big)\mathbf{1}_{A_{n-1}}\Big] = \\ &\mathbb{E}\Big[\Big(\Theta_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta_{n-1,\theta}) - p(\Theta_{n-1,\theta})\Big)\mathbf{1}_{A_{n-1}}\mathbb{E}\Big[\frac{r}{n^\rho}\Big(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\Big)|\mathcal{F}_{n-1}\Big]\Big] \\ &= 0,\end{aligned} \tag{127}$$

where the final equality follows from the fact that the $X_{m,k}$, $m,k \in \mathbb{N}$, are independent and therefore satisfy for every $x \in \mathbb{R}^d$ that

$$\mathbb{E}\left[\frac{r}{n^\rho}\left(\nabla f(x) - \nabla_\theta F^{M,n}(x)\right)|\mathcal{F}_{n-1}\right] = \frac{r}{Nn^\rho}\sum_{m=1}^{M}\mathbb{E}\left[\nabla f(x) - \nabla_\theta F(x, X_{n,m})\right] = 0. \tag{128}$$

The final term of (124) is handled using Lemma 19. Since $\overline{V}_{R,\delta}(x_0)$ is compact, the independence of the $X_{m,k}$, $m,k \in \mathbb{N}$, and Lemma 19 prove that there exists $c \in (0,\infty)$ such that

$$\mathbb{E}\left[\left|\frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right)\mathbf{1}_{A_{n-1}}\right|^2\right] \leq \frac{cr^2}{Mn^{2\rho}}. \tag{129}$$

Returning to (124), it follows from (127) and (129) that there exists $c_1 \in (0,\infty)$ such that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)^2\mathbf{1}_{A_{n-1}}\right] \leq \\
&\left(1 - \frac{r\lambda}{n^\rho}\right)^2\mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)^2\mathbf{1}_{A_{n-1}}\right] + c_1\left(1 - \frac{r\lambda}{n^\rho}\right)\frac{r}{n^\rho}\left[\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)^3\mathbf{1}_{A_{n-1}}\right] \\
&+ c_1\frac{r^2}{n^{2\rho}}\mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)^4\mathbf{1}_{A_{n-1}}\right] + c_1\frac{r^2}{Mn^{2\rho}}.
\end{aligned} \tag{130}$$

Fix $\delta_1 \in (0, \delta_0]$ which satisfies that

$$\delta_1 \leq \frac{\lambda}{2c_1} \text{ and } \delta_1^2 \leq \frac{\lambda}{2c_1 r}. \tag{131}$$

Let $\delta \in (0, \delta_1]$. We claim that inequality (130) implies that there exists some $c \in (0,\infty)$ which satisfies for every $n \in \mathbb{N}$ that

$$\mathbb{E}\left[\left(\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)\wedge 1\right)^2\mathbf{1}_{A_{n-1}}\right]^{\frac{1}{2}} \leq cn^{-\frac{\rho}{2}}. \tag{132}$$

The proof of (132) will proceed by induction. Since $\rho \in (2/3, 1)$, there exists $n_0 \geq 1$ such that for every $n \geq n_0$ it holds that

$$\left(n^\rho - (n-1)^\rho - r\lambda + \frac{r^2\lambda^2}{n^\rho}\right) \leq \left(\rho(n-1)^{\rho-1} - r\lambda + \frac{r^2\lambda^2}{n^\rho}\right) \leq -\frac{r\lambda}{2}, \tag{133}$$

where the first inequality follows from the mean value theorem and $\rho \in (2/3, 1)$ and the second inequality is obtained by choosing $n \in \mathbb{N}$ sufficiently large. Fix $n_0 \geq 1$ which satisfies (133) and define $\overline{c} \in (0,\infty)$ which satisfies that

$$\overline{c} = \max\left\{(n_0 - 1)^\rho, \frac{2c_1 r}{M\lambda}\right\}. \tag{134}$$

For the base case, the definition of $\overline{c}$ guarantees for every $n \in \{1, \ldots, n_0 - 1\}$ that

$$\mathbb{E}\left[\left(\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)\wedge 1\right)^2\mathbf{1}_{A_{n-1}}\right] \leq \overline{c}n^{-\rho}. \tag{135}$$

For the induction step, suppose that for $n \geq n_0$ we have that

$$\mathbb{E}\left[\left(\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)\wedge 1\right)^2\mathbf{1}_{A_{n-2}}\right] \leq \overline{c}(n-1)^{-\rho}. \tag{136}$$

Since the event A_{n-1} implies that

$$\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U) \leq \delta \leq 1, \tag{137}$$

it follows from an L^∞-estimate, the inclusion $A_{n-1} \subseteq A_{n-2}$, and the induction hypothesis that for every $m \in \{2,3,4\}$ it holds that

$$\mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)^m \mathbf{1}_{A_{n-1}}\right] \leq \delta^{m-2}\mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)^2 \mathbf{1}_{A_{n-2}}\right] \leq \delta^{m-2}\overline{c}(n-1)^{-\rho}. \tag{138}$$

Returning to (130), it holds that

$$\begin{aligned}\mathbb{E}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)^2 \mathbf{1}_{A_{n-1}}\right] &\leq \overline{c}\left(1-\frac{r\lambda}{n^\rho}\right)^2 (n-1)^{-\rho} + \overline{c}c_1\delta\left(1-\frac{r\lambda}{n^\rho}\right)\frac{r}{n^\rho}(n-1)^{-\rho} \\ &\quad + \overline{c}c_1\delta^2\frac{r^2}{n^{2\rho}}(n-1)^{-\rho} + c_1\frac{r^2}{Mn^{2\rho}}.\end{aligned} \tag{139}$$

After adding and subtracting $\overline{c}n^{-\rho}$, it holds that

$$\begin{aligned}&\mathbb{E}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)^2 \mathbf{1}_{A_{n-1}}\right] \leq \overline{c}n^{-\rho} \\ &+ n^{-\rho}\left(\overline{c}(n-1)^{-\rho}\left(n^\rho - (n-1)^\rho - 2r\lambda + \frac{r^2\lambda^2}{n^\rho} + c_1\delta r\left(1-\frac{r\lambda}{n^\rho}\right) + c_1\delta^2\frac{r^2}{n^\rho}\right) + c_1\frac{r^2}{Mn^\rho}\right).\end{aligned} \tag{140}$$

Since $\delta \in (0, \delta_1]$, it follows from (140) that

$$\mathbb{E}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)^2 \mathbf{1}_{A_{n-1}}\right] \leq \overline{c}n^{-\rho} + n^{-\rho}\left(\overline{c}(n-1)^{-\rho}\left(n^\rho - (n-1)^\rho - r\lambda + \frac{r^2\lambda^2}{n^\rho}\right) + c_1\frac{r^2}{Mn^\rho}\right). \tag{141}$$

Since $n \geq n_0$, the choice $\overline{c} \geq \frac{2c_1 r}{M\lambda}$, (133), and (141) prove that

$$\mathbb{E}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)^2 \mathbf{1}_{A_{n-1}}\right] \leq \overline{c}n^{-\rho} + n^{-\rho}\left(-\frac{r\lambda}{2}\overline{c}(n-1)^{-\rho} + c_1\frac{r^2}{Mn^\rho}\right) \leq \overline{c}n^{-\rho}. \tag{142}$$

Therefore, we have that

$$\mathbb{E}\left[\left(\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)\wedge 1\right)^2 \mathbf{1}_{A_{n-1}}\right] \leq \left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U)^2 \mathbf{1}_{A_{n-1}}\right] \leq \overline{c}n^{-\rho}, \tag{143}$$

which completes the induction step. Since the base case is (135), this completes the proof of Proposition 20. ∎

Proposition 20 proves the convergence of SGD to the local manifold of minima on the event that SGD remains in a basin of attraction. It remains to prove that SGD remains in the basin of attraction for large times with high probability. The first step is contained in the following proposition, which estimates the maximal excursion of SGD on the event that the dynamics do not leave a basin of attraction.

Proposition 21 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, let $\{X_{n,m}\colon \Omega \to \mathbb{R}\}_{n,m\in\mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1}$ satisfy Assumption 2. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1})]$ and assume that f satisfies Assumption 1. For every $M \in \mathbb{N}$, $\rho \in (2/3, 1)$, $r \in (0,\infty)$, $\theta \in \mathbb{R}^d$ let $\{\Theta_{n,\theta} = \Theta_{n,\theta}(M,\rho,r)\}_{n\in\mathbb{N}_0}$ be defined by $\Theta_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M}(\nabla_\theta F)(\Theta_{n-1,\theta}, X_{n,m})\right], \tag{144}$$

and for every $R, \delta \in (0,\infty)$, $x_0 \in (\mathcal{M}\cap U)$, and $n \in \mathbb{N}$ let $A_n = A_n(M,r,\rho,\theta,R,\delta,x_0) \in \mathcal{F}$ be defined by

$$A_n = \left\{\forall\, m \in \{0,\dots,n\}\ \Theta_{m,\theta} \in V_{R,\delta}(x_0) \ \ \textit{(cf. Definition 12)}\right\}. \tag{145}$$

Then for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (2/3, 1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0, \infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$, $\theta \in V_{R/2,\delta}(x_0)$ (cf. Definition 12),

$$\mathbb{E}\left[\max_{1\leq k\leq n} |\Theta_{k,\theta} - \Theta_{0,\theta}|\, \mathbf{1}_{A_{k-1}}\right] \leq \sum_{k=1}^{n} \left(\mathbb{E}\left[|\Theta_{k,\theta} - \Theta_{k-1,\theta}|^2\, \mathbf{1}_{A_{k-1}}\right]\right)^{\frac{1}{2}} \leq cr\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right). \quad (146)$$

Proof [Proof of Proposition 21] Let $\mathbf{d}(\cdot, \mathcal{M} \cap U) : \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $x \in \mathbb{R}^d$ that

$$\mathbf{d}(x, \mathcal{M} \cap U) = \inf\{|x - y| : y \in (\mathcal{M} \cap U)\}. \quad (147)$$

Let $x_0 \in (\mathcal{M} \cap U)$. Since $U \subseteq \mathbb{R}^d$ is open, fix $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. Fix $R_0, \delta_0 \in (0, \infty)$ which satisfies the conclusion of Proposition 13 for this set V. We observe that the regularity of f and the compactness of $\overline{V}_{R_0,\delta_0}(x_0)$ imply that

$$\|f\|_{\mathrm{C}^3(V_{R_0,\delta_0}(x_0))} \leq c. \quad (148)$$

Finally, fix $\mathfrak{r} \in (0, \infty)$ which satisfies the conclusion of Lemma 15. Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$. As in Proposition 20, let $\nabla_\theta F^{M,n} \colon \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be the functions which satisfy for every $(\theta, \omega) \in \mathbb{R}^d \times \Omega$ that

$$\nabla_\theta F^{M,n}(\theta) = \nabla_\theta F^{M,n}(\theta, \omega) = \frac{1}{M} \sum_{m=1}^{M} (\nabla_\theta F)(\theta, X_{n,m}(\omega)). \quad (149)$$

Let $\theta \in V_{R/2,\delta}(x_0)$, let $\Theta_{0,\theta} \colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\Theta_{0,\theta}(\omega) = \theta$, and for every $n \in \mathbb{N}$ let $\Theta_{n,\theta} \colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho} \nabla_\theta F^{M,n}(\Theta_{n-1,\theta}). \quad (150)$$

We will first prove that there exists $c \in (0, \infty)$ which satisfies that

$$\mathbb{E}\left[|\Theta_{n,\theta} - \Theta_{n-1,\theta}|^2\, \mathbf{1}_{A_{n-1}}\right]^{\frac{1}{2}} \leq c\left(\frac{r}{n^{\frac{3}{2}\rho}} + \frac{r}{n^\rho M^{\frac{1}{2}}}\right), \quad (151)$$

where we observe that the constant $c \in (0, \infty)$ can be absorbed by fixing $r \in (0, \mathfrak{r}]$ sufficiently small. It holds that

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho} \nabla f(\Theta_{n-1,\theta}) + \frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right). \quad (152)$$

Lemma 14 proves that there exists $c_1 \in (0, \infty)$ and $\varepsilon_n \colon A_{n-1} \to \mathbb{R}^d$ which satisfy that

$$|\varepsilon_n| \leq c_1 \mathbf{d}(\Theta_{n-1}, \mathcal{M} \cap U)^2, \quad (153)$$

such that on the event A_{n-1} it holds that

$$\nabla f(\Theta_{n-1,\theta}) = \big(\mathrm{Hess}\, f\big)(p(\Theta_{n-1,\theta})) \cdot (\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta})) + \varepsilon_n. \quad (154)$$

Therefore, on the event A_{n-1} it holds that

$$\begin{aligned} \Theta_{n,\theta} = \Theta_{n-1,\theta} &- \frac{r}{n^\rho}\big(\mathrm{Hess}\, f\big)(p(\Theta_{n-1,\theta})) \cdot \big(\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta})\big) - \frac{r}{n^\rho}\varepsilon_n \\ &+ \frac{r}{n^\rho}\big(\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\big). \end{aligned} \quad (155)$$

Let $\tilde{\Theta}^{M,r}_{n-1,\theta}\colon A_{n-1} \to \mathbb{R}^d$ satisfy that

$$\tilde{\Theta}^{M,r}_{n-1,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho}\big(\operatorname{Hess} f\big)(p(\Theta_{n-1,\theta}))\cdot(\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}))\,. \tag{156}$$

After taking the norm-squared of (155), on the event A_{n-1} it holds that

$$\begin{aligned}\left|\Theta_{n,\theta} - \tilde{\Theta}^{M,r}_{n-1,\theta}\right|^2 &= \frac{r^2}{n^{2\rho}}\left|\varepsilon_n\right|^2 - 2\frac{r^2}{n^{2\rho}}\varepsilon_n\cdot\big(\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\big)\\ &\quad + \frac{r^2}{n^{2\rho}}\left|\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right|^2.\end{aligned} \tag{157}$$

We will estimate (157) by taking the expectation on the event A_{n-1}. The first term on the righthand side of (157) is handled using Proposition 20 and (153). For the second term, from (125) we recall the sigma algebras $\mathcal{F}_m \subseteq \mathcal{F}$, $m \in \mathbb{N}$, which satisfy that

$$\mathcal{F}_m = \sigma\big(\{X_{1,k}\}_{k=1}^M, \ldots, \{X_{m,k}\}_{k=1}^M\big). \tag{158}$$

Since $\varepsilon_n\colon A_{n-1} \to \mathbb{R}^d$ is $\mathcal{F}_{n-1}$-measurable, it follows identically to (127) and (128) that

$$\mathbb{E}\left[\varepsilon_n\cdot\big(\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\big)\mathbf{1}_{A_{n-1}}\right] = 0. \tag{159}$$

For the final term on the righthand side of (157), the compactness of $\overline{V}_{R_0,\delta_0}(x_0)$, the independence of the $X_{m,k}$, $m,k \in \mathbb{N}$, and Lemma 19 prove that there exists $c \in (0,\infty)$ which satisfies that

$$\mathbb{E}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right|^2\mathbf{1}_{A_{n-1}}\right] \leq \frac{c}{M}. \tag{160}$$

In combination, Proposition 20 and estimates (153), (157), (159), and (160) prove that there exists $c \in (0,\infty)$ which satisfies that

$$\begin{aligned}\mathbb{E}\left[\left|\Theta_{n,\theta} - \tilde{\Theta}^{M,r}_{n-1,\theta}\right|^2\mathbf{1}_{A_{n-1}}\right] &\leq c\left(\frac{r^2\delta^2}{n^{2\rho}}\mathbb{E}\left[\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)^2\right] + \frac{r^2}{n^{2\rho}M}\right)\\ &\leq c\left(\frac{r^2\delta^2}{n^{3\rho}} + \frac{r^2}{n^{2\rho}M}\right).\end{aligned} \tag{161}$$

It follows from the definition of $\tilde{\Theta}^{M,r}_{n-1,\theta}$, (148), and the definition of the projection that, on the event A_{n-1} there exists $c \in (0,\infty)$ which satisfies that

$$\begin{aligned}\left|\tilde{\Theta}^{M,r}_{n-1,\theta} - \Theta_{n-1,\theta}\right|^2 &= \frac{r^2}{n^{2\rho}}\left|\big(\operatorname{Hess} f\big)(p(\Theta_{n-1,\theta}))\cdot(\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}))\right|^2\\ &\leq c\frac{r^2}{n^{2\rho}}\mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M}\cap U)^2.\end{aligned} \tag{162}$$

Proposition 20 proves that there exists $c \in (0,\infty)$ such that

$$\mathbb{E}\left[\left|\tilde{\Theta}^{M,r}_{n-1,\theta} - \Theta_{n-1,\theta}\right|^2\mathbf{1}_{A_{n-1}}\right] \leq \frac{cr^2}{n^{3\rho}}. \tag{163}$$

It follows from the triangle inequality, (161), and (163) that there exists $c_1 \in (0,\infty)$ which satisfies that

$$\begin{aligned}\mathbb{E}\left[\left|\Theta_{n,\theta} - \Theta_{n-1,\theta}\right|^2\mathbf{1}_{A_{n-1}}\right]^{\frac{1}{2}} &\leq \mathbb{E}\left[\left|\Theta_{n,\theta} - \tilde{\Theta}^{M,r}_{n-1,\theta}\right|^2\mathbf{1}_{A_{n-1}}\right]^{\frac{1}{2}} + \mathbb{E}\left[\left|\tilde{\Theta}^{M,r}_{n-1,\theta} - \Theta_{n-1,\theta}\right|^2\mathbf{1}_{A_{n-1}}\right]^{\frac{1}{2}}\\ &\leq c_1\left(\frac{r}{n^{\frac{3}{2}\rho}} + \frac{r}{n^\rho M^{\frac{1}{2}}}\right),\end{aligned} \tag{164}$$

which completes the proof of (151). Since for every $r \leq s \in \mathbb{N}_0$ we have $\mathbf{1}_{A_s} \leq \mathbf{1}_{A_r}$, it follows from (164), the triangle inequality, and Hölder's inequality that there exists $c_2 \in (0,\infty)$ which satisfies for every $r \in (0, \mathfrak{r}]$ that

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq k\leq n} |\Theta_{k,\theta} - \Theta_{0,\theta}|\, \mathbf{1}_{A_{k-1}}\right] &\leq \sum_{k=1}^{n} \mathbb{E}\left[|\Theta_{k,\theta} - \Theta_{k-1,\theta}|\, \mathbf{1}_{A_{k-1}}\right] \\ &\leq \sum_{k=1}^{n} \mathbb{E}\left[|\Theta_{k,\theta} - \Theta_{k-1,\theta}|^2\, \mathbf{1}_{A_{k-1}}\right]^{\frac{1}{2}} \\ &\leq c_1 r \left(\sum_{k=1}^{n} k^{-\frac{3}{2}\rho} + M^{-\frac{1}{2}} \sum_{k=1}^{n} k^{-\rho}\right) \\ &\leq c_2 r \left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right),\end{aligned} \tag{165}$$

where we have used that fact that, since $\rho \in (2/3, 1)$, there exists a $c \in (0,\infty)$ such that

$$\sum_{k=1}^{n} k^{-\frac{3}{2}\rho} + M^{-\frac{1}{2}} \sum_{k=1}^{n} k^{-\rho} \leq c\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right). \tag{166}$$

This completes the proof of Proposition 21. ■

Remark 22 *The assumption $\rho \in (2/3, 1)$ is only used to ensure the boundedness in $n \in \mathbb{N}$ of the first sum appearing on the lefthand side of* (166)*, which cannot be countered by the mini-batch size $M \in \mathbb{N}$. Every other argument in the paper applies without change to the case $\rho \in (0,1)$. In particular, because the result of Proposition* 21 *is not needed if $\mathcal{M} \cap U$ is compact, the results of Section* 6 *apply for every $\rho \in (0,1)$.*

In the following lemma and proposition, we obtain a lower bound in probability for the events $\{A_n\}_{n\in\mathbb{N}}$, $n \in \mathbb{N}_0$. The interesting observation is that Proposition 20 and Proposition 21 can be used together and inductively to obtain lower bound for these probabilities. Namely, Proposition 20 implies that, on the event A_{n-1}, the process is converging to $\mathcal{M}\cap U$ in the normal directions with high probability, and Proposition 21 can be used to estimate the probability that the solution (109) escapes the basin of attraction along the tangential directions.

Lemma 23 *Let $d \in \mathbb{N}$, let $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, and let $\mathcal{M} \cap U \subseteq \mathbb{R}^d$ be a $\mathfrak{d}$-dimensional C^2-submanifold. Then for every $x_0 \in \mathcal{M}\cap U$ there exists $R_0, \delta_0 \in (0,\infty)$ such that, for every $R \in (0, R_0]$ and $\delta \in (0, \delta_0]$,*

$$\{x \in \mathbb{R}^d \colon \mathbf{d}(x, \mathcal{M}\cap U) < \delta \text{ and } |x - x_0| \leq R - \delta\} \subseteq V_{R,\delta}(x_0). \tag{167}$$

Proof [Proof of Lemma 23] Let $x_0 \in \mathcal{M}\cap U$, let $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8), and let $R_0, \delta_0 \in (0,\infty)$ satisfy the conclusion of Proposition 13. That is, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ it holds that $\overline{V}_{R,\delta}(x_0) \subseteq V$ and that

$$V_{R,\delta}(x_0) = \{x \in \mathbb{R}^d \colon \mathbf{d}(x, \mathcal{M}\cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta\}. \tag{168}$$

Suppose that $x \in \mathbb{R}^d$ satisfies that

$$\mathbf{d}(x, \mathcal{M}\cap U) < \delta \text{ with } |x - x_0| \leq R - \delta. \tag{169}$$

The definition of the distance to $\mathcal{M} \cap U$ and $|x - x_0| \leq R - \delta$ imply that there exists a possibly non-unique $\tilde{x} \in \overline{\mathcal{M} \cap U}$ which satisfies that

$$|x - \tilde{x}| = \mathbf{d}(x, \mathcal{M} \cap U) < \delta. \tag{170}$$

The triangle inequality implies that

$$|\tilde{x} - x_0| \leq |\tilde{x} - x| + |x - x_0| < \delta + (R - \delta) < R. \tag{171}$$

It follows that $\tilde{x} \in \overline{B_R(x_0) \cap \mathcal{M} \cap U}$, and therefore that

$$\mathbf{d}(x, \mathcal{M} \cap U) = \mathbf{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta. \tag{172}$$

It follows from (169) and (172) that $x \in V_{R,\delta}(x_0)$, which completes the proof of Lemma 23. ■

Proposition 24 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F \colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, let $\{X_{n,m} \colon \Omega \to \mathbb{R}\}_{n,m \in \mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1}$ satisfy Assumption 2. Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1})]$ and assume that f satisfies Assumption 1. For every $M \in \mathbb{N}$, $\rho \in (2/3, 1)$, $r \in (0, \infty)$, $\theta \in \mathbb{R}^d$ let $\{\Theta_{n,\theta} = \Theta_{n,\theta}(M, \rho, r)\}_{n \in \mathbb{N}_0}$ be defined by $\Theta_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho M} \left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,\theta}, X_{n,m}) \right], \tag{173}$$

and for every $R, \delta \in (0, \infty)$, $x_0 \in (\mathcal{M} \cap U)$, and $n \in \mathbb{N}$ let $A_n = A_n(M, r, \rho, \theta, R, \delta, x_0) \in \mathcal{F}$ be defined by

$$A_n = \Big\{ \forall\, m \in \{0, \ldots, n\}\ \Theta_{m,\theta} \in V_{R,\delta}(x_0) \ \textit{(cf. Definition 12)} \Big\}. \tag{174}$$

Then for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (2/3, 1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0, \infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$, $\theta \in V_{R/2,\delta}(x_0)$ (cf. Definition 12),

$$\mathbb{P}[A_n] \geq c \left(\exp\left(-\frac{c}{M} \right) - M^{-1} n^{1-\rho} - \frac{r \left(1 + M^{-\frac{1}{2}} n^{1-\rho} \right)}{\left(\frac{R}{2} - 2\delta \right)_+} \right). \tag{175}$$

Proof [Proof of Proposition 24] Let $\mathbf{d}(\cdot, \mathcal{M} \cap U) : \mathbb{R}^d \to \mathbb{R}$ be defined by

$$\mathbf{d}(x, \mathcal{M} \cap U) = \inf \{ |x - y| : y \in (\mathcal{M} \cap U) \}. \tag{176}$$

Let $x_0 \in (\mathcal{M} \cap U)$. Since $U \subseteq \mathbb{R}^d$ is open, fix $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. Fix $R_0, \delta_0 \in (0, \infty)$ which satisfy the conclusion of Proposition 13 for this set V. Fix $\mathfrak{r} \in (0, \infty)$ which satisfies the conclusion of Lemma 15. Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$. As in Proposition 20, let $\nabla_\theta F^{M,n} \colon \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be the functions which satisfy for every $(\theta, \omega) \in \mathbb{R}^d \times \Omega$ that

$$\nabla_\theta F^{M,n}(\theta) = \nabla_\theta F^{M,n}(\theta, \omega) = \frac{1}{M} \sum_{m=1}^{M} (\nabla_\theta F)(\theta, X_{n,m}(\omega)). \tag{177}$$

Let $\theta \in V_{R/2,\delta}(x_0)$, let $\Theta_{0,\theta} \colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\Theta_{0,\theta}(\omega) = \theta$, and for every $n \in \mathbb{N}$ let $\Theta_{n,\theta} \colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho} \nabla_\theta F^{M,n}(\Theta_{n-1,\theta}). \tag{178}$$

Since it holds that

$$\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \geq \delta \text{ implies that } \Theta_{n,\theta} \notin V_{R,\delta}(x_0), \tag{179}$$

it follows that

$$\begin{aligned} \mathbb{P}\left[\Theta_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right] &= \mathbb{P}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \geq \delta, A_{n-1}\right] \\ &\quad + \mathbb{P}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) < \delta, \Theta_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right]. \end{aligned} \tag{180}$$

The two terms on the righthand side of (180) will be handled separately. We will first prove that there exists $c \in (0, \infty)$ which satisfies that

$$\mathbb{P}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \geq \delta, A_{n-1}\right] \leq \frac{c}{Mn^{2\rho}} \mathbb{P}\left[A_{n-1}\right] + \frac{c}{Mn^{\rho}}. \tag{181}$$

On the event A_{n-1}, it follows from Lemma 14 that there exists $\varepsilon_n \colon A_{n-1} \to \mathbb{R}^d$, $c_1 \in (0, \infty)$ such that

$$|\varepsilon_n| \leq c_1 \mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M} \cap U)^2, \tag{182}$$

and such that on the event A_{n-1} it holds that

$$\nabla f(\Theta_{n-1,\theta}) = \big(\operatorname{Hess} f\big)(p(\Theta_{n-1,\theta})) \cdot (\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta})) + \varepsilon_n. \tag{183}$$

Therefore, on the event A_{n-1}, we have that

$$\begin{aligned} \Theta_{n,\theta} &= \Theta_{n-1,\theta} - \frac{r}{n^{\rho}} \big(\operatorname{Hess} f\big)(p(\Theta_{n-1,\theta})) \cdot (\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta})) - \frac{r}{n^{\rho}} \varepsilon_n \\ &\quad + \frac{r}{n^{\rho}} \left(\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right). \end{aligned} \tag{184}$$

Lemma 15, (182), the choice of $\mathfrak{r} \in (0, \infty)$, the definition of the projection, and the triangle inequality prove that there exist $c_1, \lambda \in (0, \infty)$ such that on the event A_{n-1} it holds that

$$\begin{aligned} &\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \\ &\leq \left|\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}) - \frac{r}{n^{\rho}} \big(\operatorname{Hess} f\big)(p(\Theta_{n-1,\theta})) \cdot (\Theta_{n-1,\theta} - p(\Theta_{n-1,\theta}))\right| \\ &\quad + \left|\frac{r}{n^{\rho}} \varepsilon_n\right| + \left|\frac{r}{n^{\rho}} \left(\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right)\right| \\ &\leq \left(1 - \frac{r\lambda}{n^{\rho}}\right) \mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M} \cap U) + c_1 \frac{r}{n^{\rho}} \mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M} \cap U)^2 \\ &\quad + \frac{r}{n^{\rho}} \left|\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right|. \end{aligned} \tag{185}$$

Fix $\delta_1 \in (0, \delta_0]$ which satisfies that

$$c_1 \delta_1 \leq \frac{\lambda}{2}. \tag{186}$$

Let $\delta \in (0, \delta_1]$. On the event A_{n-1}, it follows from (185) and the choice of $\delta_1 \in (0, \delta_0]$ that

$$\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \leq \left(1 - \frac{r\lambda}{2n^{\rho}}\right) \mathbf{d}(\Theta_{n-1,\theta}, \mathcal{M} \cap U) + \frac{r}{n^{\rho}} \left|\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right|. \tag{187}$$

We therefore conclude that

$$\begin{aligned} &\mathbb{P}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \geq \delta, A_{n-1}\right] \leq \\ &\mathbb{P}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right| \geq \frac{\delta n^{\rho}}{2r}, \Theta_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] \\ &+ \mathbb{P}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla F^{M,n}(\Theta_{n-1,\theta})\right| \geq \frac{\delta\lambda}{2}, \Theta_{n-1,\theta} \in V_{R,\delta}(x_0) \setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right]. \end{aligned} \tag{188}$$

Similarly to (127) and computation (128), it follows from the independence of the random variables $X_{m,k}$, $m,k \in \mathbb{N}$, that

$$\begin{aligned}&\mathbb{P}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right| \geq \frac{\delta n^\rho}{2r}, \Theta_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] \\ &\leq \sup_{\theta \in V_{R,\frac{\delta}{2}}(x_0)} \mathbb{P}\left[\left|\nabla f(\theta) - \nabla_\theta F^{M,n}(\theta)\right| \geq \frac{\delta n^\rho}{2r} \right] \mathbb{P}\left[\Theta_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right],\end{aligned} \tag{189}$$

and that

$$\begin{aligned}&\mathbb{P}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right| \geq \frac{\delta\lambda}{2}, \Theta_{n-1,\theta} \in V_{R,\delta}(x_0) \setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] \\ &\leq \sup_{\theta \in V_{R,\delta}(x_0)\setminus V_{R,\frac{\delta}{2}}(x_0)} \mathbb{P}\left[\left|\nabla f(\theta) - \nabla_\theta F^{M,n}(\theta)\right| \geq \frac{\delta\lambda}{2} \right] \mathbb{P}\left[\Theta_{n-1,\theta} \in V_{R,\delta}(x_0) \setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right].\end{aligned} \tag{190}$$

The definition of A_{n-1}, Chebyshev's inequality, Lemma 19, and (189) prove that there exists $c \in (0,\infty)$ which satisfies that

$$\begin{aligned}\mathbb{P}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right| \geq \frac{\delta n^\rho}{2r}, \Theta_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] &\leq \frac{c}{M} \cdot \frac{4r^2}{\delta^2 n^{2\rho}} \mathbb{P}\left[A_{n-1}\right] \\ &\leq \frac{c}{Mn^{2\rho}} \mathbb{P}\left[A_{n-1}\right].\end{aligned} \tag{191}$$

In the case of (190), Proposition 20 and Chebyshev's inequality prove that, for the indicator function $\mathbf{1}_{A_{n-2}}$ of the event A_{n-2}, there exists $c \in (0,\infty)$ which satisfies that

$$\begin{aligned}\mathbb{P}\left[\Theta_{n-1,\theta} \in V_{R,\delta}(x_0) \setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] &\leq \mathbb{P}\left[\left(\mathbf{d}\left(\Theta_{n-1,\theta}, \mathcal{M} \cap U\right) \wedge 1\right)^2 \mathbf{1}_{A_{n-2}} \geq \frac{\delta^2}{4}\right] \\ &\leq \frac{4c}{\delta^2} n^{-\rho} \\ &\leq c n^{-\rho},\end{aligned} \tag{192}$$

where we have used the fact that, since $\rho \in (2/3, 1)$, there exists $c \in (0,\infty)$ such that for every $n \in \mathbb{N}$ it holds that $(n-1)^{-\rho} \leq c n^{-\rho}$. Furthermore, Chebyshev's inequality and Lemma 19 prove that there exists $c \in (0,\infty)$ which satisfies that

$$\mathbb{P}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right| \geq \frac{\delta\lambda}{2} \right] \leq \frac{c}{M} \cdot \frac{4}{\delta^2\lambda^2} \leq \frac{c}{M}. \tag{193}$$

Returning to (190), the previous two inequalities prove that there exists $c \in (0,\infty)$ which satisfies that

$$\mathbb{P}\left[\left|\nabla f(\Theta_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta_{n-1,\theta})\right| \geq \frac{\delta\lambda}{2}, \Theta_{n-1,\theta} \in V_\delta \setminus V_{\frac{\delta}{2}}, A_{n-2}\right] \leq \frac{c}{Mn^\rho}. \tag{194}$$

Combining (188), (191), and (194), there exists $c \in (0,\infty)$ such that

$$\mathbb{P}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) \geq \delta, A_{n-1}\right] \leq \frac{c}{Mn^{2\rho}} \mathbb{P}\left[A_{n-1}\right] + \frac{c}{Mn^\rho}, \tag{195}$$

which completes the proof of (181). Returning to (180), it follows from (195) that there exists $c \in (0,\infty)$ such that

$$\begin{aligned}&\mathbb{P}\left[\Theta_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right] \\ &\leq \frac{c}{Mn^{2\rho}} \mathbb{P}\left[A_{n-1}\right] + \frac{c}{Mn^\rho} + \mathbb{P}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M} \cap U) < \delta, \Theta_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right].\end{aligned} \tag{196}$$

Therefore, there exists $c \in (0,\infty)$ which satisfies that

$$\begin{aligned}\mathbb{P}[A_n] &= \mathbb{P}\left[\Theta_{n,\theta} \in V_{R,\delta}(x_0), A_{n-1}\right] \\ &\geq \left(1 - \frac{c}{Mn^{2\rho}}\right)_+ \mathbb{P}\left[A_{n-1}\right] - \frac{c}{Mn^{\rho}} - \mathbb{P}\left[\mathbf{d}(\Theta_{n,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right].\end{aligned} \tag{197}$$

We will prove inductively that (197) implies that there exists $c \in (0,\infty)$ such that for every $n \in \mathbb{N}$ it holds that

$$\mathbb{P}[A_n] \geq \prod_{k=1}^{n}\left(1 - \frac{c}{Mk^{2\rho}}\right)_+ - \sum_{k=1}^{n}\frac{c}{Mk^{\rho}} - \sum_{k=1}^{n}\mathbb{P}\left[\mathbf{d}(\Theta_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right]. \tag{198}$$

The base case $n = 0$ follows immediately from $\theta \in V_{R/2,\delta}(x_0)$. For the inductive step, suppose that (202) is satisfied for some $n \in \mathbb{N}$. It follows from (197) that

$$\begin{aligned}\mathbb{P}\left[A_{n+1}\right] \geq \left(1 - \frac{c}{M(n+1)^{2\rho}}\right)_+ \mathbb{P}\left[A_n\right] - \frac{c}{M(n+1)^{\rho}} \\ - \mathbb{P}\left[\mathbf{d}(\Theta_{n+1,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{n+1,\theta} \notin V_{R,\delta}(x_0), A_n\right].\end{aligned} \tag{199}$$

It then follows from the inductive hypothesis (202) that

$$\begin{aligned}&\mathbb{P}\left[A_{n+1}\right] \\ &\geq \prod_{k=1}^{n+1}\left(1 - \frac{c}{Mk^{2\rho}}\right)_+ \mathbb{P}\left[A_0\right] - \frac{c}{M(n+1)^{\rho}} \\ &- \mathbb{P}\left[\mathbf{d}(\Theta_{n+1,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{n+1,\theta} \notin V_{R,\delta}(x_0), A_n\right] \\ &- \left(1 - \frac{c}{M(n+1)^{2\rho}}\right)_+ \left(\sum_{k=1}^{n}\frac{c}{Mk^{\rho}} + \sum_{k=1}^{n}\mathbb{P}\left[\mathbf{d}(\Theta_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right]\right),\end{aligned} \tag{200}$$

which proves that

$$\begin{aligned}&\mathbb{P}\left[A_{n+1}\right] \\ &\geq \prod_{k=1}^{n+1}\left(1 - \frac{c}{Mk^{2\rho}}\right)_+ \mathbb{P}\left[A_0\right] - \sum_{k=1}^{n+1}\frac{c}{Mk^{\rho}} - \sum_{k=1}^{n+1}\mathbb{P}\left[\mathbf{d}(\Theta_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right].\end{aligned} \tag{201}$$

Finally, since $\theta \in V_{R/2,\delta}(x_0) \subseteq V_{R,\delta}(x_0)$ implies that $\mathbb{P}(A_0) = 1$, it holds that

$$\mathbb{P}\left[A_{n+1}\right] \geq \prod_{k=1}^{n+1}\left(1 - \frac{c}{Mk^{2\rho}}\right)_+ - \sum_{k=1}^{n+1}\frac{c}{Mk^{\rho}} - \sum_{k=1}^{n+1}\mathbb{P}\left[\mathbf{d}(\Theta_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right], \tag{202}$$

which completes the induction step, and the proof of (202). It remains only to estimate the final term on the righthand side of inequality (202). The definition of the events A_m, $m \in \mathbb{N}_0$, implies that

$$\{\mathbf{d}(\Theta_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\} \subseteq \Omega, \ k \in \mathbb{N}, \text{ are disjoint events.} \tag{203}$$

Therefore, it holds that

$$\begin{aligned}&\sum_{k=1}^{n}\mathbb{P}\left[\mathbf{d}(\Theta_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right] \\ &= \mathbb{P}\left[\coprod_{k=1}^{n}\{\mathbf{d}(\Theta_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\}\right].\end{aligned} \tag{204}$$

Lemma 23 proves that

$$\mathbb{P}\left[\prod_{k=1}^{n}\{\mathbf{d}(\Theta_{k,\theta},\mathcal{M}\cap U)<\delta,\Theta_{k,\theta}\notin V_{R,\delta}(x_0),A_{k-1}\}\right]\leq\mathbb{P}\left[\max_{1\leq k\leq n}|\Theta_{k,\theta}-x_0|\,\mathbf{1}_{A_{k-1}}>R-\delta\right]. \tag{205}$$

Since $\Theta_{0,\theta}^{M,k}\in V_{R/2,\delta}(x_0)$, the triangle inequality prove for every $k\in\{1,2,\ldots,n\}$ that

$$\begin{aligned}|\Theta_{k,\theta}-x_0|&\leq\left|\Theta_{k,\theta}-\Theta_{0,\theta}^{M,k}\right|+\left|\Theta_{0,\theta}^{M,k}-p(\Theta_{0,\theta}^{M,k})\right|+|p(\Theta_{0,\theta})-x_0|\\&\leq|\Theta_{k,\theta}-\theta|+\delta+\frac{R}{2}.\end{aligned} \tag{206}$$

Therefore, for every $k\in\{1,\ldots,n\}$, on the event $\left\{\left|\Theta_{k,\theta}-x_0\right|>R-\delta\right\}$ it holds that

$$\frac{R}{2}-2\delta<\left|\Theta_{k,\theta}-\Theta_{0,\theta}\right|. \tag{207}$$

This implies that

$$\left\{\max_{1\leq k\leq n}|\Theta_{k,\theta}-x_0|\,\mathbf{1}_{A_{k-1}}>R-\delta\right\}\subseteq\left\{\max_{1\leq k\leq n}|\Theta_{k,\theta}-\Theta_{0,\theta}|\,\mathbf{1}_{A_{k-1}}>\frac{R}{2}-2\delta\right\}. \tag{208}$$

In combination, (204), (205), and (208) prove that

$$\sum_{k=1}^{n}\mathbb{P}\left[\mathbf{d}(\Theta_{k,\theta},\mathcal{M}\cap U)<\delta,\Theta_{k,\theta}\notin V_{R,\delta}(x_0),A_{k-1}\right]\leq\mathbb{P}\left[\max_{1\leq k\leq n}|\Theta_{k,\theta}-\Theta_{0,\theta}|\,\mathbf{1}_{A_{k-1}}>\frac{R}{2}-2\delta\right]. \tag{209}$$

It follows from Proposition 21, (209), and Chebyshev's inequality that there exists $c\in(0,\infty)$ which satisfies that

$$\sum_{k=1}^{n}\mathbb{P}\left[\mathbf{d}(\Theta_{k,\theta},\mathcal{M}\cap U)<\delta,\Theta_{k,\theta}\notin V_{R,\delta}(x_0),A_{k-1}\right]\leq\frac{cr\left(1+M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2}-2\delta\right)_+}. \tag{210}$$

Returning to (202), it follows from (210) that there exists $c\in(0,\infty)$ which satisfies that

$$\mathbb{P}[A_n]\geq\prod_{k=1}^{n}\left(1-\frac{c}{Mk^{2\rho}}\right)_+-cM^{-1}n^{1-\rho}-\frac{cr\left(1+M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2}-2\delta\right)_+}, \tag{211}$$

where we have used the fact that, since $\rho\in(2/3,1)$, there exists $c\in(0,\infty)$ which satisfies that

$$\sum_{k=1}^{n}k^{-\rho}\leq cn^{1-\rho}. \tag{212}$$

Finally, it follows that, for $c\in(0,\infty)$ as in (211), for all $M\geq 2c$,

$$\log\left(\prod_{k=1}^{n}\left(1-\frac{c}{Mk^{2\rho}}\right)_+\right)\geq-\frac{2c}{M}\sum_{k=1}^{n}k^{-2\rho}. \tag{213}$$

Therefore, since $\rho\in(2/3,1)$ that the sum on the righthand side of (211) is uniformly bounded in $n\in\mathbb{N}$, the there exists $c\in(0,\infty)$ such that, for all $M\in\mathbb{N}$ sufficiently large,

$$\prod_{k=1}^{n}\left(1-\frac{c}{Mk^{2\rho}}\right)_+\geq\exp\left(-\frac{c}{M}\right). \tag{214}$$

Since the case of small $M \in \mathbb{N}$ can be absorbed into the second term on the rigthhand side of (211), we conclude from (211) and (213) that there exists $c \in (0,\infty)$ such that

$$\mathbb{P}[A_n] \geq c \left(\exp\left(-\frac{c}{M}\right) - M^{-1} n^{1-\rho} - \frac{r\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+} \right). \tag{215}$$

This completes the proof of Proposition 24. ∎

In the following theorem, we estimate the convergence of SGD with initial data sampled uniformly from a non-empty, bounded open set of $\mathbb{R}^d$. We assume that this sampling is done independently of the noise driving the algorithm.

Theorem 25 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, and let $\{X_{n,m}\colon \Omega \to \mathbb{R}\}_{n,m\in\mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1}$ satisfy Assumption 2. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1})]$ and assume that f satisfies Assumption 1. For every $M \in \mathbb{N}$, $\rho \in (2/3, 1)$, $r \in (0,\infty)$, and $\theta \in \mathbb{R}^d$ let $\{\Theta_{n,\theta} = \Theta_{n,\theta}(M,\rho,r)\}_{n\in\mathbb{N}_0}$ be defined by $\Theta_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho M} \left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,\theta}, X_{n,m}) \right]. \tag{216}$$

Let $A \subseteq \mathbb{R}^d$ be a non-empty, bounded open set and let $\Theta_0\colon \Omega \to A$ be a uniformly distributed random variable on A that is independent of $\{X_{n,m}\}_{n,m\in\mathbb{N}}$. Then for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (2/3, 1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0,\infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$, $\varepsilon \in (0,1)$,

$$\mathbb{P}\Big(\mathbf{d}\big(\Theta_{n,\Theta_0}, \mathcal{M} \cap U\big) \geq \varepsilon\Big) \leq \frac{\big|A \backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c \left(\varepsilon^{-2} n^{-\rho} + M^{-1} n^{1-\rho} + \frac{r\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+} \right). \tag{217}$$

Proof [Proof of Theorem 25] Let $x_0 \in (\mathcal{M} \cap U)$. Since $U \subseteq \mathbb{R}^d$ is open, fix $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. Fix $R_0, \delta_0 \in (0,\infty)$ that satisfy the conclusion of Proposition 13 for this set V. Fix $\mathfrak{r} \in (0,\infty)$ that satisfies the conclusions of Lemma 15 and Proposition 24. Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $M \in \mathbb{N}$. As in Proposition 20, let $\nabla_\theta F^{M,n}\colon \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be the functions which satisfy for every $(\theta, \omega) \in \mathbb{R}^d \times \Omega$ that

$$\nabla_\theta F^{M,n}(\theta) = \nabla_\theta F^{M,n}(\theta, \omega) = \frac{1}{M} \sum_{m=1}^{M} (\nabla_\theta F)(\theta, X_{n,m}(\omega)). \tag{218}$$

For every $\theta \in \mathbb{R}^d$ let $\Theta_{0,\theta}\colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\Theta_{0,\theta}(\omega) = \theta$ and for every $n \in \mathbb{N}$ let $\Theta_{n,\theta}\colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta_{n,\theta} = \Theta_{n-1,\theta} - \frac{r}{n^\rho} \nabla_\theta F^{M,n}(\Theta_{n-1,\theta}). \tag{219}$$

Let $\Theta_0\colon \Omega \to \mathbb{R}^d$ be a random variable which is uniformly distributed on A, assume that Θ_0 and $(X_{n,m})_{n,m\in\mathbb{N}}$ are independent, and for every $n \in \mathbb{N}$ let $\Theta_n\colon \Omega \to \mathbb{R}^d$ be defined by $\Theta_n = \Theta_{n,\Theta_0}$. Let $n \in \mathbb{N}$, $\varepsilon \in (0,1)$. We have that

$$\begin{aligned} \mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M} \cap U\big) \geq \varepsilon\Big) &= \mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M} \cap U\big) \geq \varepsilon, \Theta_0 \in V_{R/2,\delta}(x_0)\Big) \\ &\quad + \mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M} \cap U\big) \geq \varepsilon, \Theta_0 \notin V_{R/2,\delta}(x_0)\Big). \end{aligned} \tag{220}$$

For the second term on the righthand side of (217), it follows from the uniform distribution of Θ_0 on A that

$$\mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \notin V_{R/2,\delta}(x_0)\Big) \leq \frac{\big|A\backslash V_{R/2,\delta}(x_0)\big|}{|\lambda(A)|}. \tag{221}$$

We will now estimate the first term on the righthand side of (220). For every $m \in \mathbb{N}_0$, $\theta \in \mathbb{R}^d$ let $A_{m,\theta} \subseteq \Omega$ be the event which satisfies that that

$$A_{m,\theta} = \Big\{\forall\, k \in \{0,\ldots,m\}\ \Theta_{k,\theta} \in V_{R,\delta}(x_0)\ \Big\}, \tag{222}$$

and for every $m \in \mathbb{N}_0$ let $A_m \in \mathcal{F}$ satisfy that

$$A_m = \Big\{\forall\, k \in \{0,\ldots,m\}\ \Theta_k \in V_{R,\delta}(x_0)\Big\}. \tag{223}$$

It holds that

$$\begin{aligned}
&\mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \in V_{R/2,\delta}(x_0)\Big)\\
&= \mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \in V_{R/2,\delta}(x_0), A_{n-1}\Big)\\
&\quad + \mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \in V_{R/2,\delta}(x_0), \Omega\backslash A_{n-1}\Big).
\end{aligned} \tag{224}$$

For the second term on the righthand side of (224), it follows from Proposition 24 that there exists $c \in (0,\infty)$ such that

$$\begin{aligned}
&\mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \notin V_{R/2,\delta}(x_0), \Omega\backslash A_{n-1}\Big)\\
&\leq 1 - c\exp\Big(-\frac{c}{M}\Big) + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+},
\end{aligned} \tag{225}$$

where we have used the fact that $\rho \in (2/3, 1)$ implies that there exists $c \in (0,\infty)$ that satisfies for every $n \in \{2,3,\ldots\}$ that $n^{1-\rho} \leq c(n-1)^{1-\rho}$. For the first term on the righthand side of (224), since the random variables Θ_0 and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, it holds that

$$\begin{aligned}
&\mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \in V_{R/2,\delta}(x_0), A_{n-1}\Big)\\
&\leq \frac{\big|V_{R/2,\delta}(x_0)\cap A\big|}{|A|} \sup_{\theta\in V_{R/2,\delta}(x_0)} \mathbb{P}\Big(\mathbf{d}\big(\Theta_{n,\theta}, \mathcal{M}\cap U\big) \geq \varepsilon, A_{n-1,\theta}\Big).
\end{aligned} \tag{226}$$

Proposition 20 and Chebyshev's inequality prove that there exists $c \in (0,\infty)$ such that for every $\theta \in V_{R/2,\delta}(x_0)$ it holds that

$$\mathbb{P}\Big(\mathbf{d}\big(\Theta_{n,\theta}, \mathcal{M}\cap U\big) \geq \varepsilon, A_{n-1,\theta}\Big) \leq \varepsilon^{-2}\mathbb{E}\Big[\Big(\mathbf{d}\big(\Theta_{n,\theta}, \mathcal{M}\cap U\big)\wedge 1\Big)^2 \mathbf{1}_{A_{n-1,\theta}}\Big] \leq c\varepsilon^{-2}n^{-\rho}. \tag{227}$$

In combination (226) and (227) prove that there exists $c \in (0,\infty)$ such that

$$\mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \in V_{R/2,\delta}(x_0), A_{n-1}\Big) \leq c\varepsilon^{-2}n^{-\rho}. \tag{228}$$

Returning to (224), it follows from (225) and (228) that there exists $c \in (0,\infty)$ such that

$$\begin{aligned}
&\mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon, \Theta_0 \in V_{R/2,\delta}(x_0)\Big)\\
&\leq c\varepsilon^{-2}n^{-\rho} + 1 - c\exp\Big(-\frac{c}{M}\Big) + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}.
\end{aligned} \tag{229}$$

Returning finally to (220), it follows from (221) and (229) that there exists $c \in (0,\infty)$ such that

$$\mathbb{P}\Big(\mathbf{d}\big(\Theta_n, \mathcal{M}\cap U\big) \geq \varepsilon\Big) \leq \\ \frac{\big|A\backslash V_{R/2,\delta}(x_0)\big|}{|\lambda(A)|} + c\varepsilon^{-2}n^{-\rho} + 1 - c\exp\Big(-\frac{c}{M}\Big) + cM^{-1}n^{1-\rho} + \frac{cr\left(1+M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2}-2\delta\right)_+}. \tag{230}$$

Finally, we observe that by Taylor expansion there exists $c \in (0,\infty)$ such that

$$1 - c\exp\Big(-\frac{c}{M}\Big) \leq \frac{c}{M}, \tag{231}$$

which since $1-\rho > 0$ can be absorbed into the fifth term on the righthand side of (230). This completes the proof of Theorem 25. ∎

Remark 26 *In the conclusion of Theorem* 25*, we observe that $\delta \in (0,\infty)$ can be chosen small enough in relation to $R \in (0,\infty)$ to guarantee that $R/2 - \delta > 0$. In this way, since $M^{-1} < M^{-\frac{1}{2}}$, we conclude that for this choice of $R \in (0,\infty)$ and $\delta \in (0,1)$ there exists $c \in (0,\infty)$ such that*

$$\mathbb{P}\Big(\mathbf{d}\big(\Theta_{n,\Theta_0}, \mathcal{M}\cap U\big) \geq \varepsilon\Big) \leq \frac{\big|A\backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c\left(\varepsilon^{-2}n^{-\rho} + M^{-1}n^{1-\rho} + r\right),$$

which recovers the statement of Theorem 3 *of the introduction. The form of estimate* (217) *is nonetheless useful, however, because it quantifies the fact that $r \in (0,\infty)$ can be chosen on the order of the tangential diameter of the basin of attraction $V_{\frac{R}{2},\delta}(x_0)$.*

The next corollary extends Theorem 25 to the case of several independent copies of SGD, each with the same mini-batch size. Corollary 28 below then extends this result to a statement about the convergence of the objective function.

Corollary 27 *Let $d \in \mathbb{N}$, let $(S,\mathcal{S})$ be a measurable space, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, and let $\{X_{n,m,k}\colon \Omega \to \mathbb{R}\}_{n,m,k\in\mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1,1}$ satisfy Assumption* 2*. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1,1})]$ and assume that f satisfies Assumption* 1*. For every $k, M \in \mathbb{N}$, $\rho \in (2/3, 1)$, $r \in (0,\infty)$, and $\theta \in \mathbb{R}^d$ let $\{\Theta^k_{n,\theta} = \Theta^k_{n,\theta}(M,\rho,r)\}_{n\in\mathbb{N}_0}$ be defined by $\Theta^k_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta^k_{n,\theta} = \Theta^k_{n-1,\theta} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta^k_{n-1,\theta}, X_{n,m,k})\right]. \tag{232}$$

Let $A \subseteq \mathbb{R}^d$ be a non-empty, bounded open set and let $\{\Theta^k_0\colon \Omega \to A\}_{k\in\mathbb{N}}$ be i.i.d. uniformly distributed random variables on A that are independent of $\{X_{n,m,k}\}_{n,m,k\in\mathbb{N}}$. Then for every $x_0 \in (\mathcal{M}\cap U)$ and $\rho \in (2/3, 1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0,\infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M, K \in \mathbb{N}$, $\varepsilon \in (0,1)$ it holds that

$$\mathbb{P}\Big(\min_{k\in\{1,2,\ldots,K\}} \mathbf{d}(\Theta^k_{n,\Theta^k_0}, \mathcal{M}\cap U) \geq \varepsilon\Big) \leq \\ \left(\frac{\big|A\backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c\left(\varepsilon^{-2}n^{-\rho} + M^{-1}n^{1-\rho} + \frac{r\left(1+M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2}-2\delta\right)_+}\right)\right)^K. \tag{233}$$

Proof [Proof of Corollary 27] Let $x_0 \in (\mathcal{M} \cap U)$. Since $U \subseteq \mathbb{R}^d$ is open, fix $V \in \text{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. Fix $R_0, \delta_0 \in (0, \infty)$ which satisfy the conclusion of Proposition 13 for this set V. Fix $\mathfrak{r} \in (0, \infty)$ which satisfy the conclusions of Lemma 15 and Proposition 24. Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M, K \in \mathbb{N}$. Since the $\{\Theta^k_{n,\Theta^k_0}\}_{k\in\mathbb{N}}$ are i.i.d. we have that

$$\begin{aligned} \mathbb{P}\Big(\min_{k\in\{1,2,\ldots,K\}} \mathbf{d}(\Theta^k_{n,\Theta^k_0}, \mathcal{M} \cap U) \geq \varepsilon \Big) &= \prod_{k=1}^{K} \mathbb{P}\Big(\mathbf{d}(\Theta^k_{n,\Theta^k_0}, \mathcal{M} \cap U) \geq \varepsilon \Big) \\ &= \mathbb{P}\Big(\mathbf{d}(\Theta^1_{n,\Theta^1_0}, \mathcal{M} \cap U) \geq \varepsilon \Big)^K . \end{aligned} \tag{234}$$

Theorem 25 and (234) prove estimate (236), which completes the proof of Corollary 27. ■

Corollary 28 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, and let $\{X_{n,m,k}\colon \Omega \to \mathbb{R}\}_{n,m,k\in\mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1,1}$ satisfy Assumption 2. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1,1})]$ and assume that f satisfies Assumption 1. For every $k, M \in \mathbb{N}$, $\rho \in (2/3, 1)$, $r \in (0, \infty)$, and $\theta \in \mathbb{R}^d$ let $\{\Theta^k_{n,\theta} = \Theta^k_{n,\theta}(M, \rho, r)\}_{n\in\mathbb{N}_0}$ be defined by $\Theta^k_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta^k_{n,\theta} = \Theta^k_{n-1,\theta} - \frac{r}{n^\rho M} \left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta^k_{n-1,\theta}, X_{n,m,k}) \right]. \tag{235}$$

Let $A \subseteq \mathbb{R}^d$ be a non-empty, bounded open set and let $\{\Theta^k_0\colon \Omega \to A\}_{k\in\mathbb{N}}$ be i.i.d. uniformly distributed random variables on A that are independent of $\{X_{n,m,k}\}_{n,m,k\in\mathbb{N}}$. Then for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (2/3, 1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0, \infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M, K \in \mathbb{N}$, $\varepsilon \in (0, 1)$,

$$\begin{aligned} &\mathbb{P}\Big(\min_{k\in\{1,2,\ldots,K\}} f(\Theta^k_{n,\Theta^k_0}) - \inf_{\theta\in\mathbb{R}^d} f(\theta) \geq \varepsilon \Big) \leq \\ &\left(\frac{\big|A \backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c \left(\varepsilon^{-2} n^{-\rho} + M^{-1} n^{1-\rho} + \frac{r\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+} \right) \right)^K . \end{aligned} \tag{236}$$

Proof [Proof of Corollary 28] The proof is an immediate consequence of Corollary 27 and the local regularity of the objective function. ■

Since the minimum appearing on the lefthand side of (236) is oftentimes computationally impractical or impossible to determine, in the lemma and theorem to follow we compute this minimum using a second mini-batch approximation. The lemma below identifies a measurable selection for this minimum, and the theorem below estimates the probability of identifying the correct minimum using the mini-batch approximation.

Lemma 29 *Let $d \in \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $\{X_k\colon \Omega \to S\}_{k\in\mathbb{N}}$ be i.i.d. random variables, and let $\{\Theta^k\colon \Omega \to \mathbb{R}^d\}_{k\in\mathbb{N}}$ be i.i.d. random variables. Then for every $K, \mathfrak{M} \in \mathbb{N}$ there exists a random variable $\Theta^{K,\mathfrak{M}}\colon \Omega \to \mathbb{R}^d$ such that*

$$\frac{1}{\mathfrak{M}} \sum_{m=1}^{\mathfrak{M}} F(\Theta^{K,\mathfrak{M}}, X_m) = \Big[\min_{k\in\{1,2,\ldots,K\}} \Big(\frac{1}{\mathfrak{M}} \sum_{m=1}^{\mathfrak{M}} F(\Theta^k, X_m) \Big) \Big]. \tag{237}$$

Proof [Proof of Lemma 29] Let $K, \mathfrak{M} \in \mathbb{N}$. Let $\mathfrak{K}\colon \Omega \to \{1, 2, \ldots, \mathfrak{M}\}$ satisfy for every $\omega \in \Omega$ that

$$\mathfrak{K}(\omega) = \min\Big\{k \in \{1, 2, \ldots, K\}\colon \sum_{m=1}^{\mathfrak{M}} F(\Theta^k(\omega), X_m) = \Big[\min_{j \in \{1,2,\ldots,\mathfrak{M}\}} \Big(\sum_{m=1}^{\mathfrak{M}} F(\Theta^j, X_m)\Big)\Big]\Big\}. \tag{238}$$

Let $\mathit{\Theta}^{K,\mathfrak{M}}\colon \Omega \to \mathbb{R}^d$ be defined by

$$\mathit{\Theta}^{K,\mathfrak{M}}(\omega) = \Theta^{\mathfrak{K}(\omega)}(\omega). \tag{239}$$

It follow from (238) and (239) that $\Theta^{K,\mathfrak{M}}$ is measurable and satisfies (237), which completes the proof of Lemma 29. ∎

Theorem 30 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, and let $\{X_{n,m,k}\colon \Omega \to \mathbb{R}\}_{n,m,k \in \mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1,1}$ satisfy Assumption 2. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1,1})]$ and assume that f satisfies Assumption 1. For every $k, M \in \mathbb{N}$, $\rho \in (2/3, 1)$, $r \in (0, \infty)$, and $\theta \in \mathbb{R}^d$ let $\{\Theta^k_{n,\theta} = \Theta^k_{n,\theta}(M, \rho, r)\}_{n \in \mathbb{N}_0}$ be defined by $\Theta^k_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta^k_{n,\theta} = \Theta^k_{n-1,\theta} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta^k_{n-1,\theta}, X_{n,m,k})\right]. \tag{240}$$

Let $A \subseteq \mathbb{R}^d$ be a non-empty, bounded open set and let $\{\Theta^k_0\colon \Omega \to A\}_{k \in \mathbb{N}}$ be i.i.d. uniformly distributed random variables on A that are independent of $\{X_{n,m,k}\}_{n,m,k \in \mathbb{N}}$. For every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $\rho \in (2/3, 1)$, and $r \in (0, \infty)$ let $\{\mathit{\Theta}_n = \mathit{\Theta}(M, \mathfrak{M}, K, \rho, r)\colon \Omega \to \mathbb{R}^d\}_{n \in \mathbb{N}_0}$ be random variables which satisfy that

$$\frac{1}{\mathfrak{M}} \sum_{m=1}^{\mathfrak{M}} F(\mathit{\Theta}_n, X_{n+1,1,m}) = \Big[\min_{k \in \{1,2,\ldots,K\}} \Big(\frac{1}{\mathfrak{M}} \sum_{m=1}^{\mathfrak{M}} F(\Theta^k_{n,\Theta^k_0}, X_{n+1,1,m})\Big)\Big]. \tag{241}$$

Then for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (2/3, 1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0, \infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M, \mathfrak{M}, K \in \mathbb{N}$, $\varepsilon \in (0, 1)$,

$$\begin{aligned} &\mathbb{P}\Big(\Big[f(\mathit{\Theta}_n) - \inf\nolimits_{\theta \in \mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{cK}{\varepsilon^2 \mathfrak{M}} \\ &+ \left(\frac{\big|A \backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c\left(\varepsilon^{-2} n^{-\rho} + M^{-1} n^{1-\rho} + \frac{r\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}\right)\right)^K. \end{aligned} \tag{242}$$

Proof [Proof of Theorem 30] Let $x_0 \in (\mathcal{M} \cap U)$. Since $U \subseteq \mathbb{R}^d$ is open, fix $V \in \mathrm{Proj}(x_0)$ (cf. Definition 8) which satisfies that $V \subseteq U$. Fix $R_0, \delta_0 \in (0, \infty)$ which satisfy the conclusion of Proposition 13 for this set V. Fix $\mathfrak{r} \in (0, \infty)$ which satisfy the conclusions of Lemma 15 and Proposition 24. Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M, \mathfrak{M}, K \in \mathbb{N}$. For every $i \in \{1, 2, \ldots, K\}$ let $B'_i \subseteq \Omega$ satisfy that

$$B'_i = \Big\{\omega \in \Omega\colon f(\Theta^i_{n,\Theta^i_0}(\omega)) = \Big[\min_{k \in \{1,2,\ldots,K\}} f(\Theta^k_{n,\Theta^k_0}(\omega))\Big]\Big\}, \tag{243}$$

and let $B_1 \subseteq \Omega$ satisfy that $B_1 = B_1'$ and for every $i \in \{2, 3, \ldots, K\}$ let $B_i \subseteq \Omega$ satisfy that $B_i = B_i' \backslash \cup_{m=1}^{i-1} B_m$. Since the events B_i, $i \in \{1, 2, \ldots, K\}$, are disjoint, it holds that

$$\begin{aligned}
&\mathbb{P}\Big(\Big[f(\mathit{\Theta}_n) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \\
&= \sum_{i=1}^{K} \mathbb{P}\Big(\Big[f(\mathit{\Theta}_n) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon, B_i\Big) \\
&= \sum_{i=1}^{K} \mathbb{P}\Big(\big[f(\mathit{\Theta}_n) - f(\Theta^i_{n,\Theta_0^i}) + f(\Theta^i_{n,\Theta_0^i}) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\big] \geq \varepsilon, B_i\Big) \\
&\leq \mathbb{P}\Big(\Big[\big[\min_{k\in\{1,2,\ldots,K\}} f(\Theta^k_{n,\Theta_0^k})\big] - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \frac{\varepsilon}{2}\Big) + \sum_{i=1}^{K} \mathbb{P}\Big(\big[f(\mathit{\Theta}_n) - f(\Theta^i_{n,\Theta_0^i}) \geq \frac{\varepsilon}{2}, B_i\Big).
\end{aligned} \tag{244}$$

For the first term on the righthand side of (244), Corollary 28 proves that there exists $c \in (0, \infty)$ which satisfies that

$$\begin{aligned}
&\mathbb{P}\Big(\Big[\big[\min_{k\in\{1,2,\ldots,K\}} f(\Theta^k_{n,\Theta_0^k})\big] - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \frac{\varepsilon}{2}\Big) \\
&\leq \left(\frac{\big|A\backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c\left(\varepsilon^{-2}n^{-\rho} + M^{-1}n^{1-\rho} + \frac{r\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}\right)\right)^K.
\end{aligned} \tag{245}$$

We will now estimate the second term on the righthand side of (245). Let $\tilde{B}_j \subseteq \Omega$, $j \in \{1, 2, \ldots, K\}$, be disjoint events which satisfy that $\Omega = \coprod_{j\in\{1,2,\ldots,K\}} \tilde{B}_j$ and that

$$\tilde{B}_j \subseteq \Big\{\omega \in \Omega\colon \sum_{m=1}^{\mathfrak{M}} F(\mathit{\Theta}_n(\omega), X_{n+1,m}(\omega)) = \sum_{m=1}^{\mathfrak{M}} F(\Theta^j_{n,\Theta_0^j}(\omega), X_{n+1,m}(\omega))\Big\}. \tag{246}$$

Since the events $\tilde{B}_j$, $j \in \{1, 2, \ldots, K\}$, are disjoint, the final term of (244) satisfies that

$$\sum_{i=1}^{K} \mathbb{P}\Big(f(\mathit{\Theta}_n) - f(\Theta^i_{n,\Theta_0^i}) \geq \frac{\varepsilon}{2}, B_i\Big) = \sum_{i,j=1}^{K} \mathbb{P}\Big(f(\Theta^j_{n,\Theta_0^j}) - f(\Theta^i_{n,\Theta_0^i}) \geq \frac{\varepsilon}{2}, B_i, \tilde{B}_j\Big). \tag{247}$$

Let $F^{\mathfrak{M},n}\colon \mathbb{R}^d \times \Omega \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$, $\omega \in \Omega$ that

$$F^{\mathfrak{M},n}(\theta, \omega) = \frac{1}{\mathfrak{M}} \sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega)). \tag{248}$$

For every $i, j \in \{1, 2, \ldots, K\}$, since it holds for every $\omega \in B_i \cap \tilde{B}_j$ that

$$F^{\mathfrak{M},n}(\Theta^j_{n,\Theta_0^j}(\omega), \omega) - F^{\mathfrak{M},n}(\Theta^i_{n,\Theta_0^i}(\omega), \omega) \leq 0, \tag{249}$$

it holds for every $i, j \in \{1, 2, \ldots, K\}$ that

$$\begin{aligned}
&\mathbb{P}\Big(f(\Theta^j_{n,\Theta_0^j}) - f(\Theta^i_{n,\Theta_0^i}) \geq \frac{\varepsilon}{2}, B_i, \tilde{B}_j\Big) \\
&\leq \mathbb{P}\Big(f(\Theta^j_{n,\Theta_0^j}(\omega)) - F^{\mathfrak{M},n}(\Theta^j_{n,\Theta_0^j}(\omega), \omega) + F^{\mathfrak{M},n}(\Theta^i_{n,\Theta_0^i}(\omega), \omega) - f(\Theta^i_{n,\Theta_0^i}(\omega)) \geq \frac{\varepsilon}{2}, B_i, \tilde{B}_j\Big) \\
&\leq \mathbb{P}\Big(\Big|f(\Theta^j_{n,\Theta_0^j}(\omega)) - F^{\mathfrak{M},n}(\Theta^j_{n,\Theta_0^j}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, B_i, \tilde{B}_j\Big) \\
&\quad + \mathbb{P}\Big(\Big|f(\Theta^i_{n,\Theta_0^i}(\omega)) - F^{\mathfrak{M},n}(\Theta^i_{n,\Theta_0^i}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, B_i, \tilde{B}_j\Big).
\end{aligned} \tag{250}$$

It follows from (247) and (250) that

$$\begin{aligned}
&\sum_{i=1}^{K} \mathbb{P}\Big(f(\Theta_n) - f(\Theta^i_{n,\Theta^i_0}) \geq \frac{\varepsilon}{2}, B_i\Big) \\
&\leq \sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta^j_{n,\Theta^j_0}(\omega)) - F^{\mathfrak{M},n}(\Theta^j_{n,\Theta^j_0}(\omega),\omega)\Big| \geq \frac{\varepsilon}{4}, \tilde{B}_j\Big) \\
&\quad + \sum_{i=1}^{K} \mathbb{P}\Big(\Big|f(\Theta^i_{n,\Theta^i_0}(\omega)) - F^{\mathfrak{M},n}(\Theta^i_{n,\Theta^i_0}(\omega),\omega)\Big| \geq \frac{\varepsilon}{4}, B_i\Big).
\end{aligned} \tag{251}$$

For the first term on the righthand side of (251), it holds that

$$\begin{aligned}
&\sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta^j_{n,\Theta^j_0}(\omega)) - F^{\mathfrak{M},n}(\Theta^j_{n,\Theta^j_0}(\omega),\omega)\Big| \geq \frac{\varepsilon}{4}, \tilde{B}_j\Big) \\
&\leq \sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta^j_{n,\Theta^j_0}(\omega)) - F^{\mathfrak{M},n}(\Theta^j_{n,\Theta^j_0}(\omega),\omega)\Big| \geq \frac{\varepsilon}{4}\Big).
\end{aligned} \tag{252}$$

Since the random variables $\{\Theta^k_{n,\Theta^k_0}\}_{k\in\mathbb{N}}$ and $\{X_{n+1,1,k}\}_{k\in\mathbb{N}}$ are independent and since the distribution of Θ^1_{n,Θ^1_0} has bounded support on $\mathbb{R}^d$, for the distribution μ_n of Θ^1_{n,Θ^1_0} on $\mathbb{R}^d$, Lemma 19, Chebyshev's inequality, and the definition of $F^{\mathfrak{M},n}$ prove that that there exists $c \in (0,\infty)$ which satisfies for every $j \in \{1,\ldots,K\}$ that

$$\begin{aligned}
&\mathbb{P}\Big(\Big|f(\Theta^j_{n,\Theta^j_0}(\omega)) - F^{\mathfrak{M},n}(\Theta^j_{n,\Theta^j_0}(\omega),\omega)\Big| \geq \frac{\varepsilon}{4}\Big) \\
&= \int_{\mathbb{R}^d} \mathbb{P}\Big(\Big|f(\theta) - \frac{1}{\mathfrak{M}}\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m})\Big| \geq \frac{\varepsilon}{4}\Big)\mu_n(\mathrm{d}\,\theta) \\
&\leq \frac{c}{\varepsilon^2\mathfrak{M}}.
\end{aligned} \tag{253}$$

Therefore, it holds that

$$\sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta^j_{n,\Theta^j_0}(\omega)) - F^{\mathfrak{M},n}(\Theta^j_{n,\Theta^j_0}(\omega),\omega)\Big| \geq \frac{\varepsilon}{4}, \tilde{B}_j\Big) \leq \frac{cK}{\varepsilon^2\mathfrak{M}}. \tag{254}$$

For the second term on the righthand side of (251), it is sufficient to apply the same argument, which proves that there exists $c \in (0,\infty)$ which satisfies that

$$\sum_{i=1}^{K} \mathbb{P}\Big(\Big|f(\Theta^i_{n,\Theta^i_0}(\omega)) - F^{\mathfrak{M},n}(\Theta^i_{n,\Theta^i_0}(\omega),\omega)\Big| \geq \frac{\varepsilon}{4}, B_i\Big) \leq \frac{cK}{\varepsilon^2\mathfrak{M}}. \tag{255}$$

Returning to (247), it follows from (251) and (254) that there exists $c \in (0,\infty)$ which satisfies that

$$\sum_{i=1}^{K} \mathbb{P}\Big(f(\Theta_n) - f(\Theta^i_{n,\Theta^i_0}) \geq \frac{\varepsilon}{2}, B_i\Big) \leq \frac{cK}{\varepsilon^2\mathfrak{M}}. \tag{256}$$

Returning finally to (244), it follows from (245) and (256) that there exists $c \in (0,\infty)$ which satisfies that

$$\mathbb{P}\Big(\Big[f(\Theta_n) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{cK}{\varepsilon^2\mathfrak{M}} + \left(\frac{\big|A\backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c\left(\varepsilon^{-2}n^{-\rho} + M^{-1}n^{1-\rho} + \frac{r\left(1+M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2}-2\delta\right)_+}\right)\right)^K, \tag{257}$$

which completes the proof of Theorem 30. ∎

In the final corollary of this section, we will compute the computational efficiency of the algorithm proposed in Theorem 30. The constant implicitly depends on the computational cost of computing F and $\nabla_\theta F$ and initializing the random variable $X_{1,1,1}$, but it does not depend upon the running time $n \in \mathbb{N}$, the sampling size $K \in \mathbb{N}$, or the mini-batch sizes $M, \mathfrak{M} \in \mathbb{N}$. We observe that, for learning rates ρ close to 1, the estimate below shows that the mini-batch size can essentially be considered an order one quantity.

Corollary 31 *In the setting of Theorem* 30*, let* $x_0 \in (\mathcal{M}\cap U)$*,* $\rho \in (2/3, 1)$*, and let* $A \subseteq \mathbb{R}^d$ *be a non-empty open set. In the conclusion of Theorem* 30*, choose* $R_0 \in (0,\infty)$ *and* $\delta_0 \in (0,1)$ *such that* $R_0/2 - \delta_0 > 0$ *and assume that*

$$\frac{\big|A\backslash V_{R_0/2,\delta_0}(x_0)\big|}{|A|} \in (0,1). \tag{258}$$

Then there exist $\mathfrak{r}_1 \in (0,\mathfrak{r}]$ *and* $\{c_i \in (0,\infty)\}_{i\in\{1,2,3,4\}}$ *such that, for every* $\varepsilon,\eta \in (0,1)$*, for* $n(\varepsilon), M(\varepsilon), K(\eta), \mathfrak{M}(\varepsilon,\eta) \in \mathbb{N}$ *defined by*

$$n(\varepsilon) = c_1\varepsilon^{-2/\rho}, \quad M(\varepsilon) = c_2\varepsilon^{-4/\rho+4}, \quad \mathfrak{M}(\varepsilon,\eta) = c_3\varepsilon^{-2}\eta^{-1}\left|\log(\eta)\right|, \quad \text{and} \quad K = c_4\left|\log(\eta)\right|, \tag{259}$$

we have, for every $r \in (0,\mathfrak{r}_1]$*,*

$$\mathbb{P}\Big(\Big[f(\Theta_{n(\varepsilon)} - \inf_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \eta. \tag{260}$$

Proof [Proof of Corollary 31] We define the constant

$$\frac{\big|A\backslash V_{R_0/2,\delta_0}(x_0)\big|}{|A|} = c_1 \in (0,1). \tag{261}$$

Theorem 30 and $R_0/2 - \delta_0 > 0$ prove that there exists $\overline{c} \in (0,\infty)$ such that, for every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $\varepsilon \in (0,1)$, and $r \in (0,\mathfrak{r}]$,

$$\mathbb{P}\Big(\Big[f(\Theta_n) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{\overline{c}K}{\varepsilon^2\mathfrak{M}} + \left(c_1 + \overline{c}\left(\varepsilon^{-2}n^{-\rho} + M^{-\frac{1}{2}}n^{1-\rho} + r\right)\right)^K. \tag{262}$$

Fix $\mathfrak{r}_1 \in (0,\overline{r}]$ such that

$$c_1 + \overline{c}\mathfrak{r}_1 = c_2 \in (0,1).$$

We then have, for every $r \in (0,\mathfrak{r}_1]$,

$$\mathbb{P}\Big(\Big[f(\Theta_n) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{\overline{c}K}{\varepsilon^2\mathfrak{M}} + \left(c_2 + \overline{c}\left(\varepsilon^{-2}n^{-\rho} + M^{-\frac{1}{2}}n^{1-\rho}\right)\right)^K. \tag{263}$$

The statement now follows by a direct computation, which completes the proof of Corollary 31. ∎

6. Stochastic gradient descent - The compact case

In this section, we will analyze the converge of SGD to the manifold of local minima under the additional assumption that the manifold of local minima is compact. The essential difference in this case is that SGD cannot leave a basin of attraction along directions tangential to the manifold, and therefore the results apply for every $\rho \in (0,1)$. The proofs are essentially identical, after taking Remark 22 into account. We therefore record only the main results.

Theorem 32 *Let $d \in \mathbb{N}$, let $(S, \mathcal{S})$ be a measurable space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ be measurable, and let $\{X_{n,m,k}\colon \Omega \to \mathbb{R}\}_{n,m,k\in\mathbb{N}}$ be i.i.d. random variables. Assume that F and $X_{1,1,1}$ satisfy Assumption 2. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\theta) = \mathbb{E}[F(\theta, X_{1,1,1})]$, assume that f satisfies Assumption 1, and assume that $\mathcal{M} \cap U$ is compact. For every $k, M \in \mathbb{N}$, $\rho \in (0,1)$, $r \in (0,\infty)$, and $\theta \in \mathbb{R}^d$ let $\{\Theta^k_{n,\theta} = \Theta^k_{n,\theta}(M,\rho,r)\}_{n\in\mathbb{N}_0}$ be defined by $\Theta^k_{0,\theta} = \theta$ and, for every $n \in \mathbb{N}$,*

$$\Theta^k_{n,\theta} = \Theta^k_{n-1,\theta} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta^k_{n-1,\theta}, X_{n,m,k})\right]. \tag{264}$$

Let $A \subseteq \mathbb{R}^d$ be a non-empty, bounded open set and let $\{\Theta^k_0\colon \Omega \to A\}_{k\in\mathbb{N}}$ be i.i.d. uniformly distributed random variables on A that are independent of $\{X_{n,m,k}\}_{n,m,k\in\mathbb{N}}$. For every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $\rho \in (0,1)$, and $r \in (0,\infty)$ let $\{\Theta_n = \Theta(M,\mathfrak{M},K,\rho,r)\colon \Omega \to \mathbb{R}^d\}_{n\in\mathbb{N}_0}$ be a random variables which satisfy that

$$\frac{1}{\mathfrak{M}}\sum_{m=1}^{\mathfrak{M}} F(\Theta_n, X_{n+1,1,m}) = \Big[\min_{k\in\{1,2,\ldots,K\}} \Big(\frac{1}{\mathfrak{M}}\sum_{m=1}^{\mathfrak{M}} F(\Theta^k_{n,\Theta^k_0}, X_{n+1,1,m})\Big)\Big]. \tag{265}$$

Then for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (0,1)$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0,\infty)$ such that, for every $R \in (0, R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M, \mathfrak{M}, K \in \mathbb{N}$, $\varepsilon \in (0,1)$,

$$\begin{aligned}&\mathbb{P}\Big(\Big[f(\Theta_n) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big)\\ &\leq \frac{cK}{\varepsilon^2\mathfrak{M}} + \left(\frac{\big|A\backslash V_{R/2,\delta}(x_0)\big|}{|A|} + c\varepsilon^{-2}n^{-\rho} + 1 - c\exp\Big(-\frac{c}{M}\Big) + cM^{-1}n^{1-\rho}\right)^K.\end{aligned} \tag{266}$$

Corollary 33 *In the setting of Theorem 32, for every $x_0 \in (\mathcal{M} \cap U)$ and $\rho \in (0,1)$ there exist $R_0, \delta_0, \mathfrak{r} \in (0,\infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$ there exist $\{c_i \in (0,\infty)\}_{i\in\{1,2,3,4\}}$ such that for every $\varepsilon, \eta \in (0,1)$, for $n(\varepsilon), M(\varepsilon), K(\eta), \mathfrak{M}(\varepsilon,\eta) \in \mathbb{N}$ defined by*

$$n(\varepsilon) = c_1\varepsilon^{-2/\rho},\ \ M(\varepsilon) = c_2\varepsilon^{-2/\rho+2},\ \ \mathfrak{M}(\varepsilon,\eta) = c_3\varepsilon^{-2}\eta^{-1}\left|\log(\eta)\right|,\ \ \text{and}\ \ K = c_4\left|\log(\eta)\right|, \tag{267}$$

we have that

$$\mathbb{P}\Big(\Big[f(\Theta_{n(\varepsilon)}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \eta. \tag{268}$$

7. Applications

In this section, we prove that the Assumptions 1 and 2 are satisfied for some (simple) objective functions that arise in the training of neural networks. We will first consider a four-parameter affine-linear network with a linear activation function, for which the set of global minima is a locally smooth codimension 2 submanifold of the parameter space. We will then consider a two-parameter network with the ReLU activation function, for which the set of minima is a locally smooth codimension 1 submanifold of the parameter space.

We observe in particular that this implies the global minima are not locally unique, and therefore the established convergence results such as those proven in Dereich and Mueller-Gronbach (2015); Jentzen et al. (2018) do not apply.

Proposition 34 *Let $\varphi \in L^2([0,1])$ be finite, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\{X_{n,m}\}_{n,m\in\mathbb{N}}$ be i.i.d. random variables that are uniformly distributed on $[0,1]$, let $f\colon \mathbb{R}^4 \to \mathbb{R}$ be defined by*

$$f(\theta) = \int_0^1 |\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x)|^2 \, \mathrm{d}x, \tag{269}$$

and let $F\colon \mathbb{R}^4 \times [0,1] \to \mathbb{R}$ be defined by

$$F(\theta, x) = |\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x)|^2 . \tag{270}$$

Then f satisfies Assumption 1 and F and $X_{1,1}$ satisfy Assumption 2.

Proof [Proof of Proposition 34] Let $\varphi \in L^2([0,1])$ be finite. The finiteness of φ proves that, for every $x \in [0,1]$, we have $F(\cdot, x) \in \mathrm{C}^{0,1}_{\mathrm{loc}}(\mathbb{R}^4)$. It follows by the uniform distribution of the $X_{n,m}$, $n, m \in \mathbb{N}$, on $[0,1]$ that $f(\cdot) = \mathbb{E}[F(\cdot, X_{1,1})]$, and it follows from the L^2-integrability of φ that for every compact subset $\mathfrak{C} \subseteq \mathbb{R}^4$ it holds that

$$\sup_{\theta\in\mathfrak{C}} \mathbb{E}\left[|F(\theta, X_{1,1})|^2 + |\nabla_\theta F(\theta, X_{1,1})|^2 \right] < \infty. \tag{271}$$

It follows by the definition of f and $\varphi \in L^2([0,1])$ that $f \in \mathrm{C}^3_{\mathrm{loc}}(\mathbb{R}^4)$. It remains to characterize the set of minima of f. We first observe that when minimizing f, it is sufficient to minimize the potential over the set $\{\theta_3 \neq 0\}$. To see this, suppose that $\theta = (\theta_1, \theta_2, 0, \theta_4)$. Then for $\tilde{\theta} = (0, 0, 1, \theta_4)$ it holds that

$$f(\theta) = \int_0^1 |\theta_4 - \varphi(x)|^2 \, \mathrm{d}x = f(\tilde{\theta}). \tag{272}$$

Therefore, it holds that

$$\inf_{\theta\in\mathbb{R}^4} f(\theta) = \inf_{\theta\in\{\theta_3\neq 0\}} f(\theta). \tag{273}$$

Let $\theta \in \mathbb{R}^4 \cap \{\theta_3 \neq 0\}$ be fixed but arbitrary. An explicit computation proves the critical points of f satisfy that

$$\nabla f(\theta) = 2\int_0^1 (\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x)) \begin{pmatrix} \theta_3 x \\ \theta_3 \\ \theta_1 x + \theta_2 \\ 1 \end{pmatrix} \mathrm{d}x = 0. \tag{274}$$

For $r_k \in \mathbb{R}$, $k \in \{0, 1\}$, which satisfy that

$$r_k = \int_0^1 x^k \varphi(x) \, \mathrm{d}x, \tag{275}$$

it follows that $\theta \in \mathbb{R}^4$ satisfies equation (274) if and only if it holds that

$$\begin{cases} \frac{1}{3}\theta_1\theta_3^2 + \frac{1}{2}\theta_2\theta_3^2 + \frac{1}{2}\theta_3\theta_4 - r_1\theta_3 = 0, \\ \frac{1}{2}\theta_1\theta_3^2 + \theta_2\theta_3^2 + \theta_3\theta_4 - r_0\theta_3 = 0, \\ \frac{1}{3}\theta_1^2\theta_3 + \frac{1}{2}\theta_1\theta_2\theta_3 + \frac{1}{2}\theta_1\theta_4 - r_1\theta_1 + \frac{1}{2}\theta_1\theta_2\theta_3 + \theta_2^2\theta_3 + \theta_2\theta_4 - r_0\theta_2 = 0, \\ \frac{1}{2}\theta_1\theta_3 + \theta_2\theta_3 + \theta_4 - r_0 = 0. \end{cases} \tag{276}$$

For $\theta \in \mathbb{R}^4$ which satisfies that $\theta_3 \neq 0$, an explicit computation proves that θ satisfies system (276) if and only if it holds that

$$\theta_1\theta_3 = -6(r_0 - 2r_1) \text{ and } \theta_4 = -\theta_2\theta_3 + 4r_0 - 6r_1. \tag{277}$$

For $U \subseteq \mathbb{R}^4$ which satisfies that

$$U = \{\theta \in \mathbb{R}^4 \colon \theta_3 \neq 0\}, \tag{278}$$

for $\mathcal{M} \subseteq \mathbb{R}^4$ which satisfies that

$$\mathcal{M} = \{\theta \in \mathbb{R}^4 \colon f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^4} f(\vartheta)\}, \tag{279}$$

we claim that

$$\mathcal{M} \cap U = \{\, \theta \in \mathbb{R}^4 \colon \theta \text{ satisfies (277) and } \theta_3 \neq 0 \,\}. \tag{280}$$

Let $\theta \in \mathbb{R}^4$ satisfy (277) and $\theta_3 \neq 0$. Proceeding by contradiction, suppose that there exists $\theta_0 = (\theta_{1,0}, \theta_{2,0}, \theta_{3,0}, \theta_{4,0})$ which satisfies $\theta_{3,0} \neq 0$ such that

$$f(\theta_0) < f(\theta). \tag{281}$$

Since an explicit computation proves for every $(\theta_1, \theta_4) \in \mathbb{R}^2$ that

$$\lim_{|(\theta_1,\theta_4)| \to \infty} f(\theta_1, \theta_{2,0}, \theta_{3,0}, \theta_4) = \infty, \tag{282}$$

the identical considerations leading to (277) prove that

$$(\theta_1, \theta_4) \in \mathbb{R}^2 \mapsto f(\theta_1, \theta_{2,0}, \theta_{3,0}, \theta_4), \tag{283}$$

is uniquely minimized, owing to $\theta_{3,0} \neq 0$, by $(\theta_1, \theta_4) \in \mathbb{R}^2$ which satisfies that

$$\theta_1 = -\frac{6(r_0 - 2r_1)}{\theta_{3,0}} \text{ and } \theta_4 = -\theta_{2,0}\theta_{3,0} + 4r_0 + 6r_1. \tag{284}$$

We conclude that $\tilde{\theta}_0 \in \mathbb{R}^4$ satisfies that

$$\tilde{\theta}_0 = (-\frac{6(r_0 - 2r_1)}{\theta_{3,0}}, \theta_{2,0}, \theta_{3,0}, -\theta_{2,0}\theta_{3,0} + 4r_0 + 6r_1), \tag{285}$$

satisfies (277) and $\tilde{\theta}_{3,0} \neq 0$. Therefore, it holds that

$$f(\tilde{\theta}_0) < f(\theta_0), \tag{286}$$

which contradicts the fact that $\nabla f = 0$ on the connected set of $\theta \in \mathbb{R}^4$ which satisfies (277) and $\theta_3 \neq 0$. This proves (280). It is immediate from (277) that $\mathcal{M} \cap U$ is a non-empty, 2-dimensional, C^2-submanifold of $\mathbb{R}^4$. It remains only to prove the nondegeneracy assumption. for every $\theta \in (\mathcal{M} \cap U)$, after computing the Hessian[3], it holds that

$$\begin{aligned}
\big(\operatorname{Hess} f\big)(\theta) &= 2\int_0^1 \begin{pmatrix} \theta_3^2 x^2 & \theta_3^2 x & \theta_1\theta_3 x^2 + \theta_2\theta_3 x & \theta_3 x \\ & \theta_3^2 & \theta_1\theta_3 x + \theta_2\theta_3 & \theta_3 \\ & & (\theta_1 x + \theta_2)^2 & \theta_1 x + \theta_2 \\ & & & 1 \end{pmatrix} \mathrm{d}x \\
&= \begin{pmatrix} \frac{2}{3}\theta_3^2 & \theta_3^2 & \frac{2}{3}\theta_1\theta_3 + \theta_2\theta_3 & \theta_3 \\ & 2\theta_3^2 & \theta_1\theta_3 + 2\theta_2\theta_3 & 2\theta_3 \\ & & \frac{2}{3}\theta_1^2 + 2\theta_1\theta_2 + 2\theta_2^2 & \theta_1 + 2\theta_2 \\ & & & 2 \end{pmatrix},
\end{aligned} \tag{287}$$

[3]Due to the symmetry of the Hessian, we only write the upper diagonal.

where this equality relies upon the fact that, due to (274) and $\theta_3 \neq 0$ on $\mathcal{M} \cap U$, we have that

$$\int_0^1 (\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x))\,\mathrm{d}x = \int_0^1 (\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x))x\,\mathrm{d}x = 0. \tag{288}$$

A column-reduction, which relies on the fact that for every $\theta \in (\mathcal{M} \cap U)$ we have $\theta_3 \neq 0$, proves for every $\theta \in (\mathcal{M} \cap U)$ that

$$\operatorname{rank}((\operatorname{Hess} f)(\theta)) = 2 = \operatorname{codim}(\mathcal{M} \cap U). \tag{289}$$

This completes the proof of Proposition 34. ∎

Proposition 35 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\{X_{n,m}\}_{n,m\in\mathbb{N}}$ be i.i.d. random variables that are uniformly distributed on $[0,1]$, let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by*

$$f(\theta) = \int_0^1 |\theta_2 \max(\theta_1 x, 0) - \sin(x)|^2 \,\mathrm{d}x, \tag{290}$$

and let $F\colon \mathbb{R}^2 \times [0,1] \to \mathbb{R}$ be defined by

$$F(\theta, x) = |\theta_2 \max(\theta_1 x, 0) - \sin(x)|^2. \tag{291}$$

Then f satisfies Assumption 1 and F and $X_{1,1}$ satisfy Assumption 2.

Proof [Proof of Proposition 35] It is immediate that $F(\cdot, x) \in \mathrm{C}^{0,1}_{\mathrm{loc}}(\mathbb{R}^2)$. Since the $\{X_{n,m}\}_{n,m\in\mathbb{N}}$ are uniformly distributed on $[0,1]$, for every $\theta \in \mathbb{R}^2$,

$$f(\theta) = \mathbb{E}[F(\theta, X_{1,1})], \tag{292}$$

and, furthermore, a straightforward computation proves for every compact set $\mathfrak{C} \subseteq \mathbb{R}^2$ that

$$\sup_{\theta\in\mathfrak{C}} \mathbb{E}\left[|F(\theta, X_{1,1})|^2 + |\nabla_\theta F(\theta, X_{1,1})|^2\right] < \infty. \tag{293}$$

It remains only to characterize the minima of the objective function, and to verify the nondegeneracy condition. An explicit computation proves that, when minimizing f, it is sufficient to restrict to the set $\{\theta_1 > 0, \theta_2 > 0\}$. Let $U \subseteq \mathbb{R}^2$ satisfy that

$$U = \{\theta \in \mathbb{R}^2 \colon \theta_1 > 0, \theta_2 > 0\}. \tag{294}$$

We observe for every $\theta \in U$ that

$$f(\theta) = \int_0^1 |\theta_1\theta_2 x - \sin(x)|^2 \,\mathrm{d}x, \tag{295}$$

and for every $\theta \in U$ that

$$\nabla f(\theta) = 2\int_0^1 (\theta_1\theta_2 x - \sin(x)) \begin{pmatrix} \theta_2 x \\ \theta_1 x \end{pmatrix} \mathrm{d}x. \tag{296}$$

Therefore, for $\theta \in U$ it holds that $\nabla f(\theta) = 0$ if and only if it holds that

$$\theta_1\theta_2 = 3\int_0^1 x\sin(x)\,\mathrm{d}x = 3(\sin(1) - \cos(1)). \tag{297}$$

Let $\mathcal{M} \subseteq \mathbb{R}^2$ satisfy that

$$\mathcal{M} = \{\theta \in \mathbb{R}^2 \colon f(\theta) = \inf_{\vartheta \in \mathbb{R}^4} f(\vartheta)\}. \tag{298}$$

We claim that

$$\mathcal{M} \cap U = \{\, \theta \in \mathbb{R}^2 \colon \theta \text{ satisfies (297)},\ \theta_1 > 0, \text{ and } \theta_2 > 0\}. \tag{299}$$

Suppose that $\theta \in U$ satisfies (297). By contradiction suppose that there exists $\theta_0 = (\theta_{1,0}, \theta_{2,0}) \in \{\theta_1 > 0, \theta_2 > 0\}$ such that

$$f(\theta_0) < f(\theta). \tag{300}$$

Since $\theta_{1,0} > 0$ an explicit computation proves that

$$\lim_{\theta_2 \to \infty} f(\theta_{1,0}, \theta_2) = +\infty \ \text{ and } \ f(\theta_{1,0}, 0) > f(\theta_0). \tag{301}$$

The arguments leading from (295) to (297) prove that (301) is uniquely minimized when

$$\theta_2 = \frac{3}{\theta_{1,0}}(\sin(1) - \cos(1)). \tag{302}$$

Therefore, for $\tilde{\theta}_0 \in \mathbb{R}^2$ which satisfies that

$$\tilde{\theta}_0 = (\theta_{1,0}, \frac{3}{\theta_{1,0}}(\sin(1) - \cos(1))), \tag{303}$$

we have that $\tilde{\theta}_0 \in U$, that $\tilde{\theta}_0$ satisfies (297), and that

$$f(\tilde{\theta}_0) \leq f(\theta_0) < f(\theta). \tag{304}$$

This contradicts the fact that $\nabla f = 0$ on the connected set of $\theta \in U$ that satisfy (297). This proves (299). Since it is clear that $\mathcal{M} \cap U$ is a non-empty, 1-dimensional, C^2-submanifold of $\mathbb{R}^2$, it remains only to establish the nondegeneracy assumption. For every $\theta \in (\mathcal{M} \cap U)$ it holds that

$$\begin{aligned} \big(\operatorname{Hess} f\big)(\theta) &= 2 \begin{pmatrix} \frac{1}{3}\theta_2^2 & \frac{2}{3}\theta_1\theta_2 - (\sin(1) - \cos(1)) \\ & \frac{1}{3}\theta_1^2 \end{pmatrix} \\ &= 2 \begin{pmatrix} \frac{1}{3}\theta_2^2 & \sin(1) - \cos(1) \\ & \frac{3(\sin(1) - \cos(1))^2}{\theta_2^2} \end{pmatrix}. \end{aligned} \tag{305}$$

A column reduction and $\theta_2 \neq 0$ prove for every $\theta \in (\mathcal{M} \cap U)$ that

$$\operatorname{rank}((\operatorname{Hess} f)(\theta)) = 1 = \operatorname{codim}(\mathcal{M} \cap U). \tag{306}$$

This completes the proof of Proposition 35. ∎

Acknowledgements

The first author acknowledges financial support from the National Science Foundation Mathematical Sciences Postdoctoral Research Fellowship under Grant Number 1502731. The second author acknowledges financial support by the DFG through the CRC 1283 "Taming uncertainty and profiting from randomness and low regularity in analysis, stochastics and their applications."

References

S.-I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10:1116–1135, 2000.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15:595–627, 2014.

F. Bach and E Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. *Advances in neural information processing systems*, pages 773–781, 2013.

B. Bercu and J.-C. Fort. Generic stochastic gradient methods. *Wiley Encyclopedia of Operations Research and Management Science*, pages 1–8, 2013.

L. Bottou. Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT 2010*, pages 177–186, 2010.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Optimization for Machine Learning*, pages 351–368, 2011.

L. Bottou and Y. LeCun. Large scale online learning. *Advances in Neural Information Processing Systems*, 16, 2004.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60:223–311, 2018.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

C. Darken, J. Chang, and J. Moody. Learning rate schedules for faster stochastic gradient search. *Neural Networks for Signal Processing II: Proceedings of the 1992 IEEE Workshop*, pages 1–11, 1992.

J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M.C. Mao, M.A. Ranzato, A. Senior, P. Tucker, K. Yang, and A.Y. Ng. Large scale distributed deep networks. *Advances in Neural Information Processing Systems*, pages 1–11, 2012.

L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, and J. Williams. Recent advances in deep learning for speech research at microsoft. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

S. Dereich and T. Mueller-Gronbach. General multilevel adaptations for stochastic approximation algorithms. *arXiv:1506.05482*, 2015.

A. Dieuleveut, A. Durmus, and B. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *HAL:01565514*, 2017.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

R. L. Foote. Regularity of the distance function. *Proc. Amer. Math. Soc.*, 92:153–155, 1984.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points–online stochastic gradient for tensor decomposition. *Conference on Learning Theory*, pages 797–842, 2015.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23:2341–2368, 2013.

S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *arXiv:1308.6594*, 2013.

A. Graves. Generating sequences with recurrent neural networks. *arXiv:1308.0850*, 2013.

A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. *Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T.N. Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine*, 29:82–97, 2012.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

M. Inoue, H. Park, and M. Okada. On-line learning theory of soft committee machines with correlated hidden units steepest gradient descent and natural gradient descent. *Journal of the Physical Society of Japan*, 72:805–810, 2003.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

A. Jentzen, B. Kuckuck, A. Neufeld, and P. von Wurstemberger. Strong error analysis for stochastic gradient descent optimization algorithms. *arXiv:1801.09324*, 2018.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. *arXiv:1703.00887*, 2017.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. *European Conference on Machine Learning and Knowledge Discovery in Databases*, 9851:795–811, 2016.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

Y. Lei, T. Hu, G. Li, and K. Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *arXiv:1902.00908*, 2019.

H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31:6389–6399, 2018.

X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. *arXiv:1805.08114*, 2018.

J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *arXiv:1311.1873*, 2013.

S. Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivés partielles*, 117:87–89, 1963.

Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46:157–178, 1993.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *arXiv:1504.06298*, 2015.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19:1574–1609, 2009.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

R. Pascanu and Y. Bengio. Revisiting natural gradient for deep networks. *International Conference on Learning Representations*, 2014.

B.T. Polyak. Gradient methods for minimizing functionals. *Zh. Vychisl. Mat. Mat. Fiz.*, 3:643–653, 1963.

N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12: 145–151, 1999.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: A nonasymptotic analysis. *arXiv:1702.03849*, 2017.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv:1109.5647*, 2011.

M. Rattray, D. Saad, and S. I. Amari. Natural gradient descent for on-line learning. *Physical Review Letters*, 81:5461–5464, 1998.

S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. *International Conference on Machine Learning*, pages 314–323, 2016.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

G. Rotskoff and E. Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

S. Ruder. An overview of gradient descent optimization algorithms. *arXiv:1609.04747*, 2016.

T. Schaul, S. Zhang, and Y. LeCun. No more pesky learning rates. *arXiv:1206.1106*, 2012.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147, 2013.

C. Tang and C. Monteleoni. On the convergence rate of stochastic gradient descent for strongly convex functions. *Regularization, optimization, kernels, and support vector machines*, pages 159–175, 2015.

R. Vidal, J. Bruna, R. Giryes, and S. Soatto. Mathematics of deep learning. *arXiv:1712.04741*, 2017.

R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv:1806.01811*, 2018.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *arXiv:1707.06618*, 2017.

W. Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv:1301.3584*, 2013.

H. Zhang and W. Yin. Gradient methods for convex minimization: Better rates under weaker conditions. *arXiv:1303.4645*, 2013.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. *arXiv:1702.05575*, 2017.

D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv:1808.05671*, 2018a.

D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduction for nonconvex optimization. *arXiv:1806.07811*, 2018b.