Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

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

B. Fehrman and B. Gess and A. Jentzen

Research Report No. 2019-21
April 2019

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

CONVERGENCE RATES FOR THE STOCHASTIC GRADIENT DESCENT METHOD FOR NON-CONVEX OBJECTIVE FUNCTIONS

BENJAMIN FEHRMAN[1], BENJAMIN GESS[2], AND ARNULF JENTZEN[3]

[1]MATHEMATICAL INSTITUTE, UNIVERSITY OF OXFORD,
OXFORD, UNITED KINGDOM,
E-MAIL: BENJAMIN.FEHRMAN@MATHS.OX.AC.UK

[2] MAX PLANCK INSTITUTE FOR MATHEMATICS IN THE SCIENCES,
LEIPZIG, GERMANY,
FAKULTÄT FÜR MATHEMATIK, UNIVERSITÄT BIELEFELD,
BIELEFELD, GERMANY,
E-MAIL: BENJAMIN.GESS@MIS.MPG.DE

[3]SEMINAR FOR APPLIED MATHEMATICS, DEPARTMENT OF MATHEMATICS, ETH ZURICH
ZURICH, SWITZERLAND,
E-MAIL: ARNULF.JENTZEN@SAM.MATH.ETHZ.CH

Abstract. We prove the local convergence to minima and estimates on the rate of convergence for the stochastic gradient descent method in the case of not necessarily globally convex nor contracting objective functions. In particular, the results are applicable to simple objective functions arising in machine learning.

Contents

1. Introduction

Stochastic gradient descent algorithms (SGD), going back to [46], are the most common way to train neural networks. Despite their relevance to machine learning and much recent interest, estimates on their rate of convergence have so far only been shown under global contraction or convexity assumptions on the objective function that are often not satisfied by examples arising in machine learning. Indeed, citing from [52], “While SGD has been rigorously analyzed only

Date: April 5, 2019.

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 to the set of global minima of the objective function while avoiding such a global convexity or contractivity assumption. The relevance of the obtained results is demonstrated by the 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 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 (cf., e.g., [2, 4, 8, 13, 24, 35, 51] and the references therein). In [13, 24], the convergence of SGD with rates assuming the following contraction property for the objective function f, which is classical in stochastic approximation theory, was analyzed: There is an $L > 0$ and a zero θ^* of $\nabla_\theta f$ such that for every $\theta \in \mathbb{R}^d$ it holds that

$$(-\nabla_\theta f(\theta), \theta - \theta^*) \leq -L\|\theta - \theta^*\|^2. \tag{1.2}$$

In particular, this contraction property 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 actual objective functions arising 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 to 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 locally at sets of minima[1]. In the present work we provide estimates on the rate of convergence for SGD under assumptions avoiding a contraction property like (1.2).

Theorem 1.1. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, $\rho \in (2/3, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta, x) \in \mathbb{R}^d \times S}\colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m}\colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}, \tag{1.3}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a continuously differentiable function, assume that $f|_U\colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta \in \mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M} \cap U \cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $n \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $M \in \mathbb{N}$, $r \in (0, \infty)$ let $\Theta_n^{k,M,r}\colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$ that $(\Theta_{n-1}^{k,M,r})_{k \in \mathbb{N}}$ and $(X_{n,k})_{k \in \mathbb{N}}$ are independent, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{1,M,r}$ is continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{1,M,r}$ and

[1] We emphasize that this is disjoint from the recent works [8, 30, 53] 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.

$(X_{n,k})_{n,k\in\mathbb{N}}$ *are independent, assume for every* $n, M \in \mathbb{N}$, $r \in (0,\infty)$ *that*

$$\Theta_n^{1,M,r} = \Theta_{n-1}^{1,M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}^{1,M,r}, X_{n,m})\right], \tag{1.4}$$

and for every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $r \in (0,\infty)$ *let* $\varTheta_n^{K,M,\mathfrak{M},r}\colon \Omega \to \mathbb{R}^d$ *be a random variable which satisfies for every* $\omega \in \Omega$ *that*

$$\varTheta_n^{K,M,\mathfrak{M},r}(\omega) \in \left[\operatorname*{argmin}_{\theta\in\{\Theta_n^{k,M,r}(\omega)\colon k\in\{1,\ldots,K\}\}} \left[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega))\right]\right], \tag{1.5}$$

(cf. Lemma 5.10 *below). Then there exist* $\mathfrak{r}, c \in (0,\infty)$, $\kappa \in [0,1)$ *such that for every* $n, M, \mathfrak{M}, K \in \mathbb{N}$, $r \in (0,\mathfrak{r}]$, $\varepsilon \in (0,1]$ *it holds that*

$$\mathbb{P}\Big(\Big[f(\varTheta_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{cK}{\varepsilon^2\mathfrak{M}} + \left[\kappa + c\left(\frac{1}{\varepsilon^2 n^\rho} + \frac{n^{1-\rho}}{M^{1/2}}\right)\right]^K. \tag{1.6}$$

Remark 1.2. Theorem 1.1 is proven in Theorem 5.11 below, where the constant $\kappa \in [0,1)$ is identified precisely. The statement of Theorem 1.1 should be interpreted in the following way. We aim to minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$, where we assume that the set of minima

$$\mathcal{M} = \{\theta \in \mathbb{R}^d \colon f(\theta) = \big[\inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)\big]\}, \tag{1.7}$$

is locally smooth in the sense that there exists an open set $U \subseteq \mathbb{R}^d$ such that

$$\mathcal{M} \cap U \text{ is a non-empty } \mathfrak{d}\text{-dimensional } \mathrm{C}^1\text{-submanifold of } \mathbb{R}^d. \tag{1.8}$$

We furthermore assume that f is locally C^3 in a neighborhood of $\mathcal{M} \cap U$ and that the Hessian is maximally nondegenerate on $\mathcal{M} \cap U$ in the sense that for every $\theta \in \mathcal{M} \cap U$ it holds that

$$\operatorname{rank}\big(\big(\operatorname{Hess} f\big)(\theta)\big) = d - \mathfrak{d} = \operatorname{codim}(\mathcal{M} \cap U). \tag{1.9}$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, and let $X_{n,m}\colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables. We assume that there exists a measurable function $F\colon S \times \mathbb{R}^d \to \mathbb{R}$ satisfying for every $\theta \in \mathbb{R}^d$ that

$$f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]. \tag{1.10}$$

In particular, since it is oftentimes the case in practice that the deterministic gradient $\nabla f(\theta)$ cannot be computed or cannot be efficiently computed, the random gradient $\nabla_\theta F(\theta, X_{1,1})$ provides an efficiently computable stochastic approximation.

The initial data of SGD is sampled from a bounded open set $A \subseteq \mathbb{R}^d$ satisfying that $\mathcal{M} \cap U \cap A \neq \emptyset$. That is, for every mini-batch size $M \in \mathbb{N}$ and $r \in (0,\infty)$, the initial data $\Theta_0^{k,M,r}\colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, is uniformly distributed on A. We then compute independent solutions to SGD in the sense that for every $n \in \mathbb{N}$ it holds that $\Theta_n^{k,M,r}\colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, are i.i.d. random variables and that $\Theta_n^{1,M,r}\colon \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, satisfies that

$$\Theta_n^{1,M,r} = \Theta_{n-1}^{1,M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}^{1,M,r}, X_{n,m})\right]. \tag{1.11}$$

For a fixed terminal time $n \in \mathbb{N}$, for a sampling size $K \in \mathbb{N}$, the output of the algorithm at this point is the collection of values $\Theta_n^{k,M,r}$, $k \in \{1, 2, \ldots, K\}$. It remains to identify the value $\Theta_n^{k,M,r}$, $k \in \{1, 2, \ldots, K\}$, that minimizes the objective function.

Much as in the case of the gradient, since the objective function cannot be practically computed, for a terminal time $n \in \mathbb{N}$, for a mini-batch size $\mathfrak{M} \in \mathbb{N}$, we introduce the mini-batch approximation

$F^{\mathfrak{M},n}\colon \mathbb{R}^d \times \Omega \to \mathbb{R}$ satisfying for every $(\theta,\omega) \in \mathbb{R}^d \times \Omega$ that

$$F^{\mathfrak{M},n}(\theta,\omega) = \frac{1}{\mathfrak{M}} \sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega)). \tag{1.12}$$

We then identify the value $\Theta_n^{k,M,r}$, $k \in \{1,\ldots,K\}$, that minimizes $F^{\mathfrak{M},n}$ in the sense that we compute a random variable $\mathit{\Theta}_n^{K,M,\mathfrak{M},r}\colon \Omega \to \mathbb{R}^d$ satisfying for every $\omega \in \Omega$ that

$$\mathit{\Theta}_n^{K,M,\mathfrak{M},r}(\omega) \in \left[\operatorname*{argmin}_{\theta \in \{\Theta_n^{k,M,r}(\omega)\colon k \in \{1,2,\ldots,K\}\}} \left[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega))\right]\right]. \tag{1.13}$$

The assumption that $(\Theta_n^{k,M,r})_{k\in\{1,2,\ldots,K\}}$ and $(X_{n+1,m})_{m\in\mathbb{N}}$, are independent is only used here to prove that, with probability depending on $\mathfrak{M}, K \in \mathbb{N}$, the random variable $\mathit{\Theta}_n^{K,M,\mathfrak{M},r}$ is a minimizer of the objective function within the set $\Theta_n^{k,M,r}$, $k \in \{1,2,\ldots,K\}$.

The conclusion of Theorem 1.1 estimates the probability that $\mathit{\Theta}_n^{K,M,\mathfrak{M},r}$ is an $\varepsilon \in (0,1]$ minimizer of the objective function. Precisely, there exist $\mathfrak{r}, c \in (0,\infty)$, $\kappa \in [0,1)$ such that for every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $r \in (0,\mathfrak{r}]$, $\varepsilon \in (0,1]$ it holds that

$$\mathbb{P}\Big(\Big[f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{cK}{\varepsilon^2\mathfrak{M}} + \left[\kappa + c\left(\frac{1}{\varepsilon^2 n^\rho} + \frac{n^{1-\rho}}{M^{1/2}}\right)\right]^K. \tag{1.14}$$

The limit $\mathfrak{M} \to \infty$ corresponds to computing the minimizer of f exactly. If this can be done efficiently, then the first term on the righthand side of (1.14) vanishes.

The constant $\kappa \in [0,1)$, which we compute precisely in Theorem 5.11 below, quantifies two sources of error: the probability that the initial condition lies outside of a basin of attraction and a portion of the probability that SGD beginning in a basin of attraction fails to converge. In Remark 5.58 below and Section 6, we prove that the restriction $\rho \in (2/3, 1)$ can be extended to $\rho \in (0,1)$ under the additional assumption that $\mathcal{M} \cap U$ is a compact subset of $\mathbb{R}^d$. Finally, it is not necessary to assume that F is continuously differentiable, and this assumption can be replaced with the assumption that for every $x \in S$ we have that $F(\cdot, x)$ is a locally Lipschitz continuous function of $\theta \in \mathbb{R}^d$.

We observe that the computational efficiency of the algorithm can be estimated using Theorem 1.1. In particular, it follows from Corollary 5.12 below that there exist constants $c_i \in (0,\infty)$, $i \in \{1,2,3,4\}$, such that for every $\varepsilon, \eta \in (0,1]$, for $n(\varepsilon) \in \mathbb{N}_0, M(\varepsilon), \mathfrak{M}(\varepsilon,\eta), K(\eta) \in \mathbb{N}$ satisfying that

$$n(\varepsilon) = c_1\varepsilon^{-2/\rho}, \quad M(\varepsilon) = c_2\varepsilon^{-4/\rho+4}, \quad \mathfrak{M}(\varepsilon,\eta) = c_3\varepsilon^{-2}\eta^{-1}\left|\log(\eta)\right|, \quad \text{and} \quad K(\eta) = c_4\left|\log(\eta)\right|, \tag{1.15}$$

it holds that

$$\mathbb{P}\Big(\big[f(\mathit{\Theta}_{n(\varepsilon)}^{K(\eta),M(\varepsilon),\mathfrak{M}(\varepsilon,\eta),r}) - \inf_{\theta\in\mathbb{R}^d} f(\theta)\big] \geq \varepsilon\Big) \leq \eta. \tag{1.16}$$

For every bounded open set $A \subseteq \mathbb{R}^d$ satisfying that $\mathcal{M} \cap U \cap A$ is non-empty, for every $\varepsilon, \eta \in (0,1]$, the computational efficiency of the algorithm $\mathrm{Eff}(\varepsilon,\eta;A) \in \mathbb{N}$ satisfies that

$$\mathrm{Eff}(\varepsilon,\eta;A) = \#\text{ computations sufficient to ensure (1.16)}. \tag{1.17}$$

It follows from (1.15) that there exists $c \in (0,\infty)$ satisfying for every $\varepsilon, \eta \in (0,1]$ that

$$\mathrm{Eff}(\varepsilon,\eta;A) \leq c\big(\varepsilon^{-2}\eta^{-1}\left|\log(\eta)\right| + \varepsilon^{-6/\rho+4}\left|\log(\eta)\right|\big), \tag{1.18}$$

where the constant $c \in (0,\infty)$ depends on the computational cost of computing F and $\nabla_\theta F$ but not on the running time $n \in \mathbb{N}$, mini-batch size $M \in \mathbb{N}$, or sampling size $K \in \mathbb{N}$. Furthermore, we prove in Corollary 6.5 below that that computational efficiency can be improved in the case that the local manifold of minima is compact.

The estimate of Theorem 1.1 quantifies two sources of error. The first term on the righthand side of (1.6) quantifies the error introduced by the mini-batch approximation of the objective function. In the case that the objective function f can be efficiently computed, this error can be avoided by computing $\Theta_n^{K,M,\infty,r} : \Omega \to \mathbb{R}^d$ satisfying for every $\omega \in \Omega$ that

$$\Theta_n^{K,M,\infty,r}(\omega) \in \Big[\operatorname*{argmin}_{\theta \in \{\Theta_n^{k,M,r}(\omega) \colon k \in \{1,\dots,K\}\}} f(\theta) \Big], \tag{1.19}$$

for which it follows from Corollary 5.9 below that

$$\mathbb{P}\Big(\big[f(\Theta_n^{K,M,\infty,r}) - \inf_{\theta \in \mathbb{R}^d} f(\theta) \big] \geq \varepsilon \Big) \leq \left[\kappa + c \left(\frac{1}{\varepsilon^2 n^\rho} + \frac{n^{1-\rho}}{M^{1/2}} \right) \right]^K . \tag{1.20}$$

The second term on the rigththand side of (1.6) quantifies the failure of the solutions $\Theta_n^{k,M,r}$, $k \in \{1, 2, \dots, K\}$, to converge to within distance $\varepsilon \in (0, 1]$ to the local manifold of minima at time $n \in \mathbb{N}$. We quantify this error in Corollary 5.8 below, where we prove that

$$\mathbb{P}\Big(\big[\min_{k \in \{1,2,\dots,K\}} \inf_{\theta \in \mathcal{M} \cap U} \big| \Theta_n^{k,M,r} - \theta \big| \big] \geq \varepsilon \Big) \leq \left[\kappa + c \left(\frac{1}{\varepsilon^2 n^\rho} + \frac{n^{1-\rho}}{M^{1/2}} \right) \right]^K . \tag{1.21}$$

The methods of Corollary 5.12 below prove that there exist constants $c_i \in (0, \infty)$, $i \in \{1, 2, 3\}$, such that for every $\varepsilon, \eta \in (0, 1]$, for $n(\varepsilon) \in \mathbb{N}, M(\varepsilon), K(\eta) \in \mathbb{N}$ satisfying that

$$n(\varepsilon) = c_1 \varepsilon^{-2/\rho}, \;\; M(\varepsilon) = c_2 \varepsilon^{-4/\rho + 4}, \;\; \text{and} \;\; K(\eta) = c_3 \left| \log(\eta) \right| , \tag{1.22}$$

it holds that

$$\mathbb{P}\Big(\big[\min_{k \in \{1,2,\dots,K(\eta)\}} \inf_{\vartheta \in \mathcal{M} \cap U} \big| \Theta_{n(\varepsilon)}^{k,M(\varepsilon),r} - \vartheta \big| \big] \geq \varepsilon \Big) \leq \eta. \tag{1.23}$$

For every bounded open set $A \subseteq \mathbb{R}^d$ with $\mathcal{M} \cap U \cap A \neq \emptyset$, for every $\varepsilon, \eta \in (0, 1]$, the computational efficiency $\mathrm{Eff}_{\mathrm{SGD}}(\varepsilon, \eta; A) \in \mathbb{N}$ of (1.21) satisfies that

$$\mathrm{Eff}_{\mathrm{SGD}}(\varepsilon, \eta; A) = \#\ \text{computations sufficient to ensure (1.23)}. \tag{1.24}$$

It follows from (1.22) that for every bounded open set $A \subseteq \mathbb{R}^d$ with $\mathcal{M} \cap U \cap A \neq \emptyset$ there exists $c \in (0, \infty)$ such that for every $\varepsilon, \eta \in (0, 1]$ it holds that

$$\mathrm{Eff}_{\mathrm{SGD}}(\varepsilon, \eta; A) \leq c \big(\varepsilon^{-6/\rho + 4} \left| \log(\eta) \right| \big). \tag{1.25}$$

In the following remark, we compare this computational efficiency with that of a random sampling algorithm.

Remark 1.3. The computational efficiency $\mathrm{Eff}_{\mathrm{SGD}}$ yields a significant improvement when compared with a random sampling algorithm. Precisely, suppose that $A \subseteq \mathbb{R}^d$ is a bounded open subset with $\mathcal{M} \cap U \cap A \neq \emptyset$. Then, since $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional, C^1-submanifold of $\mathbb{R}^d$, for the Lebesgue-Borel measure $\lambda : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$, there exists $c \in (0, \infty)$ satisfying that

$$\frac{\lambda\big(\{ \theta \in A \colon \inf_{\vartheta \in \mathcal{M} \cap U} |x - \vartheta| \geq \varepsilon \} \big)}{\lambda(A)} \geq 1 - \frac{c \varepsilon^{d - \mathfrak{d}}}{\lambda(A)}. \tag{1.26}$$

If $\Theta^i : \Omega \to A$, $i \in \mathbb{N}$, are i.i.d. random variables that are continuous uniformly distributed on A, it follows from (1.26) that for every $K \in \mathbb{N}$ it holds that

$$\mathbb{P}\Big(\min_{i \in \{1,2,\dots,K\}} \inf_{\theta \in \mathcal{M} \cap U} \big| \Theta^i - \theta \big| \geq \varepsilon \Big) \geq \Big(1 - \frac{c \varepsilon^{d - \mathfrak{d}}}{\lambda(A)} \Big)^K . \tag{1.27}$$

For every $\varepsilon, \eta \in (0, 1]$, $K \in \mathbb{N}$, in order to ensure that

$$\mathbb{P}\Big(\min_{i \in \{1,2,\dots,K\}} \inf_{\theta \in \mathcal{M} \cap U} \big| \Theta^i - \theta \big| \geq \varepsilon \Big) \leq \eta, \tag{1.28}$$

it is necessary to choose $K(\varepsilon,\eta)\in\mathbb{N}$ satisfying that

$$K(\varepsilon,\eta)\geq\log\left(1-\frac{c\varepsilon^{d-\mathfrak{d}}}{\lambda(A)}\right)^{-1}\left|\log(\eta)\right|. \tag{1.29}$$

In particular, there exists $c\in(0,\infty)$ satisfying for every $\varepsilon\in(0,({}^{\lambda(A)}/{}_{2r})^{{}^1/{}_{d-\mathfrak{d}}}]$ that

$$K(\varepsilon,\eta)\geq c\varepsilon^{-(d-\mathfrak{d})}\left|\log(\eta)\right|. \tag{1.30}$$

The computational efficiency of the random sampling algorithm is therefore worse than $\mathrm{Eff}_{\mathrm{SGD}}$ whenever the codimension $d-\mathfrak{d}$ is greater than ${}^6/{}_\rho-4$. This condition is expected to be satisfied in all practical machine learning applications, where the dimension $d\in\mathbb{N}$ is large, since for $\rho\in({}^2/{}_3,1)$ we have ${}^6/{}_\rho-4<5$. In particular, this condition is satisfied for any $\rho\in({}^2/{}_3,1)$ if there exists a unique minimum and $d\geq5$.

In a non-globally stable setting, i.e. when (1.2) 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 satisfying (1.2) in a neighborhood V of the minimum, the global analysis put forward in [24] 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 basin of attraction V, outside of which no control on the dynamics can be expected. Therefore, it becomes necessary to provide estimates on the probability that SGD leaves favorable neighborhoods. Second, 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, which in turn relies on estimates on the probability of SGD leaving favorable neighborhoods in normal and tangential direction separately. In order to derive estimates on the rate of convergence, these steps are performed in a quantitative way in the proofs of this work. An intriguing observation is that the mathematical analysis of the rate of convergence relies on the use of mini-batches in order to control the loss of iterates in non-attracted regions.

In Sections 3 and 4 we provide an analysis of the deterministic gradient descent algorithm in continuous and discrete time in order to highlight the relevance of the assumptions in simplified settings. We emphasize again that, while the deterministic algorithms converge quickly, the computational costs of computing ∇f typically make the implementation of such algorithms infeasible. This is particularly the case when f takes the form (1.33) below for a measure μ that is the empirical measure of a large training set. An advantage of the stochastic algorithm is that, provided $M\in\mathbb{N}$ is not too large, the mini-batch gradient can be computed efficiently in the case of (1.34) below. The disadvantage is that, inside an attracting set, the algebraic convergence of SGD in expectation is much slower than the exponential convergence of its deterministic counterpart.

1.1. **Structure of the work.** The paper is organized as follows. We will use the local smoothness of $\mathcal{M}\cap U$, the local smoothness of the objective function f, and the maximal nondegeneracy of the Hessian to identify a basin of attraction for SGD. In Section 2, we present the geometric preliminaries that are used to identify this set. In particular, in Proposition 2.2 below we recall the existence of projections in a local neighborhoods of $\mathcal{M}\cap U$, in Proposition 2.5 below we recall the existence of local tubular neighborhoods about $\mathcal{M}\cap U$, in Lemma 2.6 below we prove a useful decomposition of ∇f into components normal and tangential to $\mathcal{M}\cap U$, and in Lemma 2.7 below we prove a contraction estimate that will be used to obtain a convergence rate for the gradient descent algorithms in discrete time.

In Section 3, for objective functions $f\colon\mathbb{R}^d\to\mathbb{R}$ satisfying the conditions of Theorem 1.1, we analyze the converge of the deterministic gradient descent algorithm in continuous time $\theta_t\in\mathbb{R}^d$,

$t \in [0, \infty)$, satisfying for every $t \in (0, \infty)$ that

$$\frac{d}{dt}\theta_t = -\nabla f(\theta_t). \tag{1.31}$$

We prove in Proposition 3.1 below that the local smoothness of $\mathcal{M} \cap U$, the local smoothness of f, and the nondegeneracy of the Hessian imply the existence of a neighborhood $V \subseteq \mathbb{R}^d$ such that for every $\theta_0 \in V$ the solution θ_t, $t \in [0, \infty)$, converges exponentially fast to $\mathcal{M} \cap U$. However, since in general neither f nor ∇f are practically computable, and since continuous gradient descent cannot be implemented, the purpose of this section is to explain in a simplified setting the role of the assumptions and the geometric arguments from Section 2.

In Section 4, for objective functions $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfying the conditions of Theorem 1.1, we analyze the converge of the deterministic gradient descent algorithm in discrete time $\theta_n \in \mathbb{R}^d$, $n \in \mathbb{N}_0$, satisfying for $\rho \in (0, 1)$, $r \in (0, \infty)$, for every $n \in \mathbb{N}$ that

$$\theta_n = \theta_{n-1} - \frac{r}{n^\rho}\nabla f(\theta_{n-1}). \tag{1.32}$$

We prove in Proposition 4.1 below that there exists a neighborhood $V \subseteq \mathbb{R}^d$ such that for every $\theta_0 \in V$ the solution θ_n, $n \in \mathbb{N}_0$, converges exponentially quickly to $\mathcal{M} \cap U$. However, while discrete gradient descent yields an implementable algorithm, the computational costs of f and ∇f in general make it practically infeasible. The purpose of this section is instead to explain how the geometric preliminaries of Section 2, and in particular Lemma 2.6 and Lemma 2.7, are applied in a simplified discrete setting.

In Section 5, we analyze the convergence of SGD to the manifold of local minima $\mathcal{M} \cap U$. In Proposition 5.2 below, we prove the convergence of (1.4) to $\mathcal{M} \cap U$ in directions normal to the manifold. Precisely, we identify a basin of attraction $V \subseteq \mathbb{R}^d$ such that, on the event that SGD remains in V, SGD converges to $\mathcal{M} \cap U$ in expectation with an algebraic rate. It remains to estimate the probability that SGD remains in the basin of attraction V.

The first step is contained in Proposition 5.3 below, which estimates the maximal excursion of SGD in expectation. Then, in Proposition 5.6 below, we estimate the probability that SGD remains in a basin of attraction V by separating this event into the event that SGD leaves V in a direction normal to $\mathcal{M} \cap U$ and the event that SGD leaves V in a direction tangential to $\mathcal{M} \cap U$. Proposition 5.2 is used to estimate the first of these events, and Proposition 5.3 is used to estimate the second. In Theorem 5.7, we combine Proposition 5.2 and Proposition 5.6 to estimate the probability that SGD converges to within distance $\varepsilon \in (0, 1]$ of $\mathcal{M} \cap U$.

In Corollary 5.8 below, we estimate the probability that $K \in \mathbb{N}$ independent copies of SGD fail to converge to within distance $\varepsilon \in (0, 1]$ of $\mathcal{M} \cap U$. In Theorem 5.11 below we prove Theorem 1.1, which relies on Lemma 5.10 below and estimates for the mini-batch approximation of the objective function. Finally, in Corollary 5.12 below, we estimate the computational efficiency of the algorithm introduced in Theorem 1.1.

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. These estimates apply, in particular, to the case when the objective function has a unique minimum. The reason for the improved estimate of Theorem 6.4 below and the improved computational efficiency of Corollary 6.5 below is that, in the compact case, SGD cannot escape a basin of attraction in directions tangential to the manifold. It is therefore sufficient to take a smaller mini-batch approximation of the gradient.

In Section 7, we prove that assumptions of Theorem 1.1 are satisfied by simple loss functions arising in machine learning applications. In particular, we show that the assumptions are satisfied by objective functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying that

$$f(\theta) = \int_S |u_\theta(x) - \varphi(x)|^p \, \mu(\,\mathrm{d}x), \tag{1.33}$$

where $\theta \in \mathbb{R}^d$, $p \in [1,\infty)$, φ a measurable function on a measurable space $(S,\mathcal{S})$, and $(u_\theta\colon S \to \mathbb{R})_{\theta\in\mathbb{R}^d}$ is a jointly-measurable artificial neural network. In this case, the function $F\colon \mathbb{R}^d \times S \to \mathbb{R}$ satisfies for every $(\theta,x) \in \mathbb{R}^d \times S$ that

$$F(\theta,x) = |u_\theta(x) - \varphi(x)|^p\,, \tag{1.34}$$

and, for a probability space $(\Omega,\mathcal{F},\mathbb{P})$, the sequence of random variables $X_{n,m}\colon \Omega \to S$, $n,m \in \mathbb{N}$, are i.i.d. with distribution μ. For the objective functions considered in Section 7.1 and Section 7.2 below, the global minima are non-unique and build locally smooth, non-compact manifolds of $\mathbb{R}^d$ on which Hessian of the objective function is maximally nondegenerate.

1.2. **Literature.** The stochastic gradient descent (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. We will therefore restrict to works that seem most relevant to the current results and refer to the following works and the references therein for further details: See, for example, [2, 3, 4, 6, 7, 9, 10, 14, 23, 28, 34, 39, 40, 42, 43, 44, 49, 50, 51, 54, 56] and the references mentioned therein for numerical simulations and proofs of convergence rates for SGD type optimization algorithms, [5, 8, 47] and the references mentioned therein for overview articles on SGD type optimization algorithms, and [11, 12, 18, 19, 21, 22, 26, 27, 48] and the references mentioned therein for applications involving neural networks and 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 [8, 56]. In the case of a strongly convex objective function these can be improved to $O(1/n)$, see [20, 37, 38].

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 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 these do not exclude convergence to saddle points or local minima. In the case of non-convex loss function rather complete results are known concerning the minimization of the gradient of the loss function. For example, the convergence of the gradient to zero with rates was shown by Lei, Hu, Li and Tang in [29] assuming a Hölder-regularity condition on the gradient of the loss function. This generalizes previous work [17] by Ghadimi, Lan, Zhang which required a second moment boundedness condition, which in turn generalized the previous works by Ghadimi, Lan [16] and by Reddi, Hefny, Sra, Poczos, Smola [45]. We note that while convergence to the global minimum with rates was obtained in [17] for the convex case, no results on the convergence of the value of the loss function have been shown in the non-convex case.

The convergence of the stochastic gradient descent method has been analysed in the literature under several additional assumptions replacing (strong) convexity, such as the error bounds condition by Luo and Tseng [33], essential strong convexity [31], weak strong convexity [36], the restricted secant inequality [55], and the quadratic growth condition Anitescu [1]. In these works, linear convergence rates are shown. In the notable contribution [25] Karimi, Nutini, and Schmidt have shown that all of these conditions imply the Polyak-Lojasiewicz (PL) inequality, introduced by Lojasiewicz in [32] and Polyak in [41], under which linear convergence of SGD is proven in [25], thus generalizing these previous works. Recently, further progress was made by Lei, Hu, Li and Tang in [29] where a boundedness assumption on the gradient of the objective function, required in [25], 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 each local minimum is a global minimum. Therefore, it is not implied by the assumptions made in the current work.

2. Geometric preliminaries

In this section, for an objective function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying the conditions of Theorem 1.1, 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 projection to $\mathcal{M} \cap U$ which is, however, only well-defined in local neighborhoods of the local manifold.

Proposition 2.1. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, \dots, d-1\}$, let $\mathcal{M} \cap U \subseteq \mathbb{R}^d$ be a non-empty, $\mathfrak{d}$-dimensional, C^1-submanifold of $\mathbb{R}^d$. Then for every $x_0 \in \mathcal{M} \cap U$ there exists an open subset $V_*(x_0) \subseteq \mathbb{R}^d$ containing x_0 satisfying that:*

(a) projections exist: for every $x \in V_*(x_0)$, there exists a unique point $x_* \in \mathcal{M} \cap U$ satisfying that

$$|x - x_*| = \mathrm{d}(x, \mathcal{M} \cap U) = \inf_{\vartheta \in \mathcal{M} \cap U} |x - \vartheta| . \tag{2.1}$$

(b) the projection map is locally C^1-smooth: the map $x \in V_*(x_0) \mapsto x_* \in \mathcal{M} \cap U$ is C^1-smooth.

Proof of Proposition 2.1. The proof is an immediate consequence of Foote [15, Lemma] and the C^1-regularity of $\mathcal{M} \cap U$. □

The following proposition proves that for every $x \in \mathcal{M} \cap U$ the tangent space $T_x(\mathcal{M} \cap U)$ and normal space $T_x(\mathcal{M} \cap U)^\perp$ to $\mathcal{M} \cap U$ at x are characterized respectively by the null space of Hessian of f and the space on which the Hessian of f is positive definite. To simplify the notation, we will write $\nabla^2 f$ for the Hessian of the objective function.

Proposition 2.2. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, 2, \dots, d-1\}$, let $U \subseteq \mathbb{R}^d$ be an open set, let $f \colon U \to \mathbb{R}$ be a three times continuously differentiable function, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \left\{ \theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)] \right\}, \tag{2.2}$$

assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, and assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$. Then for every $x \in \mathcal{M} \cap U$, there exists a $(d - \mathfrak{d})$-dimensional subspace $P_x \subseteq \mathbb{R}^d$ and a $\mathfrak{d}$-dimensional subspace $N_x \subseteq \mathbb{R}^d$ such that

$$\nabla^2 f(x)(P_x) \subseteq P_x \;\; \textit{with} \;\; \nabla^2 f(x)|_{P_x} \;\; \textit{strictly positive definite on} \;\; P_x, \tag{2.3}$$

that

$$\nabla^2 f(x)(N_x) \subseteq N_x \;\; \textit{with} \;\; \nabla^2 f(x)|_{N_x} = 0, \tag{2.4}$$

and that

$$N_x = T_x(\mathcal{M} \cap U) \;\; \textit{and} \;\; P_x = T_x(\mathcal{M} \cap U)^\perp. \tag{2.5}$$

Proof of Proposition 2.2. Let $x \in \mathcal{M} \cap U$. Since $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, the symmetry of the Hessian implies that there exist subspaces $N_x, P_x \subseteq \mathbb{R}^d$ satisfying $\mathbb{R}^d = P_x \oplus N_x$, that $\dim(P_x) = d - \mathfrak{d}$, that

$$\nabla^2 f(x)(P_x) \subseteq P_x \;\; \text{with} \;\; \nabla^2 f(x)|_{P_x} \;\; \text{strictly positive definite on} \;\; P_x, \tag{2.6}$$

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

$$\nabla^2 f(x)(N_x) \subseteq N_x \;\; \text{with} \;\; \nabla^2 f(x)|_{N_x} = 0. \tag{2.7}$$

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

$$\left. \frac{d}{dt} \nabla f(\gamma(t)) \right|_{t=0} = \nabla^2 f(x) \cdot \dot{\gamma}(0) = 0. \tag{2.8}$$

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 $\mathbb{R}^d = T_x(\mathcal{M} \cap U) \oplus T_x(\mathcal{M} \cap U)^\perp$, it holds that $P_x = T_x(\mathcal{M} \cap U)^\perp$, which completes the proof. □

In the following lemma, for a point $x \in \mathbb{R}^d$ satisfying that the projection $x_* \in \mathcal{M} \cap U$ is well-defined, we prove that the difference $x - x_* \in \mathbb{R}^d$ lies in the space normal to $\mathcal{M} \cap U$ at x_*. This fact will be used to obtain a rate of convergence for the discrete gradient descent algorithms.

Lemma 2.3. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, 2, \dots, d-1\}$, let $U \subseteq \mathbb{R}^d$ be an open set, let $f\colon U \to \mathbb{R}$ be a three times continuously differentiable function, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \bigl\{\theta \in \mathbb{R}^d\colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\bigr\}, \tag{2.9}$$

assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, and assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$. Then for every $x_0 \in \mathcal{M} \cap U$, for every open neighborhood $V_*(x_0) \subseteq \mathbb{R}^d$ containing x_0 that satisfies the conclusion of Proposition 2.1, it holds for every $x \in V_*(x_0)$ that

$$x - x_* \in T_{x_*}(\mathcal{M} \cap U)^{\perp}. \tag{2.10}$$

Proof of Lemma 2.3. Let $x_0 \in \mathcal{M} \cap U$ and let $V_*(x_0) \subseteq \mathbb{R}^d$ be an open neighborhood containing x_0 that satisfies the conclusion of Proposition 2.1. Let $x \in V_*(x_0)$. If $x \in \mathcal{M} \cap U$, the claim is immediate since then $x - x_* = 0$. 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 satisfying $\gamma(0) = x_*$. It holds that

$$\frac{d}{dt} \|x - \gamma(t)\|^2 \bigg|_{t=0} = -2\dot{\gamma}(0) \cdot (x - x_*) = 0. \tag{2.11}$$

Therefore, since the curve γ was arbitrary, it holds that $x - x_* \in T_{x_*}(\mathcal{M} \cap U)^{\perp}$, which completes the proof. □

In the following lemma, we derive a formula for the derivative of the distance function to the manifold in a neighborhood of $\mathcal{M} \cap U$. The regularity of the distance function and the formula for its differential will be used to prove the convergence of the deterministic gradient descent algorithm in continuous time.

Lemma 2.4. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, 2, \dots, d-1\}$, let $U \subseteq \mathbb{R}^d$ be an open set, let $f\colon U \to \mathbb{R}$ be a three times continuously differentiable function, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \bigl\{\theta \in \mathbb{R}^d\colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\bigr\}, \tag{2.12}$$

assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, and assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$. Then for every $x_0 \in \mathcal{M} \cap U$, for every neighborhood $V_*(x_0) \subseteq \mathbb{R}^d$ satisfying the conclusion of Proposition 2.1, it holds for every $x \in V_*(x_0) \setminus \mathcal{M} \cap U$ that

$$\nabla \mathrm{d}(x, \mathcal{M} \cap U) = \frac{x - x_*}{\|x - x_*\|}. \tag{2.13}$$

Proof of Lemma 2.4. Let $x_0 \in \mathcal{M} \cap U$ and let $V_*(x_0) \subseteq \mathbb{R}^d$ be an open neighborhood containing x_0 that satisfies the conclusion of Proposition 2.1. It follows from Proposition 2.1 that

$$x \in V_*(x_0) \mapsto \|x - x_*\|^2 = \mathrm{d}(x, \mathcal{M} \cap U)^2 \text{ is } \mathrm{C}^1\text{-smooth}. \tag{2.14}$$

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

$$\frac{\partial}{\partial x_i} \mathrm{d}(x, \mathcal{M} \cap U)^2 = \frac{\partial}{\partial x_i} \|x - x_*\|^2 = 2(x - x_*) \cdot e_i - 2(x - x_*) \cdot \frac{\partial}{\partial x_i} x_*. \tag{2.15}$$

Since $\frac{\partial}{\partial x_i} x_* \in N_{x_*}$ and since $x - x_* \in P_{x_*}$ it follows from Lemma 2.3 that

$$(x - x_*) \cdot \frac{\partial}{\partial x_i} x_* = 0. \tag{2.16}$$

Since for every $x \in V_*(x_0) \setminus \mathcal{M} \cap U$ it holds that

$$\nabla \mathrm{d}(x, \mathcal{M} \cap U)^2 = 2\,\mathrm{d}(x, \mathcal{M} \cap U) \nabla \mathrm{d}(x, \mathcal{M} \cap U) = 2(x - x_*), \tag{2.17}$$

it holds for every $x \in V_*(x_0) \setminus \mathcal{M} \cap U$ that

$$\nabla \operatorname{d}(x, \mathcal{M} \cap U) = \frac{x - x_*}{\|x - x_*\|}, \tag{2.18}$$

which completes the proof. □

We will now quantify what are local tubular neighborhoods of the local manifold $\mathcal{M} \cap U$. For every $x_0 \in \mathcal{M} \cap U$, $R, \delta \in (0, \infty)$ let $V_{R,\delta}(x_0) \subseteq \mathbb{R}^d$ satisfy that

$$V_{R,\delta}(x_0) = \{x + v\colon x \in \overline{B}_R(x_0) \cap \mathcal{M} \cap U \text{ and } v \in T_x(\mathcal{M} \cap U)^\perp \text{ with } |v| < \delta\}. \tag{2.19}$$

The useful feature of these neighborhoods it 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 $\mathcal{M} \cap U$. The following technical proposition will be used to prove Proposition 4.1 below and Lemma 5.5 below.

Proposition 2.5. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, \dots, d-1\}$, let $\mathcal{M} \cap U \subseteq \mathbb{R}^d$ be a non-empty, $\mathfrak{d}$-dimensional, C^1-submanifold of $\mathbb{R}^d$, for every $x_0 \in \mathcal{M} \cap U$, $R, \delta \in (0, \infty)$ let $V_{R,\delta}(x_0) \subseteq \mathbb{R}^d$ satisfy that*

$$V_{R,\delta}(x_0) = \{x + v\colon x \in \overline{B}_R(x_0) \cap \mathcal{M} \cap U \text{ and } v \in T_x(\mathcal{M} \cap U)^\perp \text{ with } |v| < \delta\}. \tag{2.20}$$

Then for every $x_0 \in \mathcal{M} \cap U$, for every open neighborhood $V_*(x_0) \subseteq \mathbb{R}^d$ containing x_0 that satisfies the conclusion of Proposition 2.1, there exist $R_0, \delta_0 \in (0, \infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ it holds that $\overline{V}_{R,\delta}(x_0) \subseteq V_*(x_0)$, that

$$V_{R,\delta}(x_0) = \{x \in \mathbb{R}^d\colon \operatorname{d}(x, \mathcal{M} \cap U) = \operatorname{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta\}, \tag{2.21}$$

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

$$(x + v)_* = x. \tag{2.22}$$

Proof of Proposition 2.5. Let $x_0 \in \mathcal{M} \cap U$. For every $R, \delta \in (0, \infty)$ let $\tilde{V}_{R,\delta}(x_0) \subseteq \mathbb{R}^d$ satisfy that

$$\tilde{V}_{R,\delta}(x_0) = \{x \in \mathbb{R}^d\colon \operatorname{d}(x, \mathcal{M} \cap U) = \operatorname{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta\}. \tag{2.23}$$

Let $V_*(x_0) \subseteq \mathbb{R}^d$ be an open neighborhood containing x_0 that satisfies the conclusion of Proposition 2.1. Since $U, V_*(x_0) \subseteq \mathbb{R}^d$ are open, there exist $R_0, \delta_0 \in (0, \infty)$ such that for every $R \in (0, R_0]$ it holds that

$$\overline{B}_R(x_0) \cap \mathcal{M} \subseteq \mathcal{M} \cap U, \tag{2.24}$$

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

$$V_{R,\delta}(x_0) \subseteq V_*(x_0) \text{ and } \tilde{V}_{R,\delta}(x_0) \subseteq V_*(x_0). \tag{2.25}$$

Following [15, Lemma], the normal bundle $T(\mathcal{M} \cap U)^\perp \subseteq \mathbb{R}^{2d}$ satisfies that

$$T(\mathcal{M} \cap U)^\perp := \left\{(x, v) \in \mathbb{R}^d \times \mathbb{R}^d\colon x \in \mathcal{M} \cap U \text{ and } v \in T_x(\mathcal{M} \cap U)^\perp\right\}. \tag{2.26}$$

Since $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional C^1-submanifold, it follows that $T(\mathcal{M} \cap U)^\perp \subseteq \mathbb{R}^{2d}$ is a d-dimensional C^1-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

$$D_{(x,0)}\Psi\colon T_{(x,0)}\big(T(\mathcal{M} \cap U)^\perp\big) \to T_x\mathbb{R}^d \text{ is nonsingular.} \tag{2.27}$$

It follows from the inverse function theorem that there exists $\delta_1 \in (0, (\delta_0 \wedge {R_0}/{4}))$ such that for every $R \in (0, {R_0}/{2}]$, $\delta \in (0, \delta_1]$ it holds that

$$\Psi\colon \{(x, v) \in (TM)^\perp : x \in \overline{B}_{R+2\delta_1}(x_0) \text{ and } |v| < \delta\} \to V_{R+2\delta_1,\delta}(x_0) \text{ is injective.} \tag{2.28}$$

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 $\mathrm{d}(x, \mathcal{M} \cap U) = \mathrm{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U)$ and since the choice of $R_0 \in (0, \infty)$ implies that

$$\overline{B}_R(x_0) \cap \mathcal{M} \cap U = \overline{B}_R(x_0) \cap \mathcal{M} \text{ is a closed subset of } \mathbb{R}^d, \tag{2.29}$$

it holds that $x_* \in \overline{B}_R(x_0) \cap \mathcal{M} \cap U$. Since $\mathrm{d}(x, \mathcal{M} \cap U) = \mathrm{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) = \|x - x_*\| < \delta$ and since it holds that

$$x = x_* + \|x - x_*\| \frac{x - x_*}{\|x - x_*\|}, \tag{2.30}$$

for $\frac{x-x_*}{\|x-x_*\|} \in T_x (\mathcal{M} \cap U)^\perp$ by Lemma 2.3, it holds 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 $\mathrm{d}(x, \mathcal{M} \cap U) = \mathrm{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta$. 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 $\tilde{v} \in T_{\tilde{x}} (\mathcal{M} \cap U)^\perp$ with $|\tilde{v}| < \delta$ satisfying that $x = \tilde{x} + \tilde{v}$. We will prove that $x_* = \tilde{x}$.

By contradiction, suppose that $x_* \neq \tilde{x}$. This implies that

$$\|x - x_*\| < \|x - \tilde{x}\| = \|\tilde{v}\| < \delta. \tag{2.31}$$

It follows from the triangle inequality that

$$\|x_* - \tilde{x}\| \leq \|x_* - x\| + \|x - \tilde{x}\| < 2\delta \leq 2\delta_1, \tag{2.32}$$

which proves that

$$x = \tilde{x} + \tilde{v} = x_* + (x - x_*), \tag{2.33}$$

for $x - x_* \in T_{x_*} (\mathcal{M} \cap U)^\perp$ by Lemma 2.3 with $\|x - x_*\| < \delta$. Since $\tilde{x} \in \overline{B}_R(x_0) \cap \mathcal{M} \cap U$, it follows from (2.32) that $x_* \in \overline{B}_{R+2\delta_1}(x_0) \cap \mathcal{M} \cap U$.

Since $R \in (0, {}^{R_0}/{}_2]$ and since $\delta \in (0, \delta_1]$, equation (2.33) contradicts (2.28), which states that Ψ is injective on the set

$$\{(x, v) \in (TM)^\perp : x \in B_{R+2\delta_1}(x_0) \text{ and } |v| < \delta\}. \tag{2.34}$$

We conclude that $x_* = \tilde{x}$, which implies that

$$\mathrm{d}(x, \mathcal{M} \cap U) = \mathrm{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) = \|x - x_*\| = \|\tilde{v}\| < \delta. \tag{2.35}$$

Therefore, it holds 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 (2.32) and (2.33). □

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

$$\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d} = \operatorname{codim}(\mathcal{M} \cap U). \tag{2.36}$$

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 $\mathcal{M} \cap U$. We will use the normal component to obtain a rate of convergence for the gradient descent algorithms. The contribution of the tangential component will create errors that will need to be controlled.

Lemma 2.6. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, 2, \ldots, d-1\}$, let $U \subseteq \mathbb{R}^d$ be an open set, let $f \colon U \to \mathbb{R}$ be a three times continuously differentiable function, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \left\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\right\}, \tag{2.37}$$

assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, and assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$. Then for every $x_0 \in \mathcal{M} \cap U$ there exist $R_0, \delta_0, c \in (0, \infty)$

and an open neighborhood $V_*(x_0) \subseteq U$ containing x_0 satisfying Proposition 2.1 *such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ it holds that*

$$\overline{V}_{R,\delta}(x_0) \subseteq V_*(x_0), \tag{2.38}$$

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

$$\nabla f(x) = \nabla^2 f(x_*) \cdot (x - x_*) + \varepsilon_x. \tag{2.39}$$

Proof of Lemma 2.6. Let $x_0 \in \mathcal{M} \cap U$ and $R > 0$. Since $U \subseteq \mathbb{R}^d$ is an open set, there exists an open neighborhood $V_*(x_0) \subseteq U$ containing x_0 that satisfies the conclusion of Proposition 2.1. Since $V_*(x_0)$ is open, fix $R_0, \delta_0 \in (0, \infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ it holds that

$$\overline{V}_{R,\delta}(x_0) \subseteq V_*(x_0). \tag{2.40}$$

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

$$\|f\|_{\mathrm{C}^3(V_{R,\delta}(x_0))} = \sup_{0 \le k \le 3} \left\| \nabla^k f \right\|_{L^\infty(V_{R,\delta_0}(x_0);\mathbb{R}^{(d^k)})} \le c. \tag{2.41}$$

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

$$\begin{aligned} \nabla f(x) &= \int_0^1 \nabla^2 f(x_* + s(x - x_*)) \cdot (x - x_*)\, \mathrm{d}s \\ &= \nabla^2 f(x_*) \cdot (x - x_*) + \int_0^1 \big(\nabla^2 f(x_* + s(x - x_*)) - \nabla^2 f(x_*)\big) \cdot (x - x_*)\, \mathrm{d}s. \end{aligned} \tag{2.42}$$

It follows from (2.41), the local regularity of f, and the definition of the projection that there exists $c \in (0, \infty)$ satisfying that

$$\begin{aligned} \left| \int_0^1 \big(\nabla^2 f(x_* + s(x - x_*)) - \nabla^2 f(x_*)\big) \cdot (x - x_*)\, \mathrm{d}s \right| &\le c\,\mathrm{d}(x, \mathcal{M} \cap U)^2 \int_0^1 s\, \mathrm{d}s \\ &\le c\,\mathrm{d}(x, \mathcal{M} \cap U)^2. \end{aligned} \tag{2.43}$$

After defining

$$\varepsilon_x := \int_0^1 \big(\nabla^2 f(x_* + s(x - x_*)) - \nabla^2 f(x_*)\big) \cdot (x - x_*)\, \mathrm{d}s, \tag{2.44}$$

equation (2.42) and estimate (2.43) complete the proof. □

The following lemma will play an important role in the analysis of the deterministic and stochastic gradient descent algorithms in discrete time. In the context of Lemma 2.6, for every $x \in \mathbb{R}^d$ with $x_* \in \mathcal{M} \cap U$ well-defined, the following lemma quantifies the convergence of gradient descent to $\mathcal{M} \cap U$ in the direction of the approximately normal component of the gradient $\nabla^2 f(x_*) \cdot (x - x_*)$.

Lemma 2.7. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, 2, \ldots, d-1\}$, let $U \subseteq \mathbb{R}^d$ be an open set, let $f \colon U \to \mathbb{R}$ be a three times continuously differentiable function, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}, \tag{2.45}$$

assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, and assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$. Then for every $x_0 \in \mathcal{M} \cap U$ there exist $R_0, \delta_0, \mathfrak{r}, \in (0, \infty)$, $\lambda \in (0, \infty)$ satisfying that

$$\lambda \le \max_{x \in \mathcal{M} \cap U \cap \overline{B}_R(x_0)} \left\| \nabla^2 f(x_*) \right\|, \tag{2.46}$$

and an open neighborhood $V_*(x_0) \subseteq \mathbb{R}^d$ satisfying Proposition 2.1 *such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $x \in V_{R,\delta}(x_0)$ it holds that*

$$\overline{V}_{R,\delta}(x_0) \subseteq V_*(x_0), \tag{2.47}$$

that

$$\begin{aligned} \mathrm{d}\left(x - r\nabla^2 f(x_*) \cdot (x - x_*), \mathcal{M} \cap U\right) &\leq \left|(x - x_*) - r\nabla^2 f(x_*) \cdot (x - x_*)\right| \\ &\leq (1 - \lambda r)\, \mathrm{d}(x, \mathcal{M} \cap U), \end{aligned} \tag{2.48}$$

and that

$$\left(\nabla^2 f(x_*) \cdot (x - x_*)\right) \cdot (x - x_*) \geq \lambda\, \mathrm{d}(x, \mathcal{M} \cap U)^2. \tag{2.49}$$

Proof of Lemma 2.7. Let $x_0 \in \mathcal{M} \cap U$. Since $U \subseteq \mathbb{R}^d$ is an open subset, there exists an open neighborhood $V_*(x_0) \subseteq U$ of x_0 satisfying Proposition 2.1. Fix $R_0, \delta_0 \in (0, \infty)$ such that every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ satisfies that

$$\overline{V}_{R,\delta}(x_0) \subseteq V_*(x_0). \tag{2.50}$$

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

$$\|f\|_{\mathrm{C}^3(V_{R,\delta}(x_0))} \leq c. \tag{2.51}$$

Let $x \in V_{R,\delta}(x_0)$. For the first claim, using (2.51), fix $\mathfrak{r} \in (0, \infty)$ satisfying that

$$\mathfrak{r} \left(\max_{x \in V_{R_0,\delta_0}(x_0)} \left\|\nabla^2 f(x_*)\right\| \right) \leq 1. \tag{2.52}$$

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

$$\mathrm{d}\left(x - r\nabla^2 f(x_*) \cdot (x - x_*), \mathcal{M} \cap U\right) \leq \left|(x - x_*) - r\nabla^2 f(x_*) \cdot (x - x_*)\right|. \tag{2.53}$$

Since the nondegeneracy assumption states that

$$\operatorname{rank}((\operatorname{Hess} f)(x_*)) = d - \mathfrak{d} = \operatorname{codim}(\mathcal{M} \cap U), \tag{2.54}$$

Lemma 2.3 below and (2.51) prove that there exists for $\lambda \in (0, \infty)$ satisfying that

$$\lambda \leq \max_{x \in \mathcal{M} \cap U \cap \overline{B}_R(x_0)} \left\|\nabla^2 f(x_*)\right\|, \tag{2.55}$$

for which we have that

$$\left|(x - x_*) - r\nabla^2 f(x_*) \cdot (x - x_*)\right| \leq (1 - r\lambda)\, |x - x_*| = (1 - r\lambda)\, \mathrm{d}(x, \mathcal{M} \cap U), \tag{2.56}$$

where the choice of $\mathfrak{r}$ and (2.55) guarantee that $(1 - r\lambda) \geq 0$. In combination, estimates (2.53), (2.55), and (2.56) 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 2.3, and (2.51) prove that there exists $\lambda \in (0, \infty)$ satisfying (2.55) such that

$$\left(\nabla^2 f(x_*) \cdot (x - x_*)\right) \cdot (x - x_*) \geq \lambda\, |x - x_*|^2 = \lambda\, \mathrm{d}(x, \mathcal{M} \cap U)^2, \tag{2.57}$$

which completes the proof. □

3. Continuous deterministic gradient descent

In this section, for an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfying the conditions of Theorem 1.1, 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 $\theta_t \in \mathbb{R}^d$, $t \in [0, \infty)$, satisfying for every $t \in (0, \infty)$ that

$$\frac{d}{dt}\theta_t = -\nabla f(\theta_t). \tag{3.1}$$

We will prove that the solution of (3.1) converges to the local manifold of minima $\mathcal{M} \cap U$, provided the initial condition is chosen in a sufficiently small neighborhood of $\mathcal{M} \cap U$. The proof can be outlined as follows. Given any $x_0 \in \mathcal{M} \cap U$, we first fix an open neighborhood x_0 satisfying the conclusions of Lemma 2.6 and Lemma 2.7. Then, for initial data θ_0 in this neighborhood, we quantify the convergence of the solution (3.1) to $\mathcal{M} \cap U$ in directions normal to the manifold, using the decomposition of ∇f from Lemma 2.6. Finally, after fixing smaller neighborhood about x_0, we prove that the tangential components of the gradient of ∇f do not take the trajectory from the basin of attraction.

Proposition 3.1. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, 2, \ldots, d-1\}$, let $U \subseteq \mathbb{R}^d$ be an open set, let $f\colon U \to \mathbb{R}$ be a three times continuously differentiable function, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \bigl\{\theta \in \mathbb{R}^d\colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\bigr\}, \tag{3.2}$$

assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, and assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$. Then for every $x_0 \in \mathcal{M} \cap U$ there exist $R_0, \delta_0, \lambda \in (0, \infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $\theta_0 \in V_{R/2,\delta}(x_0)$, for $\theta_t \in \mathbb{R}^d$, $t \in (0, \infty)$, satisfying that

$$\frac{d}{dt}\theta_t = -\nabla f(\theta_t), \tag{3.3}$$

it holds for every $t \in [0, \infty)$ that

$$\mathrm{d}(\theta_t, \mathcal{M} \cap U) \le \exp(-\lambda t)\, \mathrm{d}(\theta_0, \mathcal{M} \cap U). \tag{3.4}$$

Proof of Proposition 3.1. Let $x_0 \in \mathcal{M} \cap U$. Since $U \subseteq \mathbb{R}^d$ is an open set, fix a neighborhood $V_*(x_0) \subseteq U$ of x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0, \delta_0 \in (0, \infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ the set $V_{R,\delta}(x_0)$ satisfies the conclusion of Proposition 2.5 for $V_*(x_0)$. This is to say that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ we have $\overline{V}_{R,\delta}(x_0) \subseteq V_*(x_0)$ and that

$$\begin{aligned} V_{R,\delta}(x_0) &= \{x + v \in \mathbb{R}^d\colon x \in \overline{B}_R(x_0) \cap \mathcal{M} \cap U \text{ and } v \in T_x\left(\mathcal{M} \cap U\right)^{\perp} \text{ with } |v| < \delta\} \\ &= \{x \in \mathbb{R}^d\colon\ \mathrm{d}(x, \mathcal{M} \cap U) = \mathrm{d}(x, \overline{B}_R(x_0) \cap \mathcal{M} \cap U) < \delta\}. \end{aligned} \tag{3.5}$$

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

$$\|f\|_{\mathrm{C}^3(V_{R_0,\delta_0}(x_0))} \le c. \tag{3.6}$$

Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, and let $P\colon V_{R,\delta}(x_0) \to \mathcal{M} \cap U$ satisfy for every $y \in V_{R,\delta}(x_0)$ that $P(y) = y_*$. Let $\theta_0 \in V_{R/2,\delta}(x_0)$, let $\theta_t \in \mathbb{R}^d$, $t \in (0, \infty)$, satisfy that

$$\frac{d}{dt}\theta_t = -\nabla f(\theta_t), \tag{3.7}$$

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

$$\tau := \inf\{\ t \ge 0 \mid \theta_t \notin V_{R,\delta}(x_0)\ \}. \tag{3.8}$$

Lemma 2.4 and the chain rule prove that

$$\begin{cases} \frac{d}{dt}\,\mathrm{d}(\theta_t, \mathcal{M} \cap U) = -\nabla f(\theta_t) \cdot \nabla\,\mathrm{d}(\theta_t, \mathcal{M} \cap U) = -\nabla f(\theta_t) \cdot \frac{\theta_t - \theta_{t,*}}{\|\theta_t - \theta_{t,*}\|} & \text{in } (0,\tau), \\ \frac{d}{dt}\theta_{t,*} = -DP(\theta_t) \cdot \nabla f(\theta_t) & \text{in } (0,\tau), \end{cases} \tag{3.9}$$

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 2.6 and Lemma 2.7 that there exist $\lambda, c_1 \in (0,\infty)$ satisfying that

$$\nabla f(\theta_t) \cdot \frac{\theta_t - \theta_{t,*}}{\|\theta_t - \theta_{t,*}\|} \geq \lambda\,\mathrm{d}(\theta_t, \mathcal{M} \cap U) - c_1\,\mathrm{d}(\theta_t, \mathcal{M} \cap U)^2. \tag{3.10}$$

Proposition 2.1, (3.6), and $\nabla f|_{\mathcal{M} \cap U} = 0$ prove that there exists $c_2 \in (0,\infty)$ satisfying that

$$|DP(\theta_t) \cdot \nabla f(\theta_t)| \leq c_2\,\mathrm{d}(\theta_t, \mathcal{M} \cap U). \tag{3.11}$$

Returning to (3.9), it follows from (3.10) and (3.11) that

$$\begin{cases} \frac{d}{dt}\,\mathrm{d}(\theta_t, \mathcal{M} \cap U) \leq -\lambda\,\mathrm{d}(\theta_t, \mathcal{M} \cap U) + c_1\,\mathrm{d}(\theta_t, \mathcal{M} \cap U)^2 & \text{in } (0,\tau), \\ \left|\frac{d}{dt}\theta_{t,*}\right| \leq c_2\,\mathrm{d}(\theta_t, \mathcal{M} \cap U) & \text{in } (0,\tau). \end{cases} \tag{3.12}$$

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

$$c_1\delta_1 \leq \lambda/2. \tag{3.13}$$

Let $\delta \in (0,\delta_1]$. For every $t \in (0,\tau)$ it follows from (3.12) and (3.13) that

$$\frac{d}{dt}\,\mathrm{d}(\theta_t, \mathcal{M} \cap U) \leq -\frac{\lambda}{2}\,\mathrm{d}(\theta_t, \mathcal{M} \cap U). \tag{3.14}$$

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

$$\mathrm{d}(\theta_t, \mathcal{M} \cap U) \leq \mathrm{d}(\theta_0, \mathcal{M} \cap U)\exp(-\lambda t/2) \leq \delta_1 \exp(-\lambda t/2). \tag{3.15}$$

For every $t \in [0,\tau)$, it follows from (3.12) and (3.15) that

$$\max_{0 \leq t \leq \tau} |\theta_{t,*} - \theta_{0,*}| \leq c_2 \int_0^\tau \delta_1 \exp\left(-\frac{\lambda t}{2}\right) \mathrm{d}t = \frac{2c_2\delta_1}{\lambda}\left(1 - \exp\left(-\frac{\lambda\tau}{2}\right)\right) \leq \frac{2c_2\delta_1}{\lambda}. \tag{3.16}$$

Fix $\delta_2 \in (0,\delta_1]$ satisfying

$$\frac{2c_2\delta_2}{\lambda} < \frac{R}{2}. \tag{3.17}$$

Let $\delta \in (0,\delta_2]$. In combination (3.15), (3.16), $\theta_0 \in V_{R/2,\delta}(x_0)$, and the triangle inequality prove that $\theta_t \in V_{R,\delta}(x_0)$ for every $t \in (0,\infty)$. That is, we have $\tau = \infty$. Since $\theta_0 \in V_{R/2,\delta}(x_0)$ was arbitrary, this completes the proof. □

4. Discrete deterministic gradient descent

In this section, for an objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfying the conditions of Theorem 1.1, we will analyze the convergence of the following deterministic gradient descent algorithm $\theta_n \in \mathbb{R}^d$, $n \in \mathbb{N}_0$, in discrete time satisfying for a learning rate $\rho \in (0,1)$ and $r \in (0,\infty)$ that

$$\theta_n = \theta_{n-1} - \frac{r}{n^\rho}\nabla f(\theta_{n-1}). \tag{4.1}$$

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 the learning rate $r \in (0,\infty)$ sufficiently small. Since, if the learning rate 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.

In the proof, we first identify a basin of attraction using Proposition 2.1 and Proposition 2.5. In the second step, we prove that the solution (4.1) converges along the normal directions to the manifold of local minima provided the solution remains in the basin of attraction. For this, we use the normal component of ∇f from Lemma 2.6 and the quantification of the convergence from Lemma 2.7. Finally, after fixing a perhaps smaller basin of attraction, we prove that the tangential component of the gradient from Lemma 2.6 does not cause the solution (4.1) to leave the basin of attraction.

Proposition 4.1. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{1, 2, \ldots, d-1\}$, $\rho \in (0,1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $f \colon U \to \mathbb{R}$ be a three times continuously differentiable function, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \left\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\right\}, \tag{4.2}$$

assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, and assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$. 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 $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $\theta_0 \in V_{R/2,\delta}(x_0)$, for $\theta_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, satisfying that

$$\theta_n = \theta_{n-1} - \frac{r}{n^\rho} \nabla f(\theta_{n-1}), \tag{4.3}$$

it holds for every $n \in \mathbb{N}_0$ that

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \le \exp(-c n^{1-\rho})\, \mathrm{d}(x_0, \mathcal{M} \cap U). \tag{4.4}$$

Proof of Proposition 4.1. Let $x_0 \in \mathcal{M} \cap U$ and $\rho \in (0,1)$. Since $U \subseteq \mathbb{R}^d$ is open, fix a neighborhood $V_*(x_0)$ of x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0, \delta_0 \in (0, \infty)$ satisfying the conclusion of Proposition 2.5 for this set $V_*(x_0)$. In particular, the regularity of f and the compactness of $\overline{V}_{R_0,\delta_0}(x_0)$ prove that there exists $c \in (0, \infty)$ satisfying that

$$\|f\|_{\mathrm{C}^3(V_{R_0,\delta_0}(x_0))} \le c. \tag{4.5}$$

Fix $\mathfrak{r} \in (0, \infty)$ satisfying the conclusion of Lemma 2.7 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

$$\theta_n = \theta_{n-1} - \frac{r}{n^\rho} \nabla f(\theta_{n-1}), \tag{4.6}$$

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

$$\tau = \inf\{\, n \in \mathbb{N} \mid \theta_n \notin V_{R,\delta}(x_0) \,\}. \tag{4.7}$$

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

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \le |\theta_n - \theta_{n-1,*}| = \left|\theta_{n-1} - \theta_{n-1,*} - \frac{r}{n^\rho} \nabla f(\theta_{n-1})\right|. \tag{4.8}$$

Lemma 2.6 proves that there exists $c \in (0, \infty)$ such that for every $n \in \{1, \ldots, \tau\}$ there exists $\varepsilon_n \in \mathbb{R}^d$ satisfying that

$$|\varepsilon_n| \le c\, \mathrm{d}(\theta_{n-1}, \mathcal{M} \cap U)^2, \tag{4.9}$$

such that

$$\nabla f(\theta_{n-1}) = \nabla^2 f(\theta_{n-1,*}) \cdot (x - x_*) + \varepsilon_n. \tag{4.10}$$

The triangle inequality, (4.8), (4.9), and (4.10) prove that there exists $c_1 \in (0, \infty)$ such that for every $n \in \{1, \ldots, \tau\}$ it holds that

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \le \left|\theta_{n-1} - \theta_{n-1,*} - \frac{r}{n^\rho} \nabla^2 f(\theta_{n-1,*}) \cdot (\theta_{n-1} - \theta_{n-1,*})\right| + \frac{c_1 r}{n^\rho}\, \mathrm{d}(\theta_{n-1}, \mathcal{M} \cap U)^2. \tag{4.11}$$

Finally, the choice of $\mathfrak{r} \in (0, \infty)$, Lemma 2.7, and (4.11) prove that there exists $\lambda \in (0, \infty)$ such that for every $n \in \{1, \ldots, \tau\}$ it holds that

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \leq \left(1 - \frac{r\lambda}{n^\rho}\right) \mathrm{d}(\theta_{n-1}, \mathcal{M} \cap U) + \frac{c_1 r}{n^\rho} \mathrm{d}(\theta_{n-1}, \mathcal{M} \cap U)^2, \tag{4.12}$$

where the choice of $\mathfrak{r} \in (0, \infty)$ guarantees that $(1 - r\lambda) \geq 0$.

Fix $\delta_1 \in (0, \delta_0]$ satisfying that

$$c_1 \delta_1 \leq \frac{\lambda}{2}. \tag{4.13}$$

Let $\delta \in (0, \delta_1]$. It follows from (4.12) and (4.13) that for every $n \in \{1, \ldots, \tau\}$ it holds that

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \leq \left(1 - \frac{r\lambda}{2n^\rho}\right) \mathrm{d}(\theta_{n-1}, \mathcal{M} \cap U). \tag{4.14}$$

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

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \leq \prod_{k=1}^{n} \left(1 - \frac{r\lambda}{2k^\rho}\right) \mathrm{d}(\theta_0, \mathcal{M} \cap U). \tag{4.15}$$

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

$$\log\left(\prod_{k=1}^{n} \left(1 - \frac{r\lambda}{2k^\rho}\right)\right) = \sum_{k=1}^{n} \log\left(1 - \frac{r\lambda}{2k^\rho}\right) \leq -c \sum_{k=1}^{n} \frac{r\lambda}{2k^\rho} \leq -c\frac{r\lambda}{2} n^{1-\rho}, \tag{4.16}$$

it follows from (4.15) that there exists $c_2 \in (0, \infty)$ satisfying for every $n \in \{1, \ldots, \tau\}$ that

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \leq \exp\left(-c_2 n^{1-\rho}\right) \mathrm{d}(\theta_0, \mathcal{M} \cap U). \tag{4.17}$$

It remains only to show that, provided $\delta \in (0, \delta_1]$ is chosen sufficiently small, we have that $\tau = \infty$.

It follows from (4.5), (4.17), and $\nabla f|_{\mathcal{M} \cap U} = 0$ that there exists $c \in (0, \infty)$ satisfying that

$$|\theta_n - \theta_{n-1}| = \frac{r}{n^\rho} |\nabla f(\theta_{n-1})| \leq \frac{c}{n^\rho} \mathrm{d}(\theta_{n-1}, \mathcal{M} \cap U) \leq c n^{-\rho} \exp\left(-c_2 n^{1-\rho}\right) \mathrm{d}(\theta_0, \mathcal{M} \cap U). \tag{4.18}$$

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

$$|\theta_n - \theta_0| \leq c\, \mathrm{d}(\theta_0, \mathcal{M} \cap U) \sum_{k=1}^{\infty} c k^{-\rho} \exp\left(-c_2 k^{1-\rho}\right) = c_3\, \mathrm{d}(\theta_0, \mathcal{M} \cap U) < \infty. \tag{4.19}$$

Fix $\delta_2 \in (0, \delta_1]$ satisfying that

$$c_3 \delta_2 < \frac{R}{2} - 2\delta_2. \tag{4.20}$$

Let $\delta \in (0, \delta_2]$. The choice of $\delta_2 \in (0, \delta_1]$, (4.19), and the triangle inequality prove for every $n \in \{1, \ldots, \tau\}$ that

$$|\theta_n - x_0| \leq |\theta_n - \theta_0| + |\theta_0 - x_0| < c_3 \delta_2 + \frac{R}{2} + \delta_2 < R - \delta_2. \tag{4.21}$$

In combination (4.17) and (4.21) prove for every $n \in \{1, \ldots, \tau\}$ that

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) < \delta_2 \ \text{ and } \ |\theta_n - x_0| \leq R - \delta_2. \tag{4.22}$$

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

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) = \mathrm{d}(\theta_n, \overline{B}_R(x_0) \cap \mathcal{M} \cap U). \tag{4.23}$$

It follows from Proposition 2.5, 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. □

Remark 4.2. The conclusion of Proposition 4.1 can be extended to the case of $\rho = 1$ using the same techniques. In this case, in the setting of Proposition 4.1, there exists $R_0, \delta_0, \mathfrak{r}, c \in (0,\infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $\theta_0 \in V_{R/2,\delta}(x_0)$, for $\theta_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, satisfying that

$$\theta_n = \theta_{n-1} - \frac{r}{n} \nabla f(\theta_{n-1}), \tag{4.24}$$

it holds for every $n \in \mathbb{N}_0$ that

$$\mathrm{d}(\theta_n, \mathcal{M} \cap U) \leq \exp(-c \log(n))\, \mathrm{d}(x_0, \mathcal{M} \cap U). \tag{4.25}$$

The logarithm appears in estimate (4.16) in the case $\rho = 1$. The remainder of the proof is then the same, where the only additional observation is that the analogue of (4.19) is finite in the case $\rho = 1$ as well.

5. Stochastic gradient descent

In this section, in the setting of Theorem 1.1, for a learning rate $\rho \in (2/3, 1)$, for $r \in (0, \infty)$, $M \in \mathbb{N}$, for a bounded open subset $A \subseteq \mathbb{R}^d$, for a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for a measurable space $(S, \mathcal{S})$, for a jointly measurable function $F\colon S \times \Omega \to \mathbb{R}$, for $X_{n,m}\colon \Omega \to \mathbb{R}^d$, $n, m \in \mathbb{N}$, i.i.d. random variables, we will analyze the convergence of the mini-batch stochastic gradient descent algorithm $\Theta_n\colon \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}_0$, satisfying that Θ_0 is continuous uniformly distributed on A and for every $n \in \mathbb{N}$ that

$$\Theta_n = \Theta_{n-1} - \frac{r}{M n^\rho} \sum_{m=1}^{M} \nabla_\theta F(\Theta_{n-1}, X_{n,m}). \tag{5.1}$$

The role of the mini-batch size $M \in \mathbb{N}$ is to reduce the variance of the random gradient

$$\frac{1}{M} \sum_{m=1}^{M} \nabla_\theta F(\Theta_{n-1}, X_{n,m}). \tag{5.2}$$

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

Lemma 5.1. *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 $G = (G(\theta, x))_{(\theta,x)\in\mathbb{R}^{d_1}\times S}\colon \mathbb{R}^{d_1} \times S \to \mathbb{R}^{d_2}$ be a jointly measurable function, let $X_m\colon \Omega \to S$, $m \in \mathbb{N}$, be i.i.d. random variables, and assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|G(\theta, X_1)|^2\big] < \infty$. Then for every non-empty compact set $\mathfrak{C} \subseteq U$ there exists $c \in (0,\infty)$ satisfying for every $M \in \mathbb{N}$ that*

$$\sup_{\theta\in\mathfrak{C}} \Big(\mathbb{E}\Big[\Big|\frac{1}{M} \sum_{m=1}^{M} G(\theta, X_m) - \mathbb{E}\big[G(\theta, X_1)\big]\Big|^2\Big]\Big) \leq \frac{c}{M}. \tag{5.3}$$

Proof of Lemma 5.1. Let $\mathfrak{C} \subseteq U$ be compact. For every $\theta \in \mathfrak{C}$, $M \in \mathbb{N}$ it holds that

$$\begin{aligned} &\mathbb{E}\Big[\Big|\frac{1}{M} \sum_{m=1}^{M} G(\theta, X_m) - \mathbb{E}\big[G(\theta, X_1)\big]\Big|^2\Big] \\ &= \frac{1}{M^2} \sum_{i,j=1}^{M} \mathbb{E}\Big[\big(G(\theta, X_i) - \mathbb{E}\big[G(\theta, X_1)\big]\big) \cdot \big(G(\theta, X_j) - \mathbb{E}\big[G(\theta, X_1)\big]\Big]. \end{aligned} \tag{5.4}$$

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 $c \in (0, \infty)$ satisfying for every $M \in \mathbb{N}$ that

(5.5)
$$\begin{aligned}\sup_{\theta \in \mathfrak{C}} \Big(\mathbb{E}\Big[\Big| \frac{1}{M} \sum_{m=1}^{M} G(\theta, X_m) - \mathbb{E}\big[G(\theta, X_1)\big] \Big|^2 \Big] \Big) &= \sup_{\theta \in \mathfrak{C}} \Big(\frac{1}{M^2} \sum_{m=1}^{M} \mathbb{E}\Big[\Big| G(\theta, X_m) - \mathbb{E}\big[G(\theta, X_1)\big] \Big|^2 \Big] \Big) \\ &= \frac{1}{M} \sup_{\theta \in \mathfrak{C}} \Big(\mathbb{E}\Big[\Big| G(\theta, X_1) - \mathbb{E}\big[G(\theta, X_1)\big] \Big|^2 \Big] \Big) \\ &\leq \frac{c}{M},\end{aligned}$$

which completes the proof. □

In the following proposition, much like the first step of the proofs of Proposition 3.1 and Proposition 4.1, we establish the convergence of (5.1) in directions normal to the local manifold of minima. We first identify a basin of attraction for (5.1) using Proposition 2.1 and Proposition 2.5 and prove, using the gradient decomposition of Lemma 2.6 and the quantification of convergence from Lemma 2.7, that on the event that SGD does not escape this basin of attraction SGD converges to the manifold of minima in expectation.

We emphasize that the events A_n, $n \in \mathbb{N}_0$, defined in Proposition 5.2 below depend upon the quantifiers $n, M \in \mathbb{N}$, $r, R, \delta \in (0, \infty)$, $\theta \in \mathbb{R}^d$, and $x_0 \in \mathcal{M} \cap U$. However, in order to simplify the presentation, we suppress this dependence in the notation.

Proposition 5.2. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, $\rho \in (2/3, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta, x) \in \mathbb{R}^d \times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}, \tag{5.6}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta \in \mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $M \in \mathbb{N}$, $r \in (0, \infty)$, $\theta \in \mathbb{R}^d$ let $\Theta_{0,\theta}^{M,r} \in \mathbb{R}^d \colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\theta_{0,\theta}^{M,r}(\omega) = \theta$, for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$, $\theta \in \mathbb{R}^d$ let $\Theta_{n,\theta}^{M,r} \colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta_{n,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho M} \left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,\theta}^{M,r}, X_{n,m}) \right], \tag{5.7}$$

for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0, \infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M} \cap U$ let $A_n \subseteq \Omega$ be the event satisfying that

$$A_n = \Big\{\omega \in \Omega \colon \Theta_{m,\theta}^{M,r}(\omega) \in V_{R,\delta}(x_0) \ \forall\, m \in \{0, \ldots, n\}\Big\}, \tag{5.8}$$

and for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0, \infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M} \cap U$ let $\mathbf{1}_n : \Omega \to \{0, 1\}$ denote the indicator function of A_n. Then for every $x_0 \in \mathcal{M} \cap U$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0, \infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$, $\theta \in V_{R,\delta}(x_0)$ it holds that

$$\mathbb{E}\left[\Big(\mathrm{d}(\Theta_{n,\theta}^{M,r}, \mathcal{M} \cap U) \wedge 1 \Big)^2 \mathbf{1}_{n-1} \right]^{\frac{1}{2}} \leq c n^{-\frac{\rho}{2}}. \tag{5.9}$$

Proof of Proposition 5.2. Let $x_0 \in \mathcal{M} \cap U$. Since $U \subseteq \mathbb{R}^d$ is open, fix a neighborhood $V_*(x_0) \subseteq U$ of x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0, \delta_0 \in (0, \infty)$ satisfying the conclusion of Proposition 2.5 for this set $V_*(x_0)$. Finally, fix $\mathfrak{r} \in (0, \infty)$ satisfying the conclusion of Lemma 2.7.

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 $\nabla_\theta F^{M,n} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be the functions satisfying for every $(\theta, \omega) \in \mathbb{R}^d \times \Omega$ that

$$\nabla_\theta F^{M,n}(\theta) = \nabla_\theta F^{M,n}(\theta, \omega) = \frac{1}{M} \sum_{m=1}^{M} (\nabla_\theta F)(\theta, X_{n,m}(\omega)). \tag{5.10}$$

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

$$\Theta^{M,r}_{n,\theta} = \Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho} \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta}). \tag{5.11}$$

We will analyze the solution $\Theta^{M,r}_{n,\theta}$ of (5.11) on the event A_{n-1}. We observe that

$$\Theta^{M,r}_{n,\theta} = \Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho} \nabla f(\Theta^{M,r}_{n-1,\theta}) + \frac{r}{n^\rho} \left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta}) \right). \tag{5.12}$$

Since the event A_{n-1} implies that $\Theta^{M,r}_{n-1,\theta} \in V_{R,\delta}(x_0) \subseteq V_*(x_0)$, for the projection $\Theta^{M,r}_{n-1,\theta,*}$ of $\Theta^{M,r}_{n-1,\theta}$, it holds by definition of the distance to $\mathcal{M} \cap U$ that

$$\begin{aligned}
&\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U)^2 \\
&\leq \left| \Theta^{M,r}_{n,\theta} - \Theta^{M,r}_{n-1,\theta,*} \right|^2 \\
&\leq \left| \Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*} - \frac{r}{n^\rho} \nabla f(\Theta^{M,r}_{n-1,\theta}) \right|^2 \\
&\quad + 2 \left(\Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*} - \frac{r}{n^\rho} \nabla f(\Theta^{M,r}_{n-1,\theta}) \right) \cdot \frac{r}{n^\rho} \left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta}) \right) \\
&\quad + \left| \frac{r}{n^\rho} \left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta}) \right) \right|^2.
\end{aligned} \tag{5.13}$$

The three terms on the righthand side of (5.13) will be treated separately.

For the first term on the righthand side of (5.13), the choice of $\mathfrak{r} \in (0, \infty)$, Lemma 2.6, and Lemma 2.7 prove, following identically the proof leading from (4.8) to (4.12), that there exist $\lambda, c \in (0, \infty)$ such that

$$\left| \Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*} - \frac{r}{n^\rho} \nabla f(\Theta^{M,r}_{n-1,\theta}) \right| \leq \left(1 - \frac{r\lambda}{n^\rho} \right) \mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U) + c \frac{r}{n^\rho} \mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2. \tag{5.14}$$

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

$$\begin{aligned}
\left| \Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*} - \frac{r}{n^\rho} \nabla f(\Theta^{M,r}_{n-1,\theta}) \right|^2 &\leq \left(1 - \frac{r\lambda}{n^\rho} \right)^2 \mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2 \\
&\quad + c \left(1 - \frac{r\lambda}{n^\rho} \right) \frac{r}{n^\rho} \mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^3 \\
&\quad + c \frac{r^2}{n^{2\rho}} \mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^4.
\end{aligned} \tag{5.15}$$

The remaining two terms of (5.13) and the righthand side of (5.15) will be handled after taking the expectation on the event $A_{n-1} \subseteq \Omega$ satisfying that

$$A_{n-1} = \left\{ \omega \in \Omega \colon \Theta^{M,r}_{m,\theta} \in V_{R,\delta}(x_0) \; \forall \, m \in \{0, \ldots, n-1\} \right\}. \tag{5.16}$$

For the indicator function $\mathbf{1}_{n-1}$ of A_{n-1}, after returning to (5.13), it follows from (5.15) that there exists $c \in (0,\infty)$ satisfying that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] \\
&\quad \leq \left(1 - \frac{r\lambda}{n^\rho}\right)^2 \mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] \\
&\qquad + c\left(1 - \frac{r\lambda}{n^\rho}\right)\frac{r}{n^\rho}\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^3 \mathbf{1}_{n-1}\right] + c\frac{r^2}{n^{2\rho}}\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^4 \mathbf{1}_{n-1}\right] \\
&\qquad + 2\mathbb{E}\left[\left(\Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta^{M,r}_{n-1,\theta}) - \Theta^{M,r}_{n-1,\theta,*}\right) \cdot \frac{r}{n^\rho}\left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right)\mathbf{1}_{n-1}\right] \\
&\qquad + \mathbb{E}\left[\left|\frac{r}{n^\rho}\left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right)\right|^2 \mathbf{1}_{n-1}\right].
\end{aligned} \tag{5.17}$$

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

$$\mathcal{F}_m = \sigma\big(\{X_{1,k}\}_{k=1}^M, \ldots, \{X_{m,k}\}_{k=1}^M\big). \tag{5.18}$$

For the penultimate term of (5.17), since $\mathbf{1}_{n-1}$ is $\mathcal{F}_{n-1}$-measurable, properties of the conditional expectation imply that

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta^{M,r}_{n-1,\theta}) - \Theta^{M,r}_{n-1,\theta,*}\Big) \cdot \frac{r}{n^\rho}\Big(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\Big)\mathbf{1}_{n-1}\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\Big(\Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta^{M,r}_{n-1,\theta}) - \Theta^{M,r}_{n-1,\theta,*}\Big) \cdot \frac{r}{n^\rho}\Big(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\Big)\mathbf{1}_{n-1}|\mathcal{F}_{n-1}\Big]\Big] \\
&= \mathbb{E}\Big[\Big(\Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho}\nabla f(\Theta^{M,r}_{n-1,\theta}) - \Theta^{M,r}_{n-1,\theta,*}\Big)\mathbf{1}_{n-1} \cdot \mathbb{E}\Big[\frac{r}{n^\rho}\Big(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\Big)|\mathcal{F}_{n-1}\Big]\Big] \\
&= 0,
\end{aligned} \tag{5.19}$$

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

$$\mathbb{E}\left[\frac{r}{n^\rho}\left(\nabla f(x) - \nabla_\theta F^{M,n}(x)\right)|\mathcal{F}_{n-1}\right] = \frac{r}{Nn^\rho}\sum_{m=1}^M \mathbb{E}\left[\nabla f(x) - \nabla_\theta F(x, X_{n,m})\right] = 0. \tag{5.20}$$

The final term of (5.17) is handled using Lemma 5.1. Since $\overline{V}_{R,\delta}(x_0)$ is compact, the independence of the $X_{m,k}$, $m,k \in \mathbb{N}$, and Lemma 5.1 prove that there exists $c \in (0,\infty)$ such that

$$\mathbb{E}\left[\left|\frac{r}{n^\rho}\left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right)\mathbf{1}_{n-1}\right|^2\right] \leq \frac{cr^2}{Mn^{2\rho}}. \tag{5.21}$$

Returning to (5.17), it follows from (5.19) and (5.21) that there exists $c_1 \in (0,\infty)$ such that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] \leq \\
&\left(1 - \frac{r\lambda}{n^\rho}\right)^2 \mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] + c_1\left(1 - \frac{r\lambda}{n^\rho}\right)\frac{r}{n^\rho}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^3 \mathbf{1}_{n-1}\right] \\
&\quad + c_1\frac{r^2}{n^{2\rho}}\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^4 \mathbf{1}_{n-1}\right] + c_1\frac{r^2}{Mn^{2\rho}}.
\end{aligned} \tag{5.22}$$

Fix $\delta_1 \in (0,\delta_0]$ satisfying that

$$\delta_1 \leq \frac{\lambda}{2c_1} \text{ and } \delta_1^2 \leq \frac{\lambda}{2c_1 r}. \tag{5.23}$$

Let $\delta \in (0, \delta_1]$. We claim that inequality (5.22) implies that there exists some $c \in (0, \infty)$ satisfying for every $n \in \mathbb{N}$ that

$$\mathbb{E}\left[\left(\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U) \wedge 1\right)^2 \mathbf{1}_{n-1}\right]^{\frac{1}{2}} \leq c n^{-\frac{\rho}{2}}. \tag{5.24}$$

The proof of (5.24) 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

$$\left(n^\rho - (n-1)^\rho - r\lambda + \frac{r^2\lambda^2}{n^\rho}\right) \leq \left(\rho(n-1)^{\rho-1} - r\lambda + \frac{r^2\lambda^2}{n^\rho}\right) \leq -\frac{r\lambda}{2}, \tag{5.25}$$

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 $n_0 \geq 1$ satisfying (5.25) and define

$$\overline{c} := \max\left\{(n_0 - 1)^\rho, \frac{2c_1 r}{M\lambda}\right\}. \tag{5.26}$$

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

$$\mathbb{E}\left[\left(\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U) \wedge 1\right)^2 \mathbf{1}_{n-1}\right] \leq \overline{c} n^{-\rho}. \tag{5.27}$$

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

$$\mathbb{E}\left[\left(\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U) \wedge 1\right)^2 \mathbf{1}_{n-2}\right] \leq \overline{c}(n-1)^{-\rho}. \tag{5.28}$$

Since the event A_{n-1} implies that

$$\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U) \leq \delta \leq 1, \tag{5.29}$$

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

$$\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^m \mathbf{1}_{n-1}\right] \leq \delta^{m-2}\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-2}\right] \leq \delta^{m-2}\overline{c}(n-1)^{-\rho}. \tag{5.30}$$

Returning to (5.22), it holds that

$$\begin{aligned}\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] \leq& \overline{c}\left(1 - \frac{r\lambda}{n^\rho}\right)^2 (n-1)^{-\rho} + \overline{c}c_1\delta\left(1 - \frac{r\lambda}{n^\rho}\right)\frac{r}{n^\rho}(n-1)^{-\rho} \\ &+ \overline{c}c_1\delta^2\frac{r^2}{n^{2\rho}}(n-1)^{-\rho} + c_1\frac{r^2}{Mn^{2\rho}}.\end{aligned} \tag{5.31}$$

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

$$\begin{aligned}&\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] \leq \overline{c}n^{-\rho} \\ &\quad + n^{-\rho}\left(\overline{c}(n-1)^{-\rho}\left(n^\rho - (n-1)^\rho - 2r\lambda + \frac{r^2\lambda^2}{n^\rho} + c_1\delta r\left(1 - \frac{r\lambda}{n^\rho}\right) + c_1\delta^2\frac{r^2}{n^\rho}\right) + c_1\frac{r^2}{Mn^\rho}\right).\end{aligned} \tag{5.32}$$

Since $\delta \in (0, \delta_1]$, it follows from (5.32) that

$$\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] \leq \overline{c}n^{-\rho} + n^{-\rho}\left(\overline{c}(n-1)^{-\rho}\left(n^\rho - (n-1)^\rho - r\lambda + \frac{r^2\lambda^2}{n^\rho}\right) + c_1\frac{r^2}{Mn^\rho}\right). \tag{5.33}$$

Since $n \geq n_0$, the choice $\overline{c} \geq \frac{2c_1 r}{M\lambda}$, (5.25), and (5.33) prove that

$$\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U)^2 \mathbf{1}_{n-1}\right] \leq \overline{c}n^{-\rho} + n^{-\rho}\left(-\frac{r\lambda}{2}\overline{c}(n-1)^{-\rho} + c_1\frac{r^2}{Mn^\rho}\right) \leq \overline{c}n^{-\rho}. \tag{5.34}$$

Therefore, we have that

$$\mathbb{E}\left[\left(\mathrm{d}(\Theta^{M,r}_{n,\theta},\mathcal{M}\cap U)\wedge 1\right)^2 \mathbf{1}_{n-1}\right] \leq \left[\mathrm{d}(\Theta^{M,r}_{n,\theta},\mathcal{M}\cap U)^2 \mathbf{1}_{n-1}\right] \leq \overline{c} n^{-\rho}, \tag{5.35}$$

which completes the induction step. Since the base case is (5.27), this completes the proof. □

Proposition 5.2 proves the convergence of SGD to $\mathcal{M}\cap U$ on the event that SGD remains in a basin of attraction. It remains necessary to prove that, provided the mini-batch size is chosen to be sufficiently large, SGD remains in the basin of attraction for large times. We prove the first step toward this goal in the proposition below, which estimates the maximal excursion of SGD on the event that the dynamics do not leave a basin of attraction.

Proposition 5.3. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0,1,\dots,d-1\}$, $\rho \in (2/3,1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $(S,\mathcal{S})$ be a measurable space, let $F = (F(\theta,x))_{(\theta,x)\in\mathbb{R}^d\times S}\colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m}\colon \Omega \to S$, $n,m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta,X_{1,1})|^2\big] < \infty$, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta,X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)]\big\}, \tag{5.36}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta,x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U\colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|F(\theta,X_{1,1})|^2 + |(\nabla_\theta F)(\theta,X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M}\cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume for every $\theta \in \mathcal{M}\cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $M \in \mathbb{N}$, $r \in (0,\infty)$, $\theta \in \mathbb{R}^d$ let $\Theta^{M,r}_{0,\theta} \in \mathbb{R}^d \colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\theta^{M,r}_{0,\theta}(\omega) = \theta$, for every $n, M \in \mathbb{N}$, $r \in (0,\infty)$, $\theta \in \mathbb{R}^d$ let $\Theta^{M,r}_{n,\theta}\colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta^{M,r}_{n,\theta} = \Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta^{M,r}_{n-1,\theta}, X_{n,m})\right], \tag{5.37}$$

for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0,\infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M}\cap U$ let $A_n \subseteq \Omega$ be the event satisfying that

$$A_n = \Big\{\omega \in \Omega \colon \Theta^{M,r}_{m,\theta}(\omega) \in V_{R,\delta}(x_0)\ \forall\, m \in \{0,\dots,n\}\Big\}, \tag{5.38}$$

and for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0,\infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M}\cap U$ let $\mathbf{1}_n \colon \Omega \to \{0,1\}$ denote the indicator function of A_n. Then for every $x_0 \in \mathcal{M}\cap U$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0,\infty)$ such that for every $R \in (0,R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M \in \mathbb{N}$, $\theta \in V_{R/2,\delta}(x_0)$ it holds that

$$\mathbb{E}\left[\max_{1\leq k\leq n}\left|\Theta^{M,r}_{k,\theta} - \Theta^{M,r}_{0,\theta}\right|\mathbf{1}_{k-1}\right] \leq \sum_{k=1}^{n} \mathbb{E}\left[\left|\Theta^{M,r}_{k,\theta} - \Theta^{M,r}_{k-1}\right|^2 \mathbf{1}_{k-1}\right]^{\frac{1}{2}} \leq cr\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right). \tag{5.39}$$

Proof of Proposition 5.3. Let $x_0 \in \mathcal{M}\cap U$. Since $U \subseteq \mathbb{R}^d$ is open, fix a neighborhood $V_*(x_0) \subseteq U$ containing x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0, \delta_0 \in (0,\infty)$ satisfying the conclusion of Proposition 2.5 for this set $V_*(x_0)$. We observe that the regularity of f and the compactness of $\overline{V}_{R_0,\delta_0}(x_0)$ imply that

$$\|f\|_{\mathrm{C}^3(V_{R_0,\delta_0}(x_0))} \leq c. \tag{5.40}$$

Finally, fix $\mathfrak{r} \in (0,\infty)$ satisfying the conclusion of Lemma 2.7.

Let $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$. As in Proposition 5.2, let $\nabla_\theta F^{M,n} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be the functions satisfying for every $(\theta, \omega) \in \mathbb{R}^d \times \Omega$ that

$$\nabla_\theta F^{M,n}(\theta) = \nabla_\theta F^{M,n}(\theta, \omega) = \frac{1}{M} \sum_{m=1}^{M} (\nabla_\theta F)(\theta, X_{n,m}(\omega)). \tag{5.41}$$

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

$$\Theta_{n,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho} \nabla_\theta F^{M,n}(\Theta_{n-1,\theta}^{M,r}). \tag{5.42}$$

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

$$\mathbb{E}\left[\left|\Theta_{n,\theta}^{M,r} - \Theta_{n-1,\theta}^{M,r}\right|^2 \mathbf{1}_{n-1}\right]^{\frac{1}{2}} \leq c\left(\frac{r}{n^{\frac{3}{2}\rho}} + \frac{r}{n^\rho M^{\frac{1}{2}}}\right), \tag{5.43}$$

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

$$\Theta_{n,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho} \nabla f(\Theta_{n-1}^{M,r}) + \frac{r}{n^\rho}\left(\nabla f(\Theta_{n-1}^{M,r}) - \nabla F^{M,n}(\Theta_{n-1}^{M,r})\right). \tag{5.44}$$

Lemma 2.6 proves that there exists $c_1 \in (0, \infty)$ and $\varepsilon_n \colon A_{n-1} \to \mathbb{R}^d$ satisfying that

$$|\varepsilon_n| \leq c_1 \, \mathrm{d}(\Theta_{n-1}^{M,r}, \mathcal{M} \cap U)^2, \tag{5.45}$$

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

$$\nabla f(\Theta_{n-1}^{M,r}) = \nabla^2 f(\Theta_{n-1,\theta,*}^{M,r}) \cdot (\Theta_{n-1}^{M,r} - \Theta_{n-1,\theta,*}^{M,r}) + \varepsilon_n. \tag{5.46}$$

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

$$\Theta_{n,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho} \nabla^2 f(\Theta_{n-1,\theta,*}^{M,r}) \cdot (\Theta_{n-1,\theta}^{M,r} - \Theta_{n-1,\theta,*}^{M,r}) - \frac{r}{n^\rho}\varepsilon_n + \frac{r}{n^\rho}(\nabla f(\Theta_{n-1,\theta}^{M,r}) - \nabla F^{M,n}(\Theta_{n-1,\theta}^{M,r})). \tag{5.47}$$

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

$$\tilde{\Theta}_{n-1,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho} \nabla^2 f(\Theta_{n-1,\theta,*}^{M,r}) \cdot \left(\Theta_{n-1,\theta}^{M,r} - \Theta_{n-1,\theta,*}^{M,r}\right). \tag{5.48}$$

After taking the norm-squared of (5.47), on the event A_{n-1} it holds that

$$\begin{aligned} \left|\Theta_{n,\theta}^{M,r} - \tilde{\Theta}_{n-1,\theta}^{M,r}\right|^2 =& \frac{r^2}{n^{2\rho}} |\varepsilon_n|^2 - 2\frac{r^2}{n^{2\rho}} \varepsilon_n \cdot \left(\nabla f(\Theta_{n-1,\theta}^{M,r}) - \nabla F^{M,n}(\Theta_{n-1,\theta}^{M,r})\right) \\ &+ \frac{r^2}{n^{2\rho}} \left|\nabla f(\Theta_{n-1,\theta}^{M,r}) - \nabla F^{M,n}(\Theta_{n-1,\theta}^{M,r})\right|^2. \end{aligned} \tag{5.49}$$

We will estimate (5.49) by taking the expectation on the event A_{n-1}.

The first term on the righthand side of (5.49) is handled using Proposition 5.2 and (5.45). For the second term, from (5.18) we recall the sigma algebras $\mathcal{F}_m \subseteq \mathcal{F}$, $m \in \mathbb{N}$, satisfying that

$$\mathcal{F}_m = \sigma\big(\{X_{1,k}\}_{k=1}^M, \ldots, \{X_{m,k}\}_{k=1}^M\big). \tag{5.50}$$

Since $\varepsilon_n \colon A_{n-1} \to \mathbb{R}^d$ is $\mathcal{F}_{n-1}$-measurable, it follows identically to (5.19) and (5.20) that

$$\mathbb{E}\left[\varepsilon_n \cdot \left(\nabla f(\Theta_{n-1,\theta}^{M,r}) - \nabla F^{M,n}(\Theta_{n-1,\theta}^{M,r})\right) \mathbf{1}_{n-1}\right] = 0. \tag{5.51}$$

For the final term on the righthand side of (5.49), 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 5.1 prove that there exists $c \in (0, \infty)$ satisfying that

$$\mathbb{E}\left[\left|\nabla f(\Theta_{n-1,\theta}^{M,r}) - \nabla F^{M,n}(\Theta_{n-1,\theta}^{M,r})\right|^2 \mathbf{1}_{n-1}\right] \leq \frac{c}{M}. \tag{5.52}$$

In combination, Proposition 5.2 and estimates (5.45), (5.49), (5.51), and (5.52) prove that there exists $c \in (0, \infty)$ satisfying that

$$\begin{aligned}\mathbb{E}\left[\left|\Theta^{M,r}_{n,\theta} - \tilde{\Theta}^{M,r}_{n-1,\theta}\right|^2 \mathbf{1}_{n-1}\right] &\leq c\left(\frac{r^2\delta^2}{n^{2\rho}}\mathbb{E}\left[\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2\right] + \frac{r^2}{n^{2\rho}M}\right) \\ &\leq c\left(\frac{r^2\delta^2}{n^{3\rho}} + \frac{r^2}{n^{2\rho}M}\right).\end{aligned} \tag{5.53}$$

It follows from the definition of $\tilde{\Theta}^{M,r}_{n-1,\theta}$, (5.40), and the definition of the projection that, on the event A_{n-1} there exists $c \in (0, \infty)$ satisfying that

$$\left|\tilde{\Theta}^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta}\right|^2 = \frac{r^2}{n^{2\rho}}\left|\nabla^2 f(\Theta^{M,r}_{n-1,\theta,*}) \cdot (\Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*})\right|^2 \leq c\frac{r^2}{n^{2\rho}}\,\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2. \tag{5.54}$$

Proposition 5.2 proves that there exists $c \in (0, \infty)$ such that

$$\mathbb{E}\left[\left|\tilde{\Theta}^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta}\right|^2 \mathbf{1}_{n-1}\right] \leq \frac{cr^2}{n^{3\rho}}. \tag{5.55}$$

It follows from the triangle inequality, (5.53), and (5.55) that there exists $c_1 \in (0, \infty)$ satisfying that

$$\begin{aligned}\mathbb{E}\left[\left|\Theta^{M,r}_{n,\theta} - \Theta^{M,r}_{n-1,\theta}\right|^2 \mathbf{1}_{n-1}\right]^{\frac{1}{2}} \leq& \mathbb{E}\left[\left|\Theta^{M,r}_{n,\theta} - \tilde{\Theta}^{M,r}_{n-1,\theta}\right|^2 \mathbf{1}_{n-1}\right]^{\frac{1}{2}} + \mathbb{E}\left[\left|\tilde{\Theta}^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta}\right|^2 \mathbf{1}_{n-1}\right]^{\frac{1}{2}} \\ \leq& c_1\left(\frac{r}{n^{\frac{3}{2}\rho}} + \frac{r}{n^{\rho}M^{\frac{1}{2}}}\right),\end{aligned} \tag{5.56}$$

which completes the proof of (5.43).

Since for every $r \leq s \in \mathbb{N}_0$ we have $\mathbf{1}_s \leq \mathbf{1}_r$, it follows from (5.56), the triangle inequality, and Hölder's inequality that there exists $c_2 \in (0, \infty)$ satisfying for every $r \in (0, \mathfrak{r}]$ that

$$\begin{aligned}\mathbb{E}\left[\max_{1 \leq k \leq n}\left|\Theta^{M,r}_{k,\theta} - \Theta^{M,r}_{0,\theta}\right| \mathbf{1}_{k-1}\right] \leq \sum_{k=1}^{n}\mathbb{E}\left[\left|\Theta^{M,r}_{k,\theta} - \Theta^{M,r}_{k-1,\theta}\right| \mathbf{1}_{k-1}\right] \leq& \sum_{k=1}^{n}\mathbb{E}\left[\left|\Theta^{M,r}_{k,\theta} - \Theta^{M,r}_{k-1,\theta}\right|^2 \mathbf{1}_{k-1}\right]^{\frac{1}{2}} \\ \leq& c_1 r\left(\sum_{k=1}^{n} k^{-\frac{3}{2}\rho} + M^{-\frac{1}{2}}\sum_{k=1}^{n} k^{-\rho}\right) \\ \leq& c_2 r\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right),\end{aligned} \tag{5.57}$$

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

$$\sum_{k=1}^{n} k^{-\frac{3}{2}\rho} + M^{-\frac{1}{2}}\sum_{k=1}^{n} k^{-\rho} \leq c\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right). \tag{5.58}$$

This completes the proof. □

Remark 5.4. We emphasize that 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 (5.58), 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 5.3 is not needed if $\mathcal{M} \cap U$ is compact, since SGD cannot leave the basin of attraction in tangential directions, the results of Section 6 apply for $\rho \in (0, 1)$ under this additional compactness assumption.

We will next obtain a lower bound in probability for the events A_n, $n \in \mathbb{N}_0$. For this, we will first establish sufficient conditions for containment in the set $V_{R,\delta}(x_0)$. Effectively, these conditions split the normal and tangential movement of SGD in the sense that, in order to be outside the set $V_{R,\delta}(x_0)$, a point must be either distance greater than δ from $\mathcal{M} \cap U$ or be of distance roughly greater than R from x_0.

Lemma 5.5. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \dots, d-1\}$, and let $\mathcal{N} \subseteq \mathbb{R}^d$ be a $\mathfrak{d}$-dimensional C^1-submanifold. Then for every $x_0 \in \mathcal{N}$ there exists $R_0, \delta_0 \in (0, \infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$, for $V_{R,\delta}(x_0) \subseteq \mathbb{R}^d$ satisfying that*

$$V_{R,\delta}(x_0) = \{x + v \in \mathbb{R}^d \colon x \in \overline{B}_R(x_0) \cap \mathcal{N} \text{ and } v \in \big(T_x\mathcal{N}\big)^{\perp} \text{ with } |v| < \delta\}, \tag{5.59}$$

it holds that

$$\{x \in \mathbb{R}^d \colon \mathrm{d}(x, \mathcal{N}) < \delta \text{ and } |x - x_0| \leq R - \delta\ \} \subseteq V_{R,\delta}(x_0). \tag{5.60}$$

Proof of Lemma 5.5. Let $x_0 \in \mathcal{N}$, let $V_*(x_0) \subseteq \mathbb{R}^d$ be a neighborhood containing x_0 that satisfies the conclusion of Proposition 2.1, and let $R_0, \delta_0 \in (0, \infty)$ satisfy the conclusion of Proposition 2.5. That is, for every $R \in (0, R_0]$, $\delta \in (0, \delta_0]$ it holds that $\overline{V}_{R,\delta}(x_0) \subseteq V_*(x_0)$ and that

$$V_{R,\delta}(x_0) = \{x \in \mathbb{R}^d \colon \mathrm{d}(x, \mathcal{N}) = \mathrm{d}(x, \overline{B}_R(x_0) \cap \mathcal{N}) < \delta\}. \tag{5.61}$$

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

$$\mathrm{d}(x, \mathcal{N}) < \delta \ \text{ with } \ |x - x_0| \leq R - \delta. \tag{5.62}$$

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

$$|x - \tilde{x}| = \mathrm{d}(x, \mathcal{N}) < \delta. \tag{5.63}$$

The triangle inequality implies that

$$|\tilde{x} - x_0| \leq |\tilde{x} - x| + |x - x_0| < \delta + (R - \delta) < R. \tag{5.64}$$

It follows that $\tilde{x} \in \overline{\overline{B}_R(x_0) \cap \mathcal{N}}$, and therefore that

$$\mathrm{d}(x, \mathcal{N}) = \mathrm{d}(x, \overline{B}_R(x_0) \cap \mathcal{N}) < \delta. \tag{5.65}$$

It follows from (5.62) and (5.65) that $x \in V_{R,\delta}(x_0)$, which completes the proof. □

In the following proposition, we obtain a lower bound in probability for the sets A_n, $n \in \mathbb{N}_0$. The interesting observation is that Proposition 5.2 and Proposition 5.3, which obtain estimates for the solution of (5.1) conditioned on the events A_n, $n \in \mathbb{N}_0$, can be used together and inductively to obtain lower bound in probability for the events A_n, $n \in \mathbb{N}_0$. Namely, Proposition 5.2 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 5.3 can be used to estimate the probability that the solution (5.1) escapes the basin of attraction along the tangential directions.

Proposition 5.6. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \dots, d-1\}$, $\rho \in (2/3, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta, x) \in \mathbb{R}^d \times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}, \tag{5.66}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta \in \mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that

$\mathcal{M} \cap U$ is a non-empty $\mathfrak{d}$-dimensional C^1*-submanifold of $\mathbb{R}^d$, assume for every $\theta \in \mathcal{M} \cap U$ that* $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$*, for every $M \in \mathbb{N}$, $r \in (0,\infty)$, $\theta \in \mathbb{R}^d$ let $\Theta^{M,r}_{0,\theta} \in \mathbb{R}^d \colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\theta^{M,r}_{0,\theta}(\omega) = \theta$, for every $n, M \in \mathbb{N}$, $r \in (0,\infty)$, $\theta \in \mathbb{R}^d$ let $\Theta^{M,r}_{n,\theta} \colon \Omega \to \mathbb{R}^d$ satisfy that*

$$\Theta^{M,r}_{n,\theta} = \Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta^{M,r}_{n-1,\theta}, X_{n,m})\right], \tag{5.67}$$

and for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0,\infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M} \cap U$ let $A_n \subseteq \Omega$ be the event satisfying that

$$A_n = \left\{\omega \in \Omega \colon \Theta^{M,r}_{m,\theta}(\omega) \in V_{R,\delta}(x_0) \ \forall\, m \in \{0,\ldots,n\}\right\}. \tag{5.68}$$

Then for every $x_0 \in \mathcal{M} \cap U$ there exist $R_0, \delta_0, \mathfrak{r}, c \in (0,\infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M \in \mathbb{N}$, $\theta \in V_{R/2,\delta}(x_0)$ it holds that

$$\mathbb{P}[A_n] \geq \prod_{k=1}^{n}\left(1 - \frac{c}{Mk^{2\rho}}\right)_+ - cM^{-1}n^{1-\rho} - \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}. \tag{5.69}$$

Proof of Proposition 5.6. Let $x_0 \in \mathcal{M} \cap U$. Since $U \subseteq \mathbb{R}^d$ is open, fix a neighborhood $V_*(x_0)$ of x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0, \delta_0 \in (0,\infty)$ satisfying the conclusion of Proposition 2.5 for this set $V_*(x_0)$. Fix $\mathfrak{r} \in (0,\infty)$ satisfying the conclusion of Lemma 2.7.

Let $R \in (0, R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M \in \mathbb{N}$. As in Proposition 5.2, let $\nabla_\theta F^{M,n} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be the functions satisfying for every $(\theta,\omega) \in \mathbb{R}^d \times \Omega$ that

$$\nabla_\theta F^{M,n}(\theta) = \nabla_\theta F^{M,n}(\theta,\omega) = \frac{1}{M}\sum_{m=1}^{M}(\nabla_\theta F)(\theta, X_{n,m}(\omega)). \tag{5.70}$$

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

$$\Theta^{M,r}_{n,\theta} = \Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho}\nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta}). \tag{5.71}$$

Since it holds that

$$\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U) \geq \delta \ \text{ implies that } \ \Theta^{M,r}_{n,\theta} \notin V_{R,\delta}(x_0), \tag{5.72}$$

it follows that

$$\begin{aligned}\mathbb{P}\left[\Theta^{M,r}_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right] =& \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U) \geq \delta, A_{n-1}\right] \\ &+ \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right].\end{aligned} \tag{5.73}$$

The two terms on the righthand side of (5.73) will be handled separately.

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

$$\mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M} \cap U) \geq \delta, A_{n-1}\right] \leq \frac{c}{Mn^{2\rho}}\mathbb{P}\left[A_{n-1}\right] + \frac{c}{Mn^\rho}. \tag{5.74}$$

On the event A_{n-1}, it follows from Lemma 2.6 that there exists $\varepsilon_n \colon A_{n-1} \to \mathbb{R}^d$, $c_1 \in (0,\infty)$ such that

$$|\varepsilon_n| \leq c_1\, \mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M} \cap U)^2, \tag{5.75}$$

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

$$\nabla f(\Theta^{M,r}_{n-1,\theta}) = \nabla^2 f(\Theta^{M,r}_{n-1,\theta,*}) \cdot (\Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*}) + \varepsilon_n. \tag{5.76}$$

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

(5.77)
$$\Theta^{M,r}_{n,\theta} = \Theta^{M,r}_{n-1,\theta} - \frac{r}{n^\rho}\nabla^2 f(\Theta^{M,r}_{n-1,\theta,*})\cdot(\Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*}) - \frac{r}{n^\rho}\varepsilon_n + \frac{r}{n^\rho}\left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right).$$

Lemma 2.7, (5.75), 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

(5.78)
$$\begin{aligned}
&\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M}\cap U) \\
&\le \left|\Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*} - \frac{r}{n^\rho}\nabla^2 f(\Theta^{M,r}_{n-1,\theta,*})\cdot(\Theta^{M,r}_{n-1,\theta} - \Theta^{M,r}_{n-1,\theta,*})\right| \\
&\quad + \left|\frac{r}{n^\rho}\varepsilon_n\right| + \left|\frac{r}{n^\rho}\left(\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right)\right| \\
&\le \left(1 - \frac{r\lambda}{n^\rho}\right)\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M}\cap U) + c_1\frac{r}{n^\rho}\,\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M}\cap U)^2 + \frac{r}{n^\rho}\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right|.
\end{aligned}$$

Fix $\delta_1 \in (0,\delta_0]$ satisfying that

$$c_1\delta_1 \le \frac{\lambda}{2}. \tag{5.79}$$

Let $\delta \in (0,\delta_1]$. On the event A_{n-1}, it follows from (5.78) and the choice of $\delta_1 \in (0,\delta_0]$ that

$$\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M}\cap U) \le \left(1 - \frac{r\lambda}{2n^\rho}\right)\mathrm{d}(\Theta^{M,r}_{n-1,\theta}, \mathcal{M}\cap U) + \frac{r}{n^\rho}\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right|. \tag{5.80}$$

We therefore conclude that

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M}\cap U) \ge \delta, A_{n-1}\right] \le \\
&\mathbb{P}\left[\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right| \ge \frac{\delta n^\rho}{2r}, \Theta^{M,r}_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] \\
&\quad + \mathbb{P}\left[\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right| \ge \frac{\delta\lambda}{2}, \Theta^{M,r}_{n-1,\theta} \in V_{R,\delta}(x_0)\setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right].
\end{aligned} \tag{5.81}$$

Similarly to (5.19) and computation (5.20), it follows from the independence of the random variables $X_{m,k}$, $m,k\in\mathbb{N}$, that

$$\begin{aligned}
&\mathbb{P}\left[\,\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right| \ge \frac{\delta n^\rho}{2r}, \Theta^{M,r}_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] \\
&\le \sup_{\theta\in V_{R,\frac{\delta}{2}}(x_0)} \mathbb{P}\left[\,\left|\nabla f(\theta) - \nabla_\theta F^{M,n}(\theta)\right| \ge \frac{\delta n^\rho}{2r}\,\right]\mathbb{P}\left[\Theta^{M,r}_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right],
\end{aligned} \tag{5.82}$$

and that

(5.83)
$$\begin{aligned}
&\mathbb{P}\left[\,\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right| \ge \frac{\delta\lambda}{2}, \Theta^{M,r}_{n-1,\theta} \in V_{R,\delta}(x_0)\setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] \\
&\le \sup_{\theta\in V_{R,\delta}(x_0)\setminus V_{R,\frac{\delta}{2}}(x_0)} \mathbb{P}\left[\,\left|\nabla f(\theta) - \nabla_\theta F^{M,n}(\theta)\right| \ge \frac{\delta\lambda}{2}\,\right]\mathbb{P}\left[\Theta^{M,r}_{n-1,\theta} \in V_{R,\delta}(x_0)\setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right].
\end{aligned}$$

The definition of A_{n-1}, Chebyshev's inequality, Lemma 5.1, and (5.82) prove that there exists $c \in (0,\infty)$ satisfying that

$$\begin{aligned}\mathbb{P}\left[\,\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right| \geq \frac{\delta n^\rho}{2r}, \Theta^{M,r}_{n-1,\theta} \in V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] &\leq \frac{c}{M}\cdot\frac{4r^2}{\delta^2 n^{2\rho}}\mathbb{P}\left[A_{n-1}\right] \\ &\leq \frac{c}{Mn^{2\rho}}\mathbb{P}\left[A_{n-1}\right].\end{aligned}\tag{5.84}$$

In the case of (5.83), Proposition 5.2 and Chebyshev's inequality prove that, for the indicator function $\mathbf{1}_{n-2}$ of the event A_{n-2}, there exists $c \in (0,\infty)$ satisfying that

$$\begin{aligned}\mathbb{P}\left[\Theta^{M,r}_{n-1,\theta} \in V_{R,\delta}(x_0)\setminus V_{R,\frac{\delta}{2}}(x_0), A_{n-2}\right] &\leq \mathbb{P}\left[\left(\mathrm{d}\left(\Theta^{M,r}_{n-1,\theta}, \mathcal{M}\cap U\right)\wedge 1\right)^2 \mathbf{1}_{n-2} \geq \frac{\delta^2}{4}\right] \\ &\leq \frac{4c}{\delta^2}n^{-\rho} \\ &\leq cn^{-\rho},\end{aligned}\tag{5.85}$$

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 $(n-1)^{-\rho} \leq cn^{-\rho}$. Furthermore, Chebyshev's inequality and Lemma 5.1 prove that there exists $c \in (0,\infty)$ satisfying that

$$\mathbb{P}\left[\,\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right| \geq \frac{\delta\lambda}{2}\,\right] \leq \frac{c}{M}\cdot\frac{4}{\delta^2\lambda^2} \leq \frac{c}{M}.\tag{5.86}$$

Returning to (5.83), the previous two inequalities prove that there exists $c \in (0,\infty)$ satisfying that

$$\mathbb{P}\left[\,\left|\nabla f(\Theta^{M,r}_{n-1,\theta}) - \nabla_\theta F^{M,n}(\Theta^{M,r}_{n-1,\theta})\right| \geq \frac{\delta\lambda}{2}, \Theta^{M,r}_{n-1,\theta} \in V_\delta \setminus V_{\frac{\delta}{2}}, A_{n-2}\right] \leq \frac{c}{Mn^\rho}.\tag{5.87}$$

Combining (5.81), (5.84), and (5.87), there exists $c \in (0,\infty)$ such that

$$\mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M}\cap U) \geq \delta, A_{n-1}\right] \leq \frac{c}{Mn^{2\rho}}\mathbb{P}\left[A_{n-1}\right] + \frac{c}{Mn^\rho},\tag{5.88}$$

which completes the proof of (5.74).

Returning to (5.73), it follows from (5.88) that there exists $c \in (0,\infty)$ such that

$$\begin{aligned}&\mathbb{P}\left[\Theta^{M,r}_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right] \\ &\leq \frac{c}{Mn^{2\rho}}\mathbb{P}\left[A_{n-1}\right] + \frac{c}{Mn^\rho} + \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M}\cap U) < \delta, \Theta^{M,r}_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right].\end{aligned}\tag{5.89}$$

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

$$\begin{aligned}\mathbb{P}[A_n] =& \mathbb{P}\left[\Theta^{M,r}_{n,\theta} \in V_{R,\delta}(x_0), A_{n-1}\right] \\ &\geq \left(1 - \frac{c}{Mn^{2\rho}}\right)_+ \mathbb{P}\left[A_{n-1}\right] - \frac{c}{Mn^\rho} - \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n,\theta}, \mathcal{M}\cap U) < \delta, \Theta^{M,r}_{n,\theta} \notin V_{R,\delta}(x_0), A_{n-1}\right].\end{aligned}\tag{5.90}$$

We will prove inductively that (5.90) implies that there exists $c \in (0,\infty)$ such that for every $n \in \mathbb{N}$ it holds that

$$\mathbb{P}[A_n] \geq \prod_{k=1}^n\left(1 - \frac{c}{Mk^{2\rho}}\right)_+ - \sum_{k=1}^n\frac{c}{Mk^\rho} - \sum_{k=1}^n\mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M}\cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right].\tag{5.91}$$

The base case $n = 0$ follows immediately from $\theta \in V_{R/2,\delta}(x_0)$. For the inductive step, suppose that (5.95) is satisfied for some $n \in \mathbb{N}$. It follows from (5.90) that

$$\begin{aligned} \mathbb{P}\left[A_{n+1}\right] \geq \left(1 - \frac{c}{M(n+1)^{2\rho}}\right)_+ \mathbb{P}\left[A_n\right] - \frac{c}{M(n+1)^{\rho}} \\ - \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n+1,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{n+1,\theta} \notin V_{R,\delta}(x_0), A_n\right]. \end{aligned} \tag{5.92}$$

It then follows from the inductive hypothesis (5.95) that

$$\begin{aligned} &\mathbb{P}\left[A_{n+1}\right] \\ &\geq \prod_{k=1}^{n+1} \left(1 - \frac{c}{Mk^{2\rho}}\right)_+ \mathbb{P}\left[A_0\right] - \frac{c}{M(n+1)^{\rho}} \\ &- \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{n+1,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{n+1,\theta} \notin V_{R,\delta}(x_0), A_n\right] \\ &- \left(1 - \frac{c}{M(n+1)^{2\rho}}\right)_+ \left(\sum_{k=1}^{n} \frac{c}{Mk^{\rho}} + \sum_{k=1}^{n} \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right]\right), \end{aligned} \tag{5.93}$$

which proves that

$$\begin{aligned} &\mathbb{P}\left[A_{n+1}\right] \\ &\geq \prod_{k=1}^{n+1} \left(1 - \frac{c}{Mk^{2\rho}}\right)_+ \mathbb{P}\left[A_0\right] - \sum_{k=1}^{n+1} \frac{c}{Mk^{\rho}} - \sum_{k=1}^{n+1} \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right]. \end{aligned} \tag{5.94}$$

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

$$\mathbb{P}\left[A_{n+1}\right] \geq \prod_{k=1}^{n+1} \left(1 - \frac{c}{Mk^{2\rho}}\right)_+ - \sum_{k=1}^{n+1} \frac{c}{Mk^{\rho}} - \sum_{k=1}^{n+1} \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right], \tag{5.95}$$

which completes the induction step, and the proof of (5.95).

It remains only to estimate the final term on the righthand side of inequality (5.95). The definition of the events A_m, $m \in \mathbb{N}_0$, implies that

$$\{\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\} \subseteq \Omega,\ k \in \mathbb{N},\ \text{are disjoint events.} \tag{5.96}$$

Therefore, it holds that

$$\begin{aligned} &\sum_{k=1}^{n} \mathbb{P}\left[\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\right] \\ &= \mathbb{P}\left[\coprod_{k=1}^{n} \{\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\}\right]. \end{aligned} \tag{5.97}$$

Lemma 5.5 proves that

$$\mathbb{P}\left[\coprod_{k=1}^{n} \{\mathrm{d}(\Theta^{M,r}_{k,\theta}, \mathcal{M} \cap U) < \delta, \Theta^{M,r}_{k,\theta} \notin V_{R,\delta}(x_0), A_{k-1}\}\right] \leq \mathbb{P}\left[\max_{1 \leq k \leq n} \left|\Theta^{M,r}_{k,\theta} - x_0\right| \mathbf{1}_{k-1} > R - \delta\right]. \tag{5.98}$$

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

$$\begin{aligned} \left|\Theta_{k,\theta}^{M,r} - x_0\right| &\leq \left|\Theta_{k,\theta}^{M,r} - \Theta_{0,\theta}^{M,k}\right| + \left|\Theta_{0,\theta}^{M,k} - \Theta_{0,\theta,*}^{M,k}\right| + \left|\Theta_{0,\theta,*}^{M,r} - x_0\right| \\ &\leq \left|\Theta_{k,\theta}^{M,r} - \theta\right| + \delta + \frac{R}{2}. \end{aligned} \tag{5.99}$$

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

$$\frac{R}{2} - 2\delta < \left|\Theta_{k,\theta}^{M,r} - \Theta_{0,\theta}^{M,r}\right|. \tag{5.100}$$

This implies that

$$\left\{\max_{1 \leq k \leq n} \left|\Theta_{k,\theta}^{M,r} - x_0\right| \mathbf{1}_{k-1} > R - \delta\right\} \subseteq \left\{\max_{1 \leq k \leq n} \left|\Theta_{k,\theta}^{M,r} - \Theta_{0,\theta}^{M,r}\right| \mathbf{1}_{k-1} > \frac{R}{2} - 2\delta\right\}. \tag{5.101}$$

In combination, (5.97), (5.98), and (5.101) prove that

$$\sum_{k=1}^{n} \mathbb{P}\left[\mathrm{d}(\Theta_{k,\theta}^{M,r}, \mathcal{M} \cap U) < \delta, \Theta_{k,\theta}^{M,r} \notin V_{R,\delta}(x_0), A_{k-1}\right] \leq \mathbb{P}\left[\max_{1 \leq k \leq n} \left|\Theta_{k,\theta}^{M,r} - \Theta_{0,\theta}^{M,r}\right| \mathbf{1}_{k-1} > \frac{R}{2} - 2\delta\right]. \tag{5.102}$$

It follows from Proposition 5.3, (5.102), and Chebyshev's inequality that there exists $c \in (0, \infty)$ satisfying that

$$\sum_{k=1}^{n} \mathbb{P}\left[\mathrm{d}(\Theta_{k,\theta}^{M,r}, \mathcal{M} \cap U) < \delta, \Theta_{k,\theta}^{M,r} \notin V_{R,\delta}(x_0), A_{k-1}\right] \leq \frac{cr\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}. \tag{5.103}$$

Returning to (5.95), it follows from (5.103) that there exists $c \in (0, \infty)$ satisfying that

$$\mathbb{P}[A_n] \geq \prod_{k=1}^{n} \left(1 - \frac{c}{Mk^{2\rho}}\right)_+ - cM^{-1} n^{1-\rho} - \frac{cr\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}, \tag{5.104}$$

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

$$\sum_{k=1}^{n} k^{-\rho} \leq cn^{1-\rho}. \tag{5.105}$$

This completes the proof. □

We will now use Proposition 5.2 and Proposition 5.6 to estimate the probability that SGD of mini-batch size $M \in \mathbb{N}$ converges to within distance $\varepsilon \in (0, 1]$ of the manifold of local minima at time $n \in \mathbb{N}$. In the theorem, we assume that the initial condition $\Theta_0^{M,r}$ is continuous uniformly distributed on a bounded open subset $A \subseteq \mathbb{R}^d$ satisfying $\mathcal{M} \cap U \cap A \neq \emptyset$.

Theorem 5.7. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \dots, d-1\}$, $\rho \in (2/3, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $\lambda \colon \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ be the Lebesgue-Borel measure, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta, x) \in \mathbb{R}^d \times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \left\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\right\}, \tag{5.106}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta \in \mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that

$\mathcal{M}\cap U$ is a $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M}\cap U\cap A\neq\emptyset$, assume for every $\theta\in\mathcal{M}\cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta))=d-\mathfrak{d}$, for every $M\in\mathbb{N}$, $r\in(0,\infty)$ let $\Theta_0^{M,r}\colon\Omega\to\mathbb{R}^d$ be continuous uniformly distributed on A, assume for every $M\in\mathbb{N}$, $r\in(0,\infty)$ that $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, for every $M\in\mathbb{N}$, $r\in(0,\infty)$ let $\Theta_{n,\theta}^{M,r}\colon\Omega\to\mathbb{R}^d$, $n\in\mathbb{N}$, be random variables which satisfy that

$$\Theta_{n,\theta}^{M,r}=\Theta_{n-1,\theta}^{M,r}-\frac{r}{n^\rho M}\left[\sum_{m=1}^{M}(\nabla_\theta F)(\Theta_{n-1,\theta}^{M,r},X_{n,m})\right]. \tag{5.107}$$

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

$$\begin{aligned}&\mathbb{P}\Big(\operatorname{d}\big(\Theta_{n,\theta}^{M,r},\mathcal{M}\cap U\big)\geq\varepsilon\Big)\leq\\ &\frac{\lambda\big(A\backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)}+c\varepsilon^{-2}n^{-\rho}+1-\prod_{k=1}^{n}\Big(1-\frac{c}{Mk^{2\rho}}\Big)_+ +cM^{-1}n^{1-\rho}+\frac{cr\left(1+M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2}-2\delta\right)_+}.\end{aligned} \tag{5.108}$$

Proof of Theorem 5.7. Let $x_0\in\mathcal{M}\cap U\cap A$. Since $U\subseteq\mathbb{R}^d$ is open, fix a neighborhood $V_*(x_0)\subseteq U$ containing x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0,\delta_0\in(0,\infty)$ satisfying the conclusion of Proposition 2.5 for this set $V_*(x_0)$. Fix $\mathfrak{r}\in(0,\infty)$ satisfying the conclusions of Lemma 2.7 and Proposition 5.6.

Let $R\in(0,R_0]$, $\delta\in(0,\delta_0]$, $r\in(0,\mathfrak{r}]$, $M\in\mathbb{N}$. As in Proposition 5.2, let $\nabla_\theta F^{M,n}:\mathbb{R}^d\times\Omega\to\mathbb{R}^d$, $n\in\mathbb{N}$, be the functions satisfying for every $(\theta,\omega)\in\mathbb{R}^d\times\Omega$ that

$$\nabla_\theta F^{M,n}(\theta)=\nabla_\theta F^{M,n}(\theta,\omega)=\frac{1}{M}\sum_{m=1}^{M}(\nabla_\theta F)(\theta,X_{n,m}(\omega)). \tag{5.109}$$

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

$$\Theta_{n,\theta}^{M,r}=\Theta_{n-1,\theta}^{M,r}-\frac{r}{n^\rho}\nabla_\theta F^{M,n}(\Theta_{n-1,\theta}^{M,r}). \tag{5.110}$$

Let $\Theta_0^{M,r}\colon\Omega\to\mathbb{R}^d$ be a random variable which is continuous uniformly distributed on A, assume that $\Theta_0^{M,r}$ and $(X_{n,m})_{n,m\in\mathbb{N}}$ are independent, and for every $n\in\mathbb{N}$ let $\Theta_n^{M,r}\colon\Omega\to\mathbb{R}^d$ satisfy that $\Theta_n^{M,r}=\Theta_{n,\Theta_0^{M,r}}^{M,r}$.

Let $n\in\mathbb{N}$, $\varepsilon\in(0,1]$. It holds that

$$\begin{aligned}\mathbb{P}\Big(\operatorname{d}\big(\Theta_n^{M,r},\mathcal{M}\cap U\big)\geq\varepsilon\Big)=&\,\mathbb{P}\Big(\operatorname{d}\big(\Theta_n^{M,r},\mathcal{M}\cap U\big)\geq\varepsilon,\Theta_0^{M,r}\in V_{R/2,\delta}(x_0)\Big)\\ &+\mathbb{P}\Big(\operatorname{d}\big(\Theta_n^{M,r},\mathcal{M}\cap U\big)\geq\varepsilon,\Theta_0^{M,r}\notin V_{R/2,\delta}(x_0)\Big).\end{aligned} \tag{5.111}$$

For the second term on the righthand side of (5.108), it follows from the continuous uniform distribution of $\Theta_0^{M,r}$ on A that

$$\mathbb{P}\Big(\operatorname{d}\big(\Theta_n^{M,r},\mathcal{M}\cap U\big)\geq\varepsilon,\Theta_0^{M,r}\notin V_{R/2,\delta}(x_0)\Big)\leq\frac{\lambda\big(A\backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)}. \tag{5.112}$$

We will now estimate the first term on the righthand side of (5.111). For every $m\in\mathbb{N}_0$, $\theta\in\mathbb{R}^d$ let $A_{m,\theta}\subseteq\Omega$ be the event satisfying that

$$A_{m,\theta}=\Big\{\omega\in\Omega\colon\Theta_{k,\theta}^{M,r}(\omega)\in V_{R,\delta}(x_0)\;\forall\,k\in\{0,\dots,m\}\Big\}, \tag{5.113}$$

and for every $m \in \mathbb{N}_0$ let $A_m \subseteq \Omega$ be the event satisfying that

$$A_m = \Big\{\omega \in \Omega\colon \Theta_k^{M,r}(\omega) \in V_{R,\delta}(x_0)\ \forall\, k \in \{0,\ldots,m\}\Big\}. \tag{5.114}$$

It holds that

$$\begin{aligned}
&\mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, \Theta_0^{M,r} \in V_{R/2,\delta}(x_0)\Big)\\
&= \mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, \Theta_0^{M,r} \in V_{R/2,\delta}(x_0), A_{n-1}\Big)\\
&\quad + \mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, \Theta_0^{M,r} \in V_{R/2,\delta}(x_0), \Omega\backslash A_{n-1}\Big).
\end{aligned} \tag{5.115}$$

For the second term on the righthand side of (5.115), it follows from Proposition 5.6 that there exists $c \in (0,\infty)$ satisfying that

$$\begin{aligned}
&\mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, \Theta_0^{M,r} \notin V_{R/2,\delta}(x_0), \Omega\backslash A_{n-1}\Big)\\
&\leq 1 - \prod_{k=1}^{n}\Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+},
\end{aligned} \tag{5.116}$$

where we have used the fact that $\rho \in (2/3, 1)$ implies that there exists $c \in (0,\infty)$ satisfying 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 (5.115), since the random variables $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, it holds that

$$\begin{aligned}
&\mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, \Theta_0^{M,r} \in V_{R/2,\delta}(x_0), A_{n-1}\Big)\\
&\leq \frac{\lambda\big(V_{R/2,\delta}(x_0)\cap A\big)}{\lambda(A)} \sup_{\theta \in V_{R/2,\delta}(x_0)} \mathbb{P}\Big(\,\mathrm{d}\left(\Theta_{n,\theta}^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, A_{n-1,\theta}\Big).
\end{aligned} \tag{5.117}$$

For the indicator function $\mathbf{1}_{n-1,\theta}$ of $A_{n-1,\theta}$, Proposition 5.2 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

$$\mathbb{P}\Big(\,\mathrm{d}\left(\Theta_{n,\theta}^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, A_{n-1,\theta}\Big) \leq \varepsilon^{-2}\mathbb{E}\Big[\Big(\,\mathrm{d}\left(\Theta_{n,\theta}^{M,r}, \mathcal{M}\cap U\right)\wedge 1\Big)^2 \mathbf{1}_{n-1,\theta}\Big] \leq c\varepsilon^{-2}n^{-\rho}. \tag{5.118}$$

In combination (5.117) and (5.118) prove that there exists $c \in (0,\infty)$ satisfying that

$$\mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, \Theta_0^{M,r} \in V_{R/2,\delta}(x_0), A_{n-1}\Big) \leq c\varepsilon^{-2}n^{-\rho}. \tag{5.119}$$

Returning to (5.115), it follows from (5.116) and (5.119) that there exists $c \in (0,\infty)$ satisfying that

$$\begin{aligned}
&\mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon, \Theta_0^{M,r} \in V_{R/2,\delta}(x_0)\Big)\\
&\leq c\varepsilon^{-2}n^{-\rho} + 1 - \prod_{k=1}^{n}\Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}.
\end{aligned} \tag{5.120}$$

Returning finally to (5.111), it follows from (5.112) and (5.120) that there exists $c \in (0,\infty)$ satisfying that

$$\begin{aligned}
&\mathbb{P}\Big(\,\mathrm{d}\left(\Theta_n^{M,r}, \mathcal{M}\cap U\right) \geq \varepsilon\Big) \leq\\
&\frac{\lambda\big(A\backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + c\varepsilon^{-2}n^{-\rho} + 1 - \prod_{k=1}^{n}\Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+},
\end{aligned} \tag{5.121}$$

which completes the proof. □

The next corollary estimates the probability that $K \in \mathbb{N}$ independent samples of SGD with mini-batch size $M \in \mathbb{N}$ fail to to converge to within distance $\varepsilon \in (0,1]$ of the manifold of local minima $\mathcal{M} \cap U$ at time $n \in \mathbb{N}$. The proof is a straightforward consequence of Theorem 5.7 and the independence of the random variables.

Corollary 5.8. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0,1,\dots,d-1\}$, $\rho \in (2/3,1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $(S,\mathcal{S})$ be a measurable space, let $F = (F(\theta,x))_{(\theta,x)\in\mathbb{R}^d\times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n,m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta,X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta,X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)]\big\}, \tag{5.122}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta,x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|F(\theta,X_{1,1})|^2 + |(\nabla_\theta F)(\theta,X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M} \cap U \cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $n \in \mathbb{N}_0$, $M \in \mathbb{N}$, $r \in (0,\infty)$ let $\Theta_n^{k,M,r} \colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $M \in \mathbb{N}$, $r \in (0,\infty)$ that $\Theta_0^{1,M,r}$ is continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0,\infty)$ that $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, and assume for every $n, M \in \mathbb{N}$, $r \in (0,\infty)$ that

$$\Theta_n^{1,M,r} = \Theta_{n-1}^{1,M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}^{1,M,r}, X_{n,m})\right]. \tag{5.123}$$

Then for every $x_0 \in \mathcal{M} \cap U \cap A$ 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

$$\begin{gathered}\mathbb{P}\Big(\min_{k\in\{1,2,\dots,K\}} \mathrm{d}(\Theta_n^{k,M,r}, \mathcal{M}\cap U) \geq \varepsilon\Big) \leq \\ \left(\frac{\lambda(A\backslash V_{R/2,\delta}(x_0))}{\lambda(A)} + c\varepsilon^{-2}n^{-\rho} + 1 - \prod_{k=1}^{n}\Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}\right)^K.\end{gathered} \tag{5.124}$$

Proof of Corollary 5.8. Let $x_0 \in \mathcal{M} \cap U \cap A$. Since $U \subseteq \mathbb{R}^d$ is open, fix a neighborhood $V_*(x_0) \subseteq U$ containing x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0, \delta_0 \in (0,\infty)$ satisfying the conclusion of Proposition 2.5 for this set $V_*(x_0)$. Fix $\mathfrak{r} \in (0,\infty)$ satisfying the conclusions of Lemma 2.7 and Proposition 5.6.

Let $R \in (0,R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M, K \in \mathbb{N}$. Since the $\Theta_n^{k,M,r}$, $k \in \mathbb{N}$, are i.i.d. it holds that

$$\begin{aligned}\mathbb{P}\Big(\min_{k\in\{1,2,\dots,K\}} \mathrm{d}(\Theta_n^{k,M,r}, \mathcal{M}\cap U) \geq \varepsilon\Big) &= \prod_{k=1}^{K} \mathbb{P}\Big(\mathrm{d}(\Theta_n^{k,M,r}, \mathcal{M}\cap U) \geq \varepsilon\Big) \\ &= \mathbb{P}\Big(\mathrm{d}(\Theta_n^{1,M,r}, \mathcal{M}\cap U) \geq \varepsilon\Big)^K.\end{aligned} \tag{5.125}$$

Theorem 5.7 and (5.125) prove estimate (5.124), which completes the proof. □

The following corollary translates the convergence of $\Theta_n^{k,M,r}$, $k \in \{1,2,\dots,K\}$, to the local manifold of minima $\mathcal{M} \cap U$ into a statement concerning the minimization of the objective function. The proof is a consequence of Corollary 5.8 and the local regularity of the objective function.

Corollary 5.9. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \dots, d-1\}$, $\rho \in ({}^{2}\!/\!{}_{3}, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta,x)\in\mathbb{R}^d\times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)]\big\}, \tag{5.126}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M} \cap U \cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $n \in \mathbb{N}_0$, $M \in \mathbb{N}$, $r \in (0, \infty)$ let $\Theta^{k,M,r}_n \colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta^{1,M,r}_0$ is continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta^{M,r}_0$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, and assume for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$ that

$$\Theta^{1,M,r}_n = \Theta^{1,M,r}_{n-1} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta^{1,M,r}_{n-1}, X_{n,m})\right]. \tag{5.127}$$

Then for every $x_0 \in \mathcal{M} \cap U \cap A$ 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

$$\begin{split} &\mathbb{P}\Big(\Big[\big[\min_{k\in\{1,2,\dots,K\}} f(\Theta^{k,M,r}_n)\big] - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \\ &\left(\frac{\lambda\big(A\backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + c\varepsilon^{-2}n^{-\rho} + 1 - \prod_{k=1}^{n}\Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2} - 2\delta\right)_+}\right)^K. \end{split} \tag{5.128}$$

Proof of Corollary 5.9. The proof is an immediate consequence of Corollary 5.8 and the local regularity of the objective function. □

Under the assumptions and notations of Corollary 5.9, since a random variable $\varTheta^{K,M,r}_n \colon \Omega \to \mathbb{R}^d$ satisfying for every $\omega \in \Omega$ that

$$\varTheta^{K,M,r}(\omega) \in \Big[\operatorname*{argmin}_{\theta\in\{\Theta^{k,M,r}_n(\omega)\colon k\in\{1,2,\dots,K\}} f(\theta)\Big], \tag{5.129}$$

is either computationally inefficient or computationally impossible to obtain, we will prove that such a minimizer can be efficiently computed using mini-batch averages. In the following lemma, we prove that there exists a measurable selection that minimizes a mini-batch approximation.

Lemma 5.10. *Let $d \in \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F : \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_k : \Omega \to S$, $k \in \mathbb{N}$, be i.i.d. random variables, and let $\Theta^k : \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables. Then for every $\mathfrak{M} \in \mathbb{N}$ there exists a random variable $\varTheta^{\mathfrak{M}} : \Omega \to \mathbb{R}^d$ satisfying for every $\omega \in \Omega$ that*

$$\varTheta^{\mathfrak{M}}(\omega) \in \Big[\operatorname*{argmin}_{\theta\in\{\Theta^k(\omega)\colon k\in\{1,2,\dots,\mathfrak{M}\}\}} \Big[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_m)\Big]\Big]. \tag{5.130}$$

Proof of Lemma 5.10. Let $\mathfrak{M} \in \mathbb{N}$. Let $\mathfrak{m} : \Omega \to \{1, 2, \dots, \mathfrak{M}\}$ satisfy for every $\omega \in \Omega$ that

$$\mathfrak{m}(\omega) = \min\Big\{k \in \{1, 2, \dots, \mathfrak{M}\} \colon \Theta^k(\omega) \in \Big[\operatorname*{argmin}_{\theta\in\{\Theta^k(\omega)\colon k\in\{1,2,\dots,\mathfrak{M}\}\}} \Big[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_m)\Big]\Big]\Big\}. \tag{5.131}$$

Let $\varTheta^{\mathfrak{M}} : \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that

$$\varTheta^{\mathfrak{M}}(\omega) = \Theta^{\mathfrak{m}(\omega)}(\omega). \tag{5.132}$$

It follow from (5.131) and (5.132) that $\Theta^{\mathfrak{M}}$ is measurable and satisfies (5.130), which completes the proof. □

In the following theorem, we prove that the minimum appearing on the lefthand side of (5.124) can be efficiently computed using mini-batch averages of the type appearing in Lemma 5.10.

Theorem 5.11. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \dots, d-1\}$, $\rho \in (2/3, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta,x)\in\mathbb{R}^d\times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)]\big\}, \tag{5.133}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M} \cap U \cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $n \in \mathbb{N}_0$, $M \in \mathbb{N}$, $r \in (0, \infty)$ let $\Theta_n^{k,M,r} \colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$ that $(\Theta_{n-1}^{k,M,r})_{k\in\mathbb{N}}$ and $(X_{n,k})_{k\in\mathbb{N}}$ are independent, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{1,M,r}$ is continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, assume for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$ that

$$\Theta_n^{1,M,r} = \Theta_{n-1}^{1,M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}^{1,M,r}, X_{n,m})\right], \tag{5.134}$$

and for every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $r \in (0, \infty)$ let $\varTheta_n^{K,M,\mathfrak{M},r} \colon \Omega \to \mathbb{R}^d$ be a random variable which satisfies for every $\omega \in \Omega$ that

$$\varTheta_n^{K,M,\mathfrak{M},r}(\omega) \in \left[\operatorname*{argmin}_{\theta\in\{\Theta_n^{k,M,r}(\omega)\colon k\in\{1,\dots,K\}\}} \left[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega))\right]\right]. \tag{5.135}$$

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

$$\begin{aligned}
&\mathbb{P}\Big(\Big[f(\varTheta_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{cK}{\varepsilon^2\mathfrak{M}} \\
&\quad + \left(\frac{\lambda\big(A \backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + c\varepsilon^{-2}n^{-\rho} + 1 - \prod_{k=1}^{n}\Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1}n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}}n^{1-\rho}\right)}{\big(\frac{R}{2} - 2\delta\big)_+}\right)^K.
\end{aligned} \tag{5.136}$$

Proof of Theorem 5.11. Let $x_0 \in \mathcal{M} \cap U \cap A$. Since $U \subseteq \mathbb{R}^d$ is open, fix a neighborhood $V_*(x_0) \subseteq U$ of x_0 that satisfies the conclusion of Proposition 2.1. Fix $R_0, \delta_0 \in (0, \infty)$ satisfying the conclusion of Proposition 2.5 for this set $V_*(x_0)$. Fix $\mathfrak{r} \in (0, \infty)$ satisfying the conclusions of Lemma 2.7 and Proposition 5.6.

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 $i \in \{1, 2, \dots, K\}$ let $B_i' \subseteq \Omega$ satisfy that

$$B_i' = \Big\{ \omega \in \Omega \colon \Theta_n^{i,M,r}(\omega) \in \Big[\operatorname*{argmin}_{\theta \in \{\Theta_n^{k,M,r}(\omega) \colon k \in \{1,2,\dots,K\}} \big[f(\theta) \big] \Big] \Big\}, \tag{5.137}$$

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

$$\begin{aligned}
&\mathbb{P}\Big(\Big[f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta \in \mathbb{R}^d} f(\theta) \Big] \geq \varepsilon \Big) \\
&= \sum_{i=1}^{K} \mathbb{P}\Big(\Big[f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta \in \mathbb{R}^d} f(\theta) \Big] \geq \varepsilon, B_i \Big) \\
&= \sum_{i=1}^{K} \mathbb{P}\Big(\big[f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - f(\Theta_n^{i,M,r}) + f(\Theta_n^{i,M,r}) - \inf_{\theta \in \mathbb{R}^d} f(\theta) \big] \geq \varepsilon, B_i \Big) \\
&\leq \mathbb{P}\Big(\Big[\big[\min_{k \in \{1,2,\dots,K\}} f(\Theta_n^{k,M,r}) \big] - \inf\nolimits_{\theta \in \mathbb{R}^d} f(\theta) \Big] \geq \frac{\varepsilon}{2} \Big) + \sum_{i=1}^{K} \mathbb{P}\Big(\big[f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - f(\Theta_n^{i,M,r}) \geq \frac{\varepsilon}{2}, B_i \Big).
\end{aligned} \tag{5.138}$$

For the first term on the righthand side of (5.138), Corollary 5.9 proves that there exists $c \in (0, \infty)$ satisfying that

$$\begin{aligned}
&\mathbb{P}\Big(\Big[\big[\min_{k \in \{1,2,\dots,K\}} f(\Theta_n^{k,M,r}) \big] - \inf\nolimits_{\theta \in \mathbb{R}^d} f(\theta) \Big] \geq \frac{\varepsilon}{2} \Big) \\
&\leq \left(\frac{\lambda\big(A \backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + c\varepsilon^{-2} n^{-\rho} + 1 - \prod_{k=1}^{n} \Big(1 - \frac{c}{M k^{2\rho}} \Big)_+ + c M^{-1} n^{1-\rho} + \frac{cr \left(1 + M^{-\frac{1}{2}} n^{1-\rho} \right)}{\left(\frac{R}{2} - 2\delta \right)_+} \right)^K.
\end{aligned} \tag{5.139}$$

We will now estimate the second term on the righthand side of (5.139). Let $\tilde{B}_j \subseteq \Omega$, $j \in \{1, 2, \dots, K\}$, be disjoint events which satisfy that $\Omega = \coprod_{j \in \{1,2,\dots,K\}} \tilde{B}_j$ and that

$$\tilde{B}_j \subseteq \Big\{ \omega \in \Omega \colon \mathit{\Theta}_n^{K,M,\mathcal{M},r}(\omega) = \Theta_n^{j,M,r}(\omega) \Big\}. \tag{5.140}$$

Since the events $\tilde{B}_j$, $j \in \{1, 2, \dots, K\}$, are disjoint, the final term of (5.138) satisfies that

$$\sum_{i=1}^{K} \mathbb{P}\Big(f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - f(\Theta_n^{i,M,r}) \geq \frac{\varepsilon}{2}, B_i \Big) = \sum_{i,j=1}^{K} \mathbb{P}\Big(f(\Theta_n^{j,M,r}) - f(\Theta_n^{i,M,r}) \geq \frac{\varepsilon}{2}, B_i, \tilde{B}_j \Big). \tag{5.141}$$

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

$$F^{\mathfrak{M},n}(\theta, \omega) = \frac{1}{\mathfrak{M}} \sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega)). \tag{5.142}$$

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

$$F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega) - F^{\mathfrak{M},n}(\Theta_n^{i,M,r}(\omega), \omega) \leq 0, \tag{5.143}$$

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

$$\begin{aligned}
&\mathbb{P}\Big(f(\Theta_n^{j,M,r}) - f(\Theta_n^{i,M,r}) \geq \frac{\varepsilon}{2}, B_i, \tilde{B}_j\Big) \\
&\quad\leq \mathbb{P}\Big(f(\Theta_n^{j,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega) + F^{\mathfrak{M},n}(\Theta_n^{i,M,r}(\omega), \omega) - f(\Theta_n^{i,M,r}(\omega)) \geq \frac{\varepsilon}{2}, B_i, \tilde{B}_j\Big) \\
&\quad\leq \mathbb{P}\Big(\Big|f(\Theta_n^{j,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, B_i, \tilde{B}_j\Big) \\
&\qquad+ \mathbb{P}\Big(\Big|f(\Theta_n^{i,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{i,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, B_i, \tilde{B}_j\Big).
\end{aligned} \tag{5.144}$$

It follows from (5.141) and (5.144) that

$$\begin{aligned}
&\sum_{i=1}^{K} \mathbb{P}\Big(f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - f(\Theta_n^{i,M,r}) \geq \frac{\varepsilon}{2}, B_i\Big) \\
&\leq \sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta_n^{j,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, \tilde{B}_j\Big) \\
&\quad+ \sum_{i=1}^{K} \mathbb{P}\Big(\Big|f(\Theta_n^{i,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{i,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, B_i\Big).
\end{aligned} \tag{5.145}$$

For the first term on the righthand side of (5.145), it holds that

$$\begin{aligned}
&\sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta_n^{j,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, \tilde{B}_j\Big) \\
&\quad\leq \sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta_n^{j,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}\Big).
\end{aligned} \tag{5.146}$$

Since the random variables $(\Theta_n^{k,M,r})_{k\in\mathbb{N}}$ and $(X_{n+1,k})_{k\in\mathbb{N}}$ are independent, since the $(\Theta_n^{k,M,r})_{k\in\mathbb{N}}$ are identically distributed, and since the distribution of $\Theta_n^{1,M,r}$ has bounded support on $\mathbb{R}^d$, for the distribution μ_n of $\Theta_n^{1,M,r}$ on $\mathbb{R}^d$, Lemma 5.1, Chebyshev's inequality, and the definition of $F^{\mathfrak{M},n}$ prove that that there exists $c \in (0, \infty)$ satisfying for every $j \in \{1, \ldots, K\}$ that

$$\begin{aligned}
\mathbb{P}\Big(\Big|f(\Theta_n^{j,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}\Big) &= \int_{\mathbb{R}^d} \mathbb{P}\Big(\Big|f(\theta) - \frac{1}{\mathfrak{M}}\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m})\Big| \geq \frac{\varepsilon}{4}\Big)\mu_n(\mathrm{d}\,\theta) \\
&\leq \frac{c}{\varepsilon^2 \mathfrak{M}}.
\end{aligned} \tag{5.147}$$

Therefore, it holds that

$$\sum_{j=1}^{K} \mathbb{P}\Big(\Big|f(\Theta_n^{j,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{j,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, \tilde{B}_j\Big) \leq \frac{cK}{\varepsilon^2 \mathfrak{M}}. \tag{5.148}$$

For the second term on the righthand side of (5.145), it is sufficient to apply the same argument, which proves that there exists $c \in (0, \infty)$ satisfying that

$$\sum_{i=1}^{K} \mathbb{P}\Big(\Big|f(\Theta_n^{i,M,r}(\omega)) - F^{\mathfrak{M},n}(\Theta_n^{i,M,r}(\omega), \omega)\Big| \geq \frac{\varepsilon}{4}, B_i\Big) \leq \frac{cK}{\varepsilon^2 \mathfrak{M}}. \tag{5.149}$$

Returning to (5.141), it follows from (5.145) and (5.148) that there exists $c \in (0, \infty)$ satisfying that

$$\sum_{i=1}^{K} \mathbb{P}\Big(f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - f(\Theta_n^{i,M,r}) \geq \frac{\varepsilon}{2}, B_i\Big) \leq \frac{cK}{\varepsilon^2 \mathfrak{M}}. \tag{5.150}$$

Returning finally to (5.138), it follows from (5.139) and (5.150) that there exists $c \in (0, \infty)$ satisfying that

$$\begin{aligned} &\mathbb{P}\Big(\big[f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\big] \geq \varepsilon\Big) \leq \frac{cK}{\varepsilon^2 \mathfrak{M}} \\ &+ \left(\frac{\lambda\big(A \backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + c\varepsilon^{-2} n^{-\rho} + 1 - \prod_{k=1}^{n} \Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1} n^{1-\rho} + \frac{cr\left(1 + M^{-\frac{1}{2}} n^{1-\rho}\right)}{\big(\frac{R}{2} - 2\delta\big)_+}\right)^K, \end{aligned} \tag{5.151}$$

which completes the proof. □

In the final corollary of this section, we will compute the computational efficiency of the algorithm proposed in Theorem 5.11. The constant implicitly depends on the computational cost of computing F and $\nabla_\theta F$ and initializing the random variable $X_{1,1}$, 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 $M, \mathfrak{M} \in \mathbb{N}$.

Corollary 5.12. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, $\rho \in (2/3, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta,x)\in\mathbb{R}^d\times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)]\big\}, \tag{5.152}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M} \cap U \cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $n \in \mathbb{N}_0$, $M \in \mathbb{N}$, $r \in (0, \infty)$ let $\Theta_n^{k,M,r} \colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$ that $(\Theta_{n-1}^{k,M,r})_{k\in\mathbb{N}}$ and $(X_{n,k})_{k\in\mathbb{N}}$ are independent, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{1,M,r}$ is continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, assume for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$ that

$$\Theta_n^{1,M,r} = \Theta_{n-1}^{1,M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}^{1,M,r}, X_{n,m})\right], \tag{5.153}$$

and for every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $r \in (0, \infty)$ let $\mathit{\Theta}_n^{K,M,\mathfrak{M},r} \colon \Omega \to \mathbb{R}^d$ be a random variable which satisfies for every $\omega \in \Omega$ that

$$\mathit{\Theta}_n^{K,M,\mathfrak{M},r}(\omega) \in \left[\operatorname*{argmin}_{\theta\in\{\Theta_n^{k,M,r}(\omega)\colon k\in\{1,\ldots,K\}\}} \left[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega))\right]\right]. \tag{5.154}$$

Then for every $x_0 \in \mathcal{M} \cap U \cap A$ 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 $n(\varepsilon), M(\varepsilon), K(\eta), \mathfrak{M}(\varepsilon, \eta) \in \mathbb{N}$ satisfying that

$$n(\varepsilon) = c_1 \varepsilon^{-2/\rho}, \;\; M(\varepsilon) = c_2 \varepsilon^{-4/\rho+4}, \;\; \mathfrak{M}(\varepsilon, \eta) = c_3 \varepsilon^{-2} \eta^{-1} \left|\log(\eta)\right|, \;\; \textit{and} \;\; K = c_4 \left|\log(\eta)\right|, \tag{5.155}$$

it holds that

$$\mathbb{P}\Big(\Big[f(\Theta_{n(\varepsilon)}^{K(\eta),M(\varepsilon),\mathfrak{M}(\varepsilon,\eta),r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \eta. \tag{5.156}$$

Proof of Corollary 5.12. Let $x_0 \in \mathcal{M} \cap U$. Let $R_0, \delta_0, \mathfrak{r} \in (0,\infty)$ satisfy the conclusion of Theorem 5.11. Theorem 5.11 proves that there exists $\overline{c} \in (0,\infty)$ such that for every $R \in (0, R_0]$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M, \mathfrak{M}, K \in \mathbb{N}$, $\varepsilon \in (0,1]$ it holds that

$$\begin{aligned}&\mathbb{P}\Big(\Big[f(\Theta_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{\overline{c}K}{\varepsilon^2\mathfrak{M}}\\ &+\left(\frac{\lambda\big(A\backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + \overline{c}\varepsilon^{-2}n^{-\rho} + 1 - \prod_{k=1}^{n}\left(1-\frac{\overline{c}}{Mk^{2\rho}}\right)_+ + \overline{c}M^{-1}n^{1-\rho} + \frac{\overline{c}r\left(1+M^{-\frac{1}{2}}n^{1-\rho}\right)}{\left(\frac{R}{2}-2\delta\right)_+}\right)^K.\end{aligned} \tag{5.157}$$

Fix $\overline{R} \in (0, R_0]$, $\overline{\delta} \in (0, \delta_0]$ satisfying that

$$\frac{\overline{R}}{2} - 2\overline{\delta} > 0. \tag{5.158}$$

Since $\mathcal{M} \cap U \cap A \neq \emptyset$, it holds that

$$\frac{\lambda\big(A\backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} \in (0,1). \tag{5.159}$$

For every $M \in \mathbb{N}$ satisfying that $M \geq 2\overline{c}$, since $\rho \in (2/3, 1)$ there exists $c \in (0,\infty)$ satisfying that

$$\log\Big(\prod_{k=1}^{n}\left(1-\frac{\overline{c}}{Mk^{2\rho}}\right)_+\Big) \geq -\frac{c}{M}\sum_{k=1}^{M} k^{-2\rho} \geq -\frac{c}{M}, \tag{5.160}$$

and therefore for every $M \geq 2\overline{c}$ there exists $c \in (0,\infty)$ satisfying that

$$-\prod_{k=1}^{n}\left(1-\frac{\overline{c}}{Mk^{2\rho}}\right)_+ \leq -\exp\Big(-\frac{c}{M}\Big). \tag{5.161}$$

It follows from (5.158) that there exists $c \in (0,\infty)$ satisfying that

$$\overline{c}M^{-1}n^{1-\rho} + \frac{\overline{c}rM^{-\frac{1}{2}}n^{1-\rho}}{\left(\frac{\overline{R}}{2}-2\overline{\delta}\right)_+} \leq cM^{-\frac{1}{2}}n^{1-\rho}. \tag{5.162}$$

Returning to (5.157), it follows from (5.161) and (5.162) that there exists $c \in (0,\infty)$ satisfying that

$$\begin{aligned}&\mathbb{P}\Big(\Big[f(\Theta_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \leq \frac{\overline{c}K}{\varepsilon^2\mathfrak{M}}\\ &+\left(\frac{\lambda\big(A\backslash V_{\overline{R}/2,\overline{\delta}}(x_0)\big)}{\lambda(A)} + \overline{c}\varepsilon^{-2}n^{-\rho} + 1 - \exp\Big(-\frac{c}{M}\Big) + cM^{-\frac{1}{2}}n^{1-\rho} + \frac{\overline{c}r}{(\frac{\overline{R}}{2}-2\overline{\delta})_+}\right)^K.\end{aligned} \tag{5.163}$$

Let $\eta \in (0,1]$. It follows from (5.158), (5.159) and an explicit computation that there exist $c_i \in (0,\infty)$, $i \in \{1,2,3,4\}$, and $\mathfrak{r}_1 \in (0,\mathfrak{r}]$ such that for $n(\varepsilon), M(\varepsilon), \mathfrak{M}(\varepsilon,\eta), K(\eta) \in \mathbb{N}$ satisfying that

$$n(\varepsilon) = c_1\varepsilon^{-2/\rho},\ M(\varepsilon) = c_2\varepsilon^{-4/\rho+4},\ \mathfrak{M}(\varepsilon,\eta) = c_3\varepsilon^{-2}\eta^{-1}\left|\log(\eta)\right|, \text{ and } K = c_4\left|\log(\eta)\right|, \tag{5.164}$$

it holds that

$$\frac{\overline{c}K}{\varepsilon^2\mathfrak{M}(\epsilon,\eta)} \leq \frac{\eta}{2}, \tag{5.165}$$

and for every $r \in (0, \mathfrak{r}_1]$ that
(5.166)

$$\left(\frac{\lambda\big(A\backslash V_{\overline{R}/2,\overline{\delta}}(x_0)\big)}{\lambda(A)} + \overline{c}\varepsilon^{-2}n(\varepsilon)^{-\rho} + 1 - \exp\Big(-\frac{c}{M(\varepsilon)}\Big) + cM(\varepsilon)^{-\frac{1}{2}}n(\varepsilon)^{1-\rho} + \frac{\overline{c}r}{(\frac{\overline{R}}{2} - 2\overline{\delta})_+}\right)^{K(\eta)} \le \frac{\eta}{2}.$$

Returning to (5.163), it follows for every $r \in (0, \mathfrak{r}_1]$ that

$$\mathbb{P}\Big(\Big[f(\Theta_{n(\varepsilon)}^{K(\eta),M(\varepsilon),\mathfrak{M}(\varepsilon,\eta),r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \ge \varepsilon\Big) \le \eta, \tag{5.167}$$

which completes the proof. □

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. We first observe the convergence of SGD in directions normal to the manifold.

The following proposition is an immediate consequence of Proposition 5.2 and the compactness of $\mathcal{M} \cap U$, where the essential difference in the compact case is that $R \in (0,\infty)$ can be chosen arbitrarily large. In particular, by compactness, for every $x_0 \in \mathcal{M} \cap U$ there exists $R_0 \in (0,\infty)$ such that for every $R_1, R_2 \in [R_0,\infty)$, $\delta \in (0,\infty)$ it holds that $V_{R_1,\delta}(x_0) = V_{R_2,\delta}(x_0)$. Furthermore, it follows from Remark 5.4 that the results apply to $\rho \in (0,1)$.

Proposition 6.1. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0,1,\dots,d-1\}$, $\rho \in (0,1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $(S,\mathcal{S})$ be a measurable space, let $F = (F(\theta,x))_{(\theta,x)\in\mathbb{R}^d\times S}\colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m}\colon \Omega \to S$, $n,m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta,X_{1,1})|^2\big] < \infty$, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta,X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d\colon [f(\theta) = \inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)]\big\}, \tag{6.1}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta,x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U\colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|F(\theta,X_{1,1})|^2 + |(\nabla_\theta F)(\theta,X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M}\cap U$ is a non-empty compact $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume for every $\theta \in \mathcal{M}\cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $M \in \mathbb{N}$, $r \in (0,\infty)$, $\theta \in \mathbb{R}^d$ let $\Theta_{0,\theta}^{M,r} \in \mathbb{R}^d\colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\theta_{0,\theta}^{M,r}(\omega) = \theta$, for every $n, M \in \mathbb{N}$, $r \in (0,\infty)$, $\theta \in \mathbb{R}^d$ let $\Theta_{n,\theta}^{M,r}\colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta_{n,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,\theta}^{M,r}, X_{n,m})\right], \tag{6.2}$$

for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0,\infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M}\cap U$ let $A_n \subseteq \Omega$ be the event satisfying that

$$A_n = \Big\{\omega \in \Omega\colon \Theta_{m,\theta}^{M,r}(\omega) \in V_{R,\delta}(x_0)\ \forall\, m \in \{0,\dots,n\}\Big\}, \tag{6.3}$$

and for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0,\infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M}\cap U$ let $\mathbf{1}_n : \Omega \to \{0,1\}$ denote the indicator function of A_n. Then for every $x_0 \in \mathcal{M}\cap U$ there exist $\delta_0, \mathfrak{r}, c \in (0,\infty)$ such that for every $R \in (0,\infty)$, $\delta \in (0,\delta_0]$, $r \in (0,\mathfrak{r}]$, $n, M \in \mathbb{N}$, $\theta \in V_{R,\delta}(x_0)$ it holds that

$$\mathbb{E}\left[\Big(\mathrm{d}(\Theta_{n,\theta}^{M,r}, \mathcal{M}\cap U) \wedge 1\Big)^2 \mathbf{1}_{n-1}\right]^{\frac{1}{2}} \le c n^{-\frac{\rho}{2}}. \tag{6.4}$$

Proof of Proposition 6.1. The proof is an immediate consequence of Proposition 5.2 and the compactness of $\mathcal{M} \cap U$. □

We will now obtain a lower bound in probability for the events A_m, $m \in \mathbb{N}$. It follows from Proposition 5.6 and the compactness of $\mathcal{M} \cap U$ that for every $x_0 \in \mathcal{M} \cap U$ there exist $\delta_0, \mathfrak{r}, c \in (0, \infty)$ such that the conclusion of Proposition 5.6 is satisfied for every $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, and $R \in (0, \infty)$ for this constant $c \in (0, \infty)$. That is, since for every $R_1, R_2 \in (0, \infty)$ sufficiently large we have $V_{R_1,\delta}(x_0) = V_{R_2,\delta}(x_0)$, it holds that the constant can be chosen independently of $R \in (0, \infty)$.

The proof of the following proposition is then an immediate consequence of Proposition 5.6, after using the fact that the constant $c \in (0, \infty)$ is independent of $R \in (0, \infty)$ and after passing to the limit $R \to \infty$. The improvement in the estimate, when compared to Proposition 5.6, is a result of the fact that SGD cannot leave the basin of attraction along the directions tangential to the manifold.

Proposition 6.2. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, $\rho \in (0, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m} \colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}, \tag{6.5}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U \colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta \in \mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a non-empty compact $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $M \in \mathbb{N}$, $r \in (0, \infty)$, $\theta \in \mathbb{R}^d$ let $\Theta_{0,\theta}^{M,r} \in \mathbb{R}^d \colon \Omega \to \mathbb{R}^d$ satisfy for every $\omega \in \Omega$ that $\theta_{0,\theta}^{M,r}(\omega) = \theta$, for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$, $\theta \in \mathbb{R}^d$ let $\Theta_{n,\theta}^{M,r} \colon \Omega \to \mathbb{R}^d$ satisfy that

$$\Theta_{n,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,\theta}^{M,r}, X_{n,m})\right], \tag{6.6}$$

and for every $n, M \in \mathbb{N}$, $r, R, \delta \in (0, \infty)$, $\theta \in \mathbb{R}^d$, $x_0 \in \mathcal{M} \cap U$ let $A_n \subseteq \Omega$ be the event satisfying that

$$A_n = \Big\{\omega \in \Omega \colon \Theta_{m,\theta}^{M,r}(\omega) \in V_{R,\delta}(x_0)\ \forall\, m \in \{0, \ldots, n\}\Big\}. \tag{6.7}$$

Then for every $x_0 \in \mathcal{M} \cap U$ there exist $\delta_0, \mathfrak{r}, c \in (0, \infty)$ such that for every $R \in (0, \infty)$, $\delta \in (0, \delta_0]$, $r \in (0, \mathfrak{r}]$, $n, M \in \mathbb{N}$, $\theta \in V_{R/2,\delta}(x_0)$ it holds that

$$\mathbb{P}[A_n] \geq \prod_{k=1}^{n} \Big(1 - \frac{c}{M k^{2\rho}}\Big)_+ - c M^{-1} n^{1-\rho}. \tag{6.8}$$

Proof of Proposition 6.2. The proof is an immediate consequence of Proposition 5.6 and the compactness of $\mathcal{M} \cap U$. □

The following theorem proves the convergence of SGD with initial data sampled from a uniform distribution on a bounded open set $A \subseteq \mathbb{R}^d$ satisfying $\mathcal{M} \cap U \cap A \neq \emptyset$. The proof is an immediate consequence of Theorem 5.7, Proposition 6.1, and Proposition 6.2.

Theorem 6.3. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, $\rho \in (0, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $\lambda : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ be the Lebesgue-Borel measure, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be a* *measurable function, let $X_{n,m}\colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}, \tag{6.9}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U\colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta \in \mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a compact $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M} \cap U \cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, *for every $M \in \mathbb{N}$, $r \in (0, \infty)$ let $\Theta_0^{M,r}\colon \Omega \to \mathbb{R}^d$ be continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m \in \mathbb{N}}$ are independent, and for every $M \in \mathbb{N}$, $r \in (0, \infty)$ let $\Theta_{n,\theta}^{M,r}\colon \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be random variables which satisfy that*

$$\Theta_{n,\theta}^{M,r} = \Theta_{n-1,\theta}^{M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1,\theta}^{M,r}, X_{n,m})\right]. \tag{6.10}$$

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

$$\begin{aligned} &\mathbb{P}\Big(\mathrm{d}\left(\Theta_{n,\theta}^{M,r}, \mathcal{M} \cap U\right) \geq \varepsilon\Big) \\ &\leq \frac{\lambda\big(A \backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + c\varepsilon^{-2} n^{-\rho} + 1 - \prod_{k=1}^{n} \Big(1 - \frac{c}{M k^{2\rho}}\Big)_+ + c M^{-1} n^{1-\rho}. \end{aligned} \tag{6.11}$$

Proof of Theorem 6.3. The proof is an immediate consequence of Theorem 5.7, Proposition 6.1, and Proposition 6.2. □

The following theorem estimates probability that $K \in \mathbb{N}$ independent solutions of SGD with initial data sampled from a uniform distribution on a compact set $A \subseteq \mathbb{R}^d$ satisfying that $\mathcal{M} \cap U \cap A$ is non-empty fail to converge to within distance $\varepsilon \in (0, 1]$ to the local manifold of minima at time $n \in \mathbb{N}$. The convergence is measured by minimizing a mini-batch average of the objective function. The proof is a consequence of Theorem 6.3 and the arguments leading from Theorem 5.7 to Theorem 5.11.

Theorem 6.4. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \ldots, d-1\}$, $\rho \in (0, 1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta, x))_{(\theta, x) \in \mathbb{R}^d \times S}\colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m}\colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d \colon [f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^d} f(\vartheta)]\big\}, \tag{6.12}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U\colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta \in \mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M} \cap U$ is a compact $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M} \cap U \cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, *for every $n \in \mathbb{N}_0$, $M \in \mathbb{N}$, $r \in (0, \infty)$ let $\Theta_n^{k,M,r}\colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $n, M \in \mathbb{N}$, $r \in (0, \infty)$ that $(\Theta_{n-1}^{k,M,r})_{k \in \mathbb{N}}$ and $(X_{n,k})_{k \in \mathbb{N}}$ are independent, assume for every $M \in \mathbb{N}$, $r \in (0, \infty)$ that $\Theta_0^{1,M,r}$*

is continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0,\infty)$ that $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, assume for every $n, M \in \mathbb{N}$, $r \in (0,\infty)$ that

$$\Theta_n^{1,M,r} = \Theta_{n-1}^{1,M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}^{1,M,r}, X_{n,m})\right], \tag{6.13}$$

and for every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $r \in (0,\infty)$ let $\mathit{\Theta}_n^{K,M,\mathfrak{M},r}\colon \Omega \to \mathbb{R}^d$ be a random variable which satisfies for every $\omega \in \Omega$ that

$$\mathit{\Theta}_n^{K,M,\mathfrak{M},r}(\omega) \in \left[\operatorname*{argmin}_{\theta\in\{\Theta_n^{k,M,r}(\omega)\colon k\in\{1,\dots,K\}\}}\left[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega))\right]\right]. \tag{6.14}$$

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

$$\begin{aligned} &\mathbb{P}\Big(\Big[f(\mathit{\Theta}_n^{K,M,\mathfrak{M},r}) - \inf\nolimits_{\theta\in\mathbb{R}^d} f(\theta)\Big] \geq \varepsilon\Big) \\ &\leq \frac{cK}{\varepsilon^2\mathfrak{M}} + \left(\frac{\lambda\big(A\backslash V_{R/2,\delta}(x_0)\big)}{\lambda(A)} + c\varepsilon^{-2}n^{-\rho} + 1 - \prod_{k=1}^{n}\Big(1 - \frac{c}{Mk^{2\rho}}\Big)_+ + cM^{-1}n^{1-\rho}\right)^K. \end{aligned} \tag{6.15}$$

Proof of Theorem 6.4. The proof is an immediate consequence of Theorem 6.3, Theorem 5.7, and Theorem 5.11. □

In the final proposition of this section, we prove that the computation efficiency of the SGD algorithm proposed in Theorem 6.4 is improved by the compactness of $\mathcal{M}\cap U$. The improvement is due to the fact that the mini-batch size $M \in \mathbb{N}$ can be chosen smaller in the compact case, since the mini-batch size no longer needs to account for the possibility that SGD leaves a basin of attraction along directions tangential to the local manifold of minima.

Corollary 6.5. *Let $d \in \mathbb{N}$, $\mathfrak{d} \in \{0, 1, \dots, d-1\}$, $\rho \in (0,1)$, let $U \subseteq \mathbb{R}^d$ be an open set, let $A \subseteq \mathbb{R}^d$ be a bounded open set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $F = (F(\theta,x))_{(\theta,x)\in\mathbb{R}^d\times S}\colon \mathbb{R}^d \times S \to \mathbb{R}$ be a measurable function, let $X_{n,m}\colon \Omega \to S$, $n, m \in \mathbb{N}$, be i.i.d. random variables which satisfy for every $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_{1,1})|^2\big] < \infty$, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for every $\theta \in \mathbb{R}^d$ that $f(\theta) = \mathbb{E}\big[F(\theta, X_{1,1})\big]$, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy that*

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d\colon [f(\theta) = \inf\nolimits_{\vartheta\in\mathbb{R}^d} f(\vartheta)]\big\}, \tag{6.16}$$

assume for every $x \in S$ that $\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}$ is a locally Lipschitz continuous function, assume that $f|_U\colon U \to \mathbb{R}$ is a three times continuously differentiable function, assume for every non-empty compact set $\mathfrak{C} \subseteq U$ that $\sup_{\theta\in\mathfrak{C}} \mathbb{E}\big[|F(\theta, X_{1,1})|^2 + |(\nabla_\theta F)(\theta, X_{1,1})|^2\big] < \infty$, assume that $\mathcal{M}\cap U$ is a compact $\mathfrak{d}$-dimensional C^1-submanifold of $\mathbb{R}^d$, assume that $\mathcal{M}\cap U\cap A \neq \emptyset$, assume for every $\theta \in \mathcal{M}\cap U$ that $\operatorname{rank}((\operatorname{Hess} f)(\theta)) = d - \mathfrak{d}$, for every $n \in \mathbb{N}_0$, $M \in \mathbb{N}$, $r \in (0,\infty)$ let $\Theta_n^{k,M,r}\colon \Omega \to \mathbb{R}^d$, $k \in \mathbb{N}$, be i.i.d. random variables, assume for every $n, M \in \mathbb{N}$, $r \in (0,\infty)$ that $(\Theta_{n-1}^{k,M,r})_{k\in\mathbb{N}}$ and $(X_{n,k})_{k\in\mathbb{N}}$ are independent, assume for every $M \in \mathbb{N}$, $r \in (0,\infty)$ that $\Theta_0^{1,M,r}$ is continuous uniformly distributed on A, assume for every $M \in \mathbb{N}$, $r \in (0,\infty)$ that $\Theta_0^{M,r}$ and $\big(X_{n,m}\big)_{n,m\in\mathbb{N}}$ are independent, assume for every $n, M \in \mathbb{N}$, $r \in (0,\infty)$ that

$$\Theta_n^{1,M,r} = \Theta_{n-1}^{1,M,r} - \frac{r}{n^\rho M}\left[\sum_{m=1}^{M} (\nabla_\theta F)(\Theta_{n-1}^{1,M,r}, X_{n,m})\right], \tag{6.17}$$

and for every $n, M, \mathfrak{M}, K \in \mathbb{N}$, $r \in (0,\infty)$ let $\Theta_n^{K,M,\mathfrak{M},r} \colon \Omega \to \mathbb{R}^d$ be a random variable which satisfies for every $\omega \in \Omega$ that

$$\Theta_n^{K,M,\mathfrak{M},r}(\omega) \in \left[\operatorname*{argmin}_{\theta \in \{\Theta_n^{k,M,r}(\omega) \colon k \in \{1,\ldots,K\}\}} \left[\sum_{m=1}^{\mathfrak{M}} F(\theta, X_{n+1,m}(\omega)) \right] \right]. \tag{6.18}$$

Then for every $x_0 \in \mathcal{M} \cap U \cap A$ 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 $n(\varepsilon), M(\varepsilon), K(\eta), \mathfrak{M}(\varepsilon,\eta) \in \mathbb{N}$ satisfying that

$$n(\varepsilon) = c_1 \varepsilon^{-2/\rho}, \;\; M(\varepsilon) = c_2 \varepsilon^{-2/\rho + 2}, \;\; \mathfrak{M}(\varepsilon,\eta) = c_3 \varepsilon^{-2} \eta^{-1} \left|\log(\eta)\right|, \;\; \textit{and} \;\; K = c_4 \left|\log(\eta)\right|, \tag{6.19}$$

it holds that

$$\mathbb{P}\Big(\Big[f(\Theta_{n(\varepsilon)}^{K(\eta),M(\varepsilon),\mathfrak{M}(\varepsilon,\eta),r}) - \inf\nolimits_{\theta \in \mathbb{R}^d} f(\theta) \Big] \geq \varepsilon \Big) \leq \eta. \tag{6.20}$$

Proof of Corollary 6.5. The proof is an immediate consequence of Theorem 6.4 and the proof of Corollary 5.12. □

7. Applications

In this section, we prove that the conditions of Theorem 1.1 are satisfied for some (simple) objective functions $f \colon \mathbb{R}^d \to \mathbb{R}$ of the type (1.33) that arise in the training of neural networks. We will consider the case of a four-parameter affine-linear network with a linear activation function and the case of a two-parameter network with the ReLU activation function. We will prove that the set of global minima are respectively a codimension 2 submanifold of the parameter space, and a codimension 1 submanifold. This implies, in particular, that the global minima are not locally unique, and that the established convergence results, such as those proven in [13, 24], do not apply.

7.1. A four-parameter network with a linear activation function.

In this section, we show that the conditions of Theorem 1.1 are satisfied by a four-parameter affine-linear network with a linear activation function.

Proposition 7.1. *Let $\varphi \in L^2([0,1])$ be finite, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_{n,m} \colon \Omega \to [0,1]$, $n, m \in \mathbb{N}$, be i.i.d. random variables that are continuous uniformly distributed on $[0,1]$, let $f \colon \mathbb{R}^4 \to \mathbb{R}$ be the function which satisfies for every $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \mathbb{R}^4$ that*

$$f(\theta) = \int_0^1 \left| \theta_3 \theta_1 x + \theta_3 \theta_2 + \theta_4 - \varphi(x) \right|^2 \mathrm{d}x, \tag{7.1}$$

and let $F \colon \mathbb{R}^4 \times [0,1] \to \mathbb{R}$ be the function that satisfies for every $\theta \in \mathbb{R}^4$, $x \in [0,1]$ that

$$F(\theta, x) = \left| \theta_3 \theta_1 x + \theta_3 \theta_2 + \theta_4 - \varphi(x) \right|^2. \tag{7.2}$$

Then the functions f, F and the random variables $X_{n,m}$, $n, m \in \mathbb{N}$, satisfy the conditions of Theorem 1.1.

Proof of Proposition 7.1. 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

$$\sup_{\theta \in \mathfrak{C}} \mathbb{E}\left[\left| F(\theta, X_{1,1}) \right|^2 + \left| \nabla_\theta F(\theta, X_{1,1}) \right|^2 \right] < \infty. \tag{7.3}$$

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

$$f(\theta) = \int_0^1 |\theta_4 - \varphi(x)|^2 \, \mathrm{d}x = f(\tilde{\theta}). \tag{7.4}$$

Therefore, it holds that

$$\inf_{\theta \in \mathbb{R}^4} f(\theta) = \inf_{\theta \in \{\theta_3 \neq 0\}} f(\theta). \tag{7.5}$$

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

$$\nabla f(\theta) = 2 \int_0^1 (\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x)) \begin{pmatrix} \theta_3 x \\ \theta_3 \\ \theta_1 x + \theta_2 \\ 1 \end{pmatrix} \mathrm{d}x = 0. \tag{7.6}$$

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

$$r_k = \int_0^1 x^k \varphi(x) \, \mathrm{d}x, \tag{7.7}$$

it follows that $\theta \in \mathbb{R}^4$ satisfies equation (7.6) if and only if it holds that

$$\begin{cases} \frac{1}{3}\theta_1\theta_3^2 + \frac{1}{2}\theta_2\theta_3^2 + \frac{1}{2}\theta_3\theta_4 - r_1\theta_3 = 0, \\ \frac{1}{2}\theta_1\theta_3^2 + \theta_2\theta_3^2 + \theta_3\theta_4 - r_0\theta_3 = 0, \\ \frac{1}{3}\theta_1^2\theta_3 + \frac{1}{2}\theta_1\theta_2\theta_3 + \frac{1}{2}\theta_1\theta_4 - r_1\theta_1 + \frac{1}{2}\theta_1\theta_2\theta_3 + \theta_2^2\theta_3 + \theta_2\theta_4 - r_0\theta_2 = 0, \\ \frac{1}{2}\theta_1\theta_3 + \theta_2\theta_3 + \theta_4 - r_0 = 0. \end{cases} \tag{7.8}$$

For $\theta \in \mathbb{R}^4$ satisfying that $\theta_3 \neq 0$, an explicit computation proves that θ satisfies system (7.8) if and only if it holds that

$$\theta_1\theta_3 = -6(r_0 - 2r_1) \quad \text{and} \quad \theta_4 = -\theta_2\theta_3 + 4r_0 - 6r_1. \tag{7.9}$$

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

$$U = \{\theta \in \mathbb{R}^4 \colon \theta_3 \neq 0\}, \tag{7.10}$$

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

$$\mathcal{M} = \{\theta \in \mathbb{R}^4 \colon f(\theta) = \inf_{\vartheta \in \mathbb{R}^4} f(\vartheta)\}, \tag{7.11}$$

we claim that

$$\mathcal{M} \cap U = \{\, \theta \in \mathbb{R}^4 \colon \theta \text{ satisfies (7.9) and } \theta_3 \neq 0 \,\}. \tag{7.12}$$

Let $\theta \in \mathbb{R}^4$ satisfy (7.9) and $\theta_3 \neq 0$. By contradiction, suppose that there exists $\theta_0 = (\theta_{1,0}, \theta_{2,0}, \theta_{3,0}, \theta_{4,0})$ satisfying $\theta_{3,0} \neq 0$ such that

$$f(\theta_0) < f(\theta). \tag{7.13}$$

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

$$\lim_{|(\theta_1, \theta_4)| \to \infty} f(\theta_1, \theta_{2,0}, \theta_{3,0}, \theta_4) = \infty, \tag{7.14}$$

the identical considerations leading to (7.9) prove that

$$(\theta_1, \theta_4) \in \mathbb{R}^2 \mapsto f(\theta_1, \theta_{2,0}, \theta_{3,0}, \theta_4), \tag{7.15}$$

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

$$\theta_1 = -\frac{6(r_0 - 2r_1)}{\theta_{3,0}} \;\text{ and }\; \theta_4 = -\theta_{2,0}\theta_{3,0} + 4r_0 + 6r_1. \tag{7.16}$$

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

$$\tilde{\theta}_0 = (-\frac{6(r_0 - 2r_1)}{\theta_{3,0}}, \theta_{2,0}, \theta_{3,0}, -\theta_{2,0}\theta_{3,0} + 4r_0 + 6r_1), \tag{7.17}$$

satisfies (7.9) and $\tilde{\theta}_{3,0} \neq 0$. Therefore, it holds that

$$f(\tilde{\theta}_0) < f(\theta_0), \tag{7.18}$$

which contradicts the fact that $\nabla f = 0$ on the connected set of $\theta \in \mathbb{R}^4$ satisfying (7.9) and $\theta_3 \neq 0$. This proves (7.12). It is immediate from (7.9) that $\mathcal{M} \cap U$ is a non-empty, 2-dimensional, C^1-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[2], it holds that

$$\begin{aligned}\nabla^2 f(\theta) =& 2\int_0^1 \begin{pmatrix} \theta_3^2 x^2 & \theta_3^2 x & \theta_1\theta_3 x^2 + \theta_2\theta_3 x & \theta_3 x \\ & \theta_3^2 & \theta_1\theta_3 x + \theta_2\theta_3 & \theta_3 \\ & & (\theta_1 x + \theta_2)^2 & \theta_1 x + \theta_2 \\ & & & 1 \end{pmatrix} \mathrm{d}x \\ =& \begin{pmatrix} \frac{2}{3}\theta_3^2 & \theta_3^2 & \frac{2}{3}\theta_1\theta_3 + \theta_2\theta_3 & \theta_3 \\ & 2\theta_3^2 & \theta_1\theta_3 + 2\theta_2\theta_3 & 2\theta_3 \\ & & \frac{2}{3}\theta_1^2 + 2\theta_1\theta_2 + 2\theta_2^2 & \theta_1 + 2\theta_2 \\ & & & 2 \end{pmatrix},\end{aligned} \tag{7.19}$$

where this equality relies upon the fact that, due to (7.6) and $\theta_3 \neq 0$ on $\mathcal{M} \cap U$, we have that

$$\int_0^1 (\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x))\,\mathrm{d}x = \int_0^1 (\theta_3\theta_1 x + \theta_3\theta_2 + \theta_4 - \varphi(x))x\,\mathrm{d}x = 0. \tag{7.20}$$

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

$$\operatorname{rank}((\operatorname{Hess} f)(\theta)) = 2 = \operatorname{codim}(\mathcal{M} \cap U). \tag{7.21}$$

This completes the proof. □

7.2. **A two parameter network with the ReLU activation function.**

In this section, we show that the conditions of Theorem 1.1 are satisfied by a two-parameter affine-linear network with the ReLU activation function.

Proposition 7.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_{n,m}\colon \Omega \to [0,1]$, $n, m \in \mathbb{N}$, be i.i.d. random variables that are continuous uniformly distributed on $[0,1]$, let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be the function which satisfies for every $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ that*

$$f(\theta) = \int_0^1 |\theta_2 \max(\theta_1 x, 0) - \sin(x)|^2 \,\mathrm{d}x, \tag{7.22}$$

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

and let $F\colon \mathbb{R}^2 \times [0,1] \to \mathbb{R}$ *be the function which satisfies for every* $\theta \in \mathbb{R}^2$*,* $x \in [0,1]$ *that*

$$F(\theta, x) = |\theta_2 \max(\theta_1 x, 0) - \sin(x)|^2 . \tag{7.23}$$

Then the functions f*,* F *and the random variables* $X_{n,m}$*,* $n, m \in \mathbb{N}$*, satisfy the conditions of Theorem* 1.1*.*

Proof of Proposition 7.2*.* 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 $\theta \in \mathbb{R}^2$ it holds that

$$f(\theta) = \mathbb{E}[F(\theta, X_{1,1})], \tag{7.24}$$

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

$$\sup_{\theta \in \mathfrak{C}} \mathbb{E}\left[|F(\theta, X_{1,1})|^2 + |\nabla_\theta F(\theta, X_{1,1})|^2\right] < \infty. \tag{7.25}$$

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

$$U = \{\theta \in \mathbb{R}^2 \colon \theta_1 > 0, \theta_2 > 0\}. \tag{7.26}$$

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

$$f(\theta) = \int_0^1 |\theta_1 \theta_2 x - \sin(x)|^2 \, \mathrm{d}x, \tag{7.27}$$

and for every $\theta \in U$ that

$$\nabla f(\theta) = 2 \int_0^1 (\theta_1 \theta_2 x - \sin(x)) \begin{pmatrix} \theta_2 x \\ \theta_1 x \end{pmatrix} \mathrm{d}x. \tag{7.28}$$

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

$$\theta_1 \theta_2 = 3 \int_0^1 x \sin(x) \, \mathrm{d}x = 3(\sin(1) - \cos(1)). \tag{7.29}$$

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

$$\mathcal{M} = \{\theta \in \mathbb{R}^2 \colon f(\theta) = \inf\nolimits_{\vartheta \in \mathbb{R}^4} f(\vartheta)\}. \tag{7.30}$$

We claim that

$$\mathcal{M} \cap U = \{\ \theta \in \mathbb{R}^2 \colon \theta \text{ satisfies (7.29)},\ \theta_1 > 0,\ \text{and } \theta_2 > 0\}. \tag{7.31}$$

Suppose that $\theta \in U$ satisfies (7.29). By contradiction, suppose that there exists $\theta_0 = (\theta_{1,0}, \theta_{2,0}) \in \{\theta_1 > 0, \theta_2 > 0\}$ such that

$$f(\theta_0) < f(\theta). \tag{7.32}$$

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

$$\lim_{\theta_2 \to \infty} f(\theta_{1,0}, \theta_2) = +\infty \ \text{ and } \ f(\theta_{1,0}, 0) > f(\theta_0). \tag{7.33}$$

The arguments leading from (7.27) to (7.29) prove that (7.33) is uniquely minimized when

$$\theta_2 = \frac{3}{\theta_{1,0}}(\sin(1) - \cos(1)). \tag{7.34}$$

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

$$\tilde{\theta}_0 = (\theta_{1,0}, \frac{3}{\theta_{1,0}}(\sin(1) - \cos(1))), \tag{7.35}$$

we have that $\tilde{\theta}_0 \in U$, that $\tilde{\theta}_0$ satisfies (7.29), and that

$$f(\tilde{\theta}_0) \leq f(\theta_0) < f(\theta). \tag{7.36}$$

This contradicts the fact that $\nabla f = 0$ on the connected set of $\theta \in U$ satisfying (7.29). This proves (7.31). Since it is clear that $\mathcal{M} \cap U$ is a non-empty, 1-dimensional, C^1-submanifold of $\mathbb{R}^2$, it remains only to establish the nondegeneracy assumption.

For every $\theta \in \mathcal{M} \cap U$ it holds that

$$\begin{aligned} \nabla^2 f(\theta) &= 2 \begin{pmatrix} \frac{1}{3}\theta_2^2 & \frac{2}{3}\theta_1\theta_2 - (\sin(1) - \cos(1)) \\ & \frac{1}{3}\theta_1^2 \end{pmatrix} \\ &= 2 \begin{pmatrix} \frac{1}{3}\theta_2^2 & \sin(1) - \cos(1) \\ & \dfrac{3(\sin(1) - \cos(1))^2}{\theta_2^2} \end{pmatrix}. \end{aligned} \tag{7.37}$$

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

$$\operatorname{rank}((\operatorname{Hess} f)(\theta)) = 1 = \operatorname{codim}(\mathcal{M} \cap U). \tag{7.38}$$

This completes the proof. □

References

[1] M. Anitescu. Degenerate Nonlinear Programming with a Quadratic Growth Condition. 10(4):1116–1135.

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

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

[4] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

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

[6] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag/Springer, Heidelberg, 2010.

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

[8] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. 60(2):223–311.

[9] L. Bottou and Y. LeCun. Large scale online learning. *In Thrun, Sebastian, Saul, Lawrence, and Schölkopf, Bernhard (eds.), Advances in Neural In- formation Processing Systems 16. MIT Press, Cambridge, MA*, 2004.

[10] 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.

[11] 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 (NIPS)*, pages 1–11, 2012.

[12] 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. *ICASSP 2013*, 2013.

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

[14] A. Dieuleveut, A. Durmus, and B. Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *preprint, hal-01565514*, 2017.

[15] R. L. Foote. Regularity of the distance function. *Proc. Amer. Math. Soc.*, 92(1):153–155, 1984.

[16] S. Ghadimi and G. Lan. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. 23(4):2341–2368.

[17] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization.

[18] A. Graves. Generating sequences with recurrent neural networks. *preprint, arXiv:1308.0850*, 2013.

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

[20] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. 69(2-3):169–192.

[21] 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, IEEE*, 29(6):82–97, 2012.

[22] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[23] 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(4):805–810, 2003.

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

[25] H. Karimi, J. Nutini, and M. Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-\LOjasiewicz Condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, pages 795–811. Springer-Verlag.

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

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] Y. LeCun, L. Bottou, G. Orr, and K Muller. Efficient backprop. *In Orr, G. and K., Muller (eds.), Neural Networks: Tricks of the trade. Springer*, pages 9–50, 1998.

[29] Y. Lei, T. Hu, G. Li, and K. Tang. Stochastic Gradient Descent for Nonconvex Learning without Bounded Gradient Assumptions.

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

[31] J. Liu, S.J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An Asynchronous Parallel Stochastic Coordinate Descent Algorithm.

[32] S. Lojasiewicz. A topological property of real analytic subsets. 117:87–89.

[33] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. 46(1):157–178.

[34] E. Mizutani and S. Dreyfus. An analysis on negative curvature induced by singularity in multi-layer neural-network learning. *Advances in Neural Information Processing Systems*, pages 1669–1677, 2010.

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

[36] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization.

[37] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. 19(4):1574–1609.

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

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

[40] L. Pillaud-Vivien, A. Rudi, and F. Bach. Exponential convergence of testing error for stochastic gradient methods. *preprint, hal-01662278*, 2017.

[41] B.T. Polyak. Gradient methods for minimizing functionals. page 12.

[42] N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks : the official journal of the International Neural Network Society*, 12(1):145–151, 1999.

[43] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *preprint, arXiv:1109.5647*, 2012.

[44] M. Rattray, D. Saad, and S. I. Amari. Natural gradient descent for on-line learning. *Physical Review Letters*, 81(24):5461–5464, 1998.

[45] S.J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic Variance Reduction for Nonconvex Optimization. In *International Conference on Machine Learning*, pages 314–323.

[46] H. Robbins and S. Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985.

[47] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv:1609.04747*, page 12 pages, 2016.

[48] T. Schaul, S. Zhang, and Y. LeCun. Generating sequences with recurrent neural networks. *preprint, arXiv:1206.1106*, 2012.

[49] 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 (ICML-13)*, pages 1139–1147, 2013.

[50] R.S. Sutton. Two problems with backpropagation and other steepest-descent learning procedures for networks. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society, Hillsdale, NJ: Erlbaum*, 1986.

[51] C. Tang and C. Monteleoni. On the convergence rate of stochastic gradient descent for strongly convex functions. *Regularization, optimization, kernels, and support vector machines, Chapman & Hall/CRC Mach. Learn. Pattern Recogn. Ser. CRC Press, Boca Raton, FL*, pages 159–175, 2015.

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

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

[54] W. Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *preprint, arXiv:1301.3584*, 2011.

[55] H. Zhang and W. Yin. Gradient methods for convex minimization: Better rates under weaker conditions.

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