Almost Sure Convergence of Stochastic Gradient Methods under Gradient Domination

Simon Weissmann*
Institute of Mathematics
University of Mannheim
68138 Mannheim, Germany
simon.weissmann@uni-mannheim.de

Sara Klein*
Institute of Mathematics
University of Mannheim
68138 Mannheim, Germany
sara.klein@uni-mannheim.de

Waïss Azizian
CNRS, Grenoble INP
Université Grenoble Alpes
38000 Grenoble, France
waiss.azizian@univ-grenoble-alpes.fr

Leif Döring
Institute of Mathematics
University of Mannheim
68138 Mannheim, Germany
leif.doering@uni-mannheim.de

Abstract

Stochastic gradient methods are among the most important algorithms in training machine learning problems. While classical assumptions such as strong convexity allow a simple analysis they are rarely satisfied in applications. In recent years, global and local gradient domination properties have shown to be a more realistic replacement of strong convexity. They were proved to hold in diverse settings such as (simple) policy gradient methods in reinforcement learning and training of deep neural networks with analytic activation functions. We prove almost sure convergence rates $f(X_n) - f^* \in o\big(n^{-\frac{1}{4\beta-1}+\epsilon}\big)$ of the last iterate for stochastic gradient descent (with and without momentum) under global and local β-gradient domination assumptions. The almost sure rates get arbitrarily close to recent rates in expectation. Finally, we demonstrate how to apply our results to the training task in both supervised and reinforcement learning.

1 Introduction

First-order methods to minimize an objective function f have played a central role in the success of machine learning. This is accompanied by a growing interest in convergence statements particularly for stochastic gradient methods in different settings. To ensure convergence to the global optimum some kind of convexity assumption on the objective function is required. Especially in machine learning problems the standard (strong) convexity assumption is nearly never fulfilled. However, it is well known that achieving convergence towards global optima is still possible under a weaker assumption, namely under the gradient domination property, often referred to as Polyak-Łojasiewicz (PL)-inequality [45]. Also in reinforcement learning, multiple results have shown that the objective function for policy gradient methods, under specific parametrizations, fulfills a weak type of gradient domination and therefore provably achieve convergence towards the global optimum [34, 33, 11, 21]. Improving the understanding of rates and optimal step size choices for stochastic first order methods is of significant interest for the machine learning and reinforcement learning community. Many classical results identify convergence rates for the expected error $\mathbb{E}[f(X_n) - f^*]$. In the present

*These authors contributed equally to this work.

17th European Workshop on Reinforcement Learning (EWRL 2024).

article, we focus on almost sure convergence rates for the error $f(X_n) - f^*$ in stochastic gradient schemes under weak gradient domination. The contribution of this work is as follows:

(i) Under global gradient domination with parameter β (on the entire function domain), we prove that the last iterate of stochastic gradient descent (SGD) and stochastic heavy ball (SHB) converge almost surely and in expectation towards the global optimum with rate arbitrarily close to $o(n^{-\frac{1}{4\beta-1}})$. The almost sure and expectation rates of convergence that we obtain depend on the gradient domination parameter β and are the same for both algorithms and convergence types. For SGD this rate is arbitrarily close to the tight upper bound known in expectation [12], while the almost sure convergence rate is new for the (weak) gradient domination assumption (see Theorem 4.1 and discussion afterwards). To the best of our knowledge for SHB this is the first convergence result towards global optima under (weak) gradient domination, for both almost sure convergence and convergence in expectation (see Theorem 4.2).

(ii) We consider the case where the gradient domination property holds only locally, either around stationary points or around global minima. We provide the first local convergence rates under these settings: we prove that SGD remains within the good local region with high probability and, conditioned on this event, we obtain converges rates almost surely and in expectation (see Theorem 5.1 and Theorem 5.2).

(iii) Our local setting covers generic classes of functions. In particular, we demonstrate that it encompasses the training task of deep neural networks with analytic activation functions in supervised learning. Our result illustrates that the iterates of SGD are likely to become trapped in areas of local minima when the step size is is small. We verify under mild conditions, that SGD converges to local minima with given convergence speed (see Corollary 6.1).

(iv) Finally, we derive novel convergence results for policy gradient methods in reinforcement learning. We show that local gradient domination holds around the global optimum for the softmax parametrization resulting in the first local convergence rate for stochastic policy gradient with arbitrary batch-size (see Corollary 7.2).

We summarize the contributions of this paper in Table 1. These findings are also illustrated in a numerical toy experiment in Appendix E, where we have implemented SGD and SHB for monomials with increasing degree.

Table 1: Summary of known and new results. Table presents convergence rates for tuned step size ($\epsilon > 0$ arbitrarily small). Dom.: gradient domination holds locally or globally; local*: additional assumption on γ_1 required and results holds only locally. a.s.: almost surely; $\mathbb{E}$: in expectation. Ref.: for some cited results minor adjustments are necessary.

β	**Step size**	**Rate**	**Dom.**	**Algo.**	**Conv.**	**Ref.**
$\frac{1}{2}$	$\Theta\left(n^{-1+\epsilon}\right)$	$o\left(n^{-1+\epsilon}\right)$	global	SGD	a.s.	4.1 (i); e.g. [27, Thm. 1]
					$\mathbb{E}$	4.1 (ii); e.g. [20, Thm. 3]
				SHB	a.s	4.2 (i); e.g. [27, Thm. 2]
					$\mathbb{E}$	4.2 (ii); e.g. [25, Thm. 4.3]
			local*	SGD	a.s.	5.1 (ii); 5.2 (ii)
					$\mathbb{E}$	5.1 (iii); 5.2 (iii); e.g. [35, Thm. 4]
$(\frac{1}{2}, 1]$	$\Theta\left(n^{-\frac{2\beta}{4\beta-1}}\right)$	$o\left(n^{-\frac{1}{4\beta-1}+\epsilon}\right)$	global	SGD	a.s.	4.1 (i)
					$\mathbb{E}$	4.1 (ii); e.g. [12, Cor. 1]
				SHB	a.s.	4.2 (i)
					$\mathbb{E}$	4.2 (ii)
			local*	SGD	a.s.	5.1 (ii); 5.2 (ii)
					$\mathbb{E}$	5.1 (iii); 5.2 (iii)

Related work. We give a brief overview of related literature. For a more detailed literature review and classification of our contributions we refer the reader to Appendix A.1.

Usually, convergence rates for stochastic gradient methods are analysed in expectation or with high-probability bounds. Though originally, motivated by [46], research commenced with the quest for

almost sure convergence rates for gradient methods. In recent years, almost sure convergence rates towards global optima have been derived under (strong) convexity [49, 27]. The authors in [49] also analyzed almost sure convergence rates for SHB but under the assumption of convexity and [27] studied SHB under (strong) convexity and in a non-convex setting. Under the global gradient domination property almost sure convergence to a critical point is provided in [7], though without a rate.

There has been a lot of effort to derive local convergence guarantees for stochastic first order methods. It was shown in [35] that, subject to certain assumptions on the objective function, the SGD scheme converges almost surely towards a local minimum. By assuming strong convexity within a neighbourhood of a local minimum, they further established a local convergence rate conditioned on the event of remaining in this neighbourhood. We extend ideas from the strongly convex setting to local gradient domination and derive local convergence rates, both in the almost sure and expectation sense. In our analysis we distinguish the cases where the local gradient domination property holds in a neighbourhood of a local minimum or in the neighbourhood of the global optimum respectively.

For the application in the training of DNNs, it is worth noting that local convergence of SGD has been analyzed under stronger variants of gradient domination [52, 1]. Due to the stronger form of gradient domination, specific sub-classes of DNNs need to be considered to verify these assumptions whereas our result is only constrained to analytic activation functions.

For the application in reinfordement learning, recent results showed that choosing the tabular softmax parametrization in policy gradient (PG) algorithms results in objective functions which fulfill a non-uniform gradient domination property [34, 55, 21]. While convergence of PG for exact gradients is well understood, convergence rates for stochastic PG are rare and mostly require very large batch sizes [9, 21, 10]. It is noteworthy that a similar local analysis for stochastic policy gradient under entropy regularization is presented in [10]. Their local result is also based on [35], but requires an increasing batch size sequence to obtain $O(\frac{1}{n})$-convergence towards the regularized optimum with high probability. In contrast, we consider both the unregularized and entropy regularized setting and observe that one can also achieve convergence arbitrarily close to $o(\frac{1}{n})$ without the need for an increasing batch size. Moreover, the local convergence occurs almost surely on an event with high probability.

2 Mathematical Background - Optimization under Gradient Domination

We consider the problem of solving the minimization problem of the form

$$\min_{x\in\mathbb{R}^d} f(x)\,, \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ denotes the objective function of interest. Throughout this paper we assume that the objective function is bounded from below by $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$ and satisfies the classical L-smoothness assumption (either locally or globally):

Assumption 2.1. The objective function $f:\mathbb{R}^d\to\mathbb{R}$ is differentiable and the gradient ∇f is

(i) globally L-Lipschitz continuous, i.e. there exists $L>0$ such that $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$ for all $x,y\in\mathbb{R}^d$.

(ii) locally L-Lipschitz continuous, i.e. for all $R>0$ there exists $L(R)>0$ such that $\|\nabla f(x)-\nabla f(y)\|\le L(R)\|x-y\|$ for all $x,y\in\mathbb{R}^d$ with $|x|,|y|\le R$.

Using (global) L-smoothness, the descent lemma provides the inequality

$$f(y)\le f(x)+\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|y-x\|^2 \tag{2}$$

which is a fundamental instrument to analyze first order optimization methods. As a motivation recall the iterative update generated by gradient descent with constant step size $\gamma\le\frac{1}{L}$, i.e.

$$x_{n+1}=x_n-\gamma\nabla f(x_n),\quad x_1\in\mathbb{R}^d\,.$$

Applying (2) and the iteration scheme yields the iterative descent property

$$[f(x_{n+1})-f^*]\le[f(x_n)-f^*]-\frac{\gamma}{2}\|\nabla f(x_n)\|^2\,.$$

Under further strong convexity assumption it is classical to show that the gradient descent algorithm converges to a global minimum at a linear rate. In order to derive a convergence rate without assuming convexity of f one can use dominating relations of the gradient $\nabla f(x)$ with respect to the optimality gap $f(x) - f^*$. In particular, as demonstrated in [19] it is nowadays well-known that gradient descent converges linearly under the PL-condition which assumes that there exists $c > 0$ such that for all $x \in \mathbb{R}^d$ there holds

$$\|\nabla f(x)\| \geq c(f(x) - f^*)^\beta \tag{3}$$

with exponent $\beta = 1/2$ [45]. It is worthwhile to emphasize that (3) is weaker than strong convexity, the classical textbook assumption that fails for many applications. Under the PL-condition the iterative descent property can be written as a recursion:

$$[f(x_{n+1}) - f^*] \leq \left(1 - \frac{\gamma c}{2}\right)[f(x_n) - f^*] \ .$$

In fact, there are many works analyzing (stochastic) first order methods under the weaker Łojasiewicz condition formulated in (local) areas around stationary points x^* and exponents $\beta \in [1/2, 1]$ [22, 12, 48, 51]. For general $\beta \in [1/2, 1]$ the recursive descent property reads as

$$[f(x_{n+1}) - f^*] \leq [f(x_n) - f^*] - \frac{\gamma c}{2}[f(x_n) - f^*]^{2\beta}$$

leading to sub-linear convergence for $\beta > 1/2$.

For the purpose of our analysis of stochastic gradient methods, we collect the following types of global and local gradient domination properties.

Definition 2.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable with $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$.

1. We say that f satisfies the global gradient domination property with parameter $\beta \in [\frac{1}{2}, 1]$ if there exists $c > 0$ such that for all $x \in \mathbb{R}^d$ it holds true that
$$\|\nabla f(x)\| \geq c(f(x) - f^*)^\beta.$$

2. Let $x^* \in \mathbb{R}^d$ be a stationary point, i.e. $\nabla f(x^*) = 0$. We say that f satisfies a local gradient domination property in x^* with parameter $\beta_{x^*} \in [\frac{1}{2}, 1]$ if there exist a radius $r_{x^*} > 0$ and a constant $c_{x^*} > 0$ such that
$$\|\nabla f(x)\| \geq c_{x^*}|f(x) - f(x^*)|^{\beta_{x^*}}$$
for all $x \in \mathcal{B}_{r_{x^*}}(x^*) = \{y \in \mathbb{R}^d : \|x^* - y\| \leq r_{x^*}\}$. We say that f satisfies a local gradient domination property in f^* with parameter $\beta \in [\frac{1}{2}, 1]$ if there exist a radius $r > 0$ and a constant $c > 0$ such that
$$\|\nabla f(x)\| \geq c(f(x) - f^*)^\beta$$
for all $x \in \mathcal{B}^*_r = \{y \in \mathbb{R}^d : f(y) - f^* \leq r\}$.

Remark 2.3. If $\beta = \frac{1}{2}$, we will call the gradient domination *strong* since it is implied by strong convexity. In contrast, we call the gradient domination *weak* for $\beta \in (\frac{1}{2}, 1]$. Moreover, note that for the local gradient domination property in x^* the parameters r and c may depend on x^*. Furthermore, we emphasize that for the definition of the local gradient domination in f^* we do not require the existence of $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$.

The PL-condition mentioned above is a special case of the general global gradient domination property for $\beta = \frac{1}{2}$. In [30] it has been demonstrated that all analytic functions satisfy the local gradient domination property, emphasizing the particular significance of the local case. Further, it has been proved that all overparametrized neural networks fulfill the local gradient domination property [26]. See also [31, 8, 15] and references therein for the application of (strong) gradient domination to (deep) neural networks. In [12, 2, 4, 59] examples of functions are discussed that fulfill the (weak) gradient domination property. For instance, one-dimensional monomials $f(x) = |x|^p$, $p \geq 2$, satisfy the weak global gradient domination property with $\beta = \frac{p-1}{p}$. We refer to [12, Appendix A] for a longer list of globally gradient dominated functions including convex and non-convex functions. Notably, in reinforcement learning it is known that the tabular softmax parametrization leads to a parametrized value function that satisfies the so-called "non-uniform" PL-inequality [34, 33, 21]. In Section 7 we will show how this non-uniform gradient domination implies local gradient domination for f^*. This renders our local analysis of stochastic gradient methods specifically applicable in RL. As mentioned earlier, since every analytic function already satisfies local gradient domination, we expect that the local analysis can encompass further parametrizations, such as neural networks.

Assumptions on the Stochastic First Order Oracle. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an underlying probability space. In general, we assume that we can access the exact gradient $\nabla f(x)$ through a stochastic first order oracle $V : \mathbb{R}^d \times M \to \mathbb{R}^d$ defined by

$$V(x, m) = \nabla f(x) + Z(x, m), \quad x \in \mathbb{R}^d,\ m \in M\,, \tag{4}$$

where $(M, \mathcal{M})$ is a measurable space, $Z : \mathbb{R}^d \times M \to \mathbb{R}^d$ is a state dependent $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{M}/\mathcal{B}(\mathbb{R}^d)$-measurable mapping describing the error to the exact gradient ∇f. The stochastic gradient evaluation is then modelled through $V(x, \zeta)$, where the random variable $\zeta : \Omega \to M$ is independent of the state $x \in \mathbb{R}^d$. We make the following unbiasedness and second moment assumption:

Assumption 2.4. We assume that for each $x \in \mathbb{R}^d$ it holds that

$$\mathbb{E}[Z(x, \zeta)] := \int_\Omega Z(x, \zeta(\omega))\mathrm{d}\mathbb{P}(\omega) = 0$$

and there exist non-negative constants A, B and C such that for all $x \in \mathbb{R}^d$ it holds that

$$\mathbb{E}[\|V(x, \zeta)\|^2] \le A(f(x) - f^*) + B\|\nabla f(x)\|^2 + C\,. \tag{ABC}$$

It is worth noting that the (ABC) assumption is a generalization of the bounded variance assumption that appears for $A = B = 0$. It was introduced by [20] as expected smoothness condition and shown to be the weakest assumption among many others.

Stochastic Gradient Methods. The following two classical optimization algorithms will be analyzed in this article. Both algorithms are described as discrete time stochastic process $(X_n)_{n\in\mathbb{N}}$ driven by noisy gradient evaluations in (4). In each iteration, we assume that the stochastic first order oracle is accessed through the evaluation of ζ_{n+1} which is a copy of ζ independent from the current state X_n.

The stochastic gradient descent (SGD) scheme is given by the stochastic update

$$X_{n+1} = X_n - \gamma_n\, V(X_n, \zeta_{n+1})\,,$$

where X_1 is a $\mathbb{R}^d$-valued random vector which denotes the initial state. To keep the notation simple, we will introduce $V_{n+1}(X_n) := V(X_n, \zeta_{n+1})$ suppressing the explicit noise representation through $(\zeta_n)_{n\in\mathbb{N}}$ in the following. The iterative update formula then reads as

$$X_{n+1} = X_n - \gamma_n\, V_{n+1}(X_n). \tag{SGD}$$

The iterative scheme of stochastic heavy ball (SHB) is defined by

$$X_{n+1} = X_n - \gamma_n V_{n+1}(X_n) + \nu(X_n - X_{n-1})\,, \tag{SHB}$$

with initial $\mathbb{R}^d$-valued random vector X_1. The additional summand is called the momentum term with momentum parameter $\nu \in [0, 1)$. In both cases, $(\gamma_n)_{n\in\mathbb{N}}$ denotes a sequence of positive step sizes and we denote by $(\mathcal{F}_n)_{n\in\mathbb{N}}$ the natural filtration induced by the process $(X_n)_{n\in\mathbb{N}}$. Note that we adopt the convention where the set of natural numbers $\mathbb{N}$ refers to the positive integers excluding zero. When necessary, we explicitly define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

Under smoothness of the function f one can use the global gradient domination property in Definition 2.2 together with the (ABC) to derive an iterative inequality of the form

$$\begin{aligned}&\mathbb{E}[f(X_{n+1}) - f^* \mid \mathcal{F}_n]\\ &\le \Big(1 + \frac{LA\gamma_n^2}{2}\Big)(f(X_n) - f^*) - \Big(\gamma_n - \frac{BL\gamma_n^2}{2}\Big)c^2(f(X_n) - f^*)^{2\beta} + \frac{LC\gamma_n^2}{2}\,,\end{aligned} \tag{5}$$

where (X_n) is generated by (SGD). We provide more details in Appendix A.3. Typically, the expectation is taken on both sides of the inequality to derive a convergence rate in expectation by working with recursive inequalities. In this article we push the argument further. We combine smoothness and gradient domination with a variant of the Robbins-Siegmund Theorem (see Lemma A.2) to derive almost sure convergence rates.

3 Preliminary Discussion on Super-Martingale Convergence Rates

In the previous section, we have sketched how to combine the global gradient domination property with smoothness to derive a recursive inequality of the form

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 + c_1\gamma_n)Y_n - c_2\gamma_n Y_n^{2\beta} + c_3\gamma_n^2\,,$$

where $Y_n := f(X_n) - f^*$. For analysing these inequalities, we must deal separately with the strong gradient domination case ($\beta = \frac{1}{2}$) and the weak gradient domination case ($\beta > \frac{1}{2}$) to avoid divisions by zero. For the former case the recursive inequality simplifies, whereas a more complex analysis is required for the latter. To establish almost sure convergence rates we employ convergence lemmas for super-martingales based on the Robbins-Sigmund Theorem. This methodology has been introduced in [49] and further utilized in [27] to analyze SGD and SHB under (strong) convexity. In the following, we illustrate how to extend the arguments to convergence under the global gradient domination property. Here is our super-martingale result that also encompasses [27, Lemma 1] when $\beta = \frac{1}{2}$ for completeness:

Lemma 3.1. *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of non-negative random variables on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with natural filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and suppose there exists $\beta \in [\frac{1}{2}, 1]$, $c_1, c_3 \geq 0$ and $c_2 > 0$ such that*

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 + c_1\gamma_n^2)Y_n - c_2\gamma_n Y_n^{2\beta} + c_3\gamma_n^2\,,$$

for all $n \geq 1$, where $\gamma_n = \Theta(\frac{1}{n^\theta})$ for some fixed $\theta \in \left(\frac{1}{2}, 1\right)$. Then, for any

$$\eta \in \begin{cases} \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2 - 2\theta, 1) & : \beta = \frac{1}{2} \end{cases},$$

$(Y_n)_{n\in\mathbb{N}}$ vanishes almost surely with $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$.

4 Convergence for Global Gradient Domination Property

In this part of the paper, we provide our global convergence result in a non-convex and globally smooth setting. Combining the recursive inequality (5) with the super-martingale convergence result from Lemma 3.1 leads to the following theorem. The detailed proof is given in Appendix C.

Theorem 4.1. *Suppose Assumption 2.1 (i) and Assumption 2.4 are fulfilled and let f satisfy the global gradient domination property from Definition 2.2 with $\beta \in [\frac{1}{2}, 1]$. Denote by $(X_n)_{n\in\mathbb{N}}$ the sequence generated by* (SGD) *using a step size $\gamma_n = \Theta(\frac{1}{n^\theta})$ with $\theta \in \left(\frac{1}{2}, 1\right)$. For any*

$$\eta \in \begin{cases} \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2 - 2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

$$(i)\;\; f(X_n) - f^* \in o\left(\frac{1}{n^{1-\eta}}\right),\ a.s., \quad \text{and} \quad (ii)\;\; \mathbb{E}[f(X_n) - f^*] \in o\left(\frac{1}{n^{1-\eta}}\right).$$

To the best of our knowledge our theorem presents the first convergence rate for SGD under weak gradient domination with respect to almost sure convergence.

It is natural to ask which θ leads to the best convergence rate. Optimising for η yields an optimal choice $\theta = \frac{2\beta}{4\beta-1}$ to achieve the best possible rate of convergence. This specific choice yields a lower bound of the interval given by $2 - 2\theta = \frac{\theta+2\beta-2}{2\beta-1} = 1 - \frac{1}{4\beta-1}$ and therefore an almost sure convergence of the form $o(\frac{1}{n^p})$ where p is arbitrarily close to $\frac{1}{4\beta-1}$ (see also Table 1). We emphasize that the rate we obtain is arbitrarily close to the one obtained in [14, 12] in expectation and is tight according to [12, Prop. 2].

Roughly speaking, our result guarantees a faster convergence rate for "stronger" gradient domination properties (i.e. for smaller β). Indeed, as $2 - 2\theta > \frac{\theta+2\beta-2}{2\beta-1}$ for β sufficiently close to $\frac{1}{2}$ our result is

consistent to the one presented in [27, Thm. 1] by replacing the μ-strongly convex assumption with the strong gradient domination property with $\beta = \frac{1}{2}$.

Similar arguments can be used to derive almost sure convergence rates for SHB under global gradient domination:

Theorem 4.2. *Suppose Assumption 2.1 (i) and Assumption 2.4 are fulfilled and let f satisfy the global gradient domination property from Definition 2.2 with $\beta \in [\frac{1}{2}, 1]$. Denote by $(X_n)_{n\in\mathbb{N}}$ the sequence generated by* (SHB) *using a step size $\gamma_n = \Theta(\frac{1}{n^\theta})$ for $\theta \in (\frac{1}{2}, 1)$. For any*

$$\eta \in \begin{cases} \left(\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2-2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

$$(i)\ \ f(X_n) - f^* \in o\left(\frac{1}{n^{1-\eta}}\right),\ a.s., \quad \text{and} \quad (ii)\ \ \mathbb{E}[f(X_n) - f^*] \in o\left(\frac{1}{n^{1-\eta}}\right).$$

To the best of our knowledge, our result gives the first convergence proof of SHB to global optima under weak gradient domination, with rates for almost sure convergence and convergence of expectations. The resulting convergence rate using the optimized step size are summarized in Table 1. In the strong gradient domination setting our rate in expectation gets arbitrarily close to the $O(\frac{1}{n})$ convergence obtained in [25]. It is noteworthy that the utilization of SHB in our analysis does not yield a superior convergence rate compared to SGD. This arises from the proof technique and aligns with the findings in [27, 49] where the authors similarly achieve no acceleration. In general, for deterministic settings acceleration of gradient methods can achieve improvements of convergence rates [51]. Though in the special case of gradient domination with $\beta = \frac{1}{2}$, HB as well as Nesterov cannot accelerate in the deterministic setting as shown in [56].

5 Convergence for Local Gradient Domination Property

In this section, we want to generalize this analysis [35] under local strong convexity to the weaker local gradient domination property for different cases of β. We consider the two cases of local gradient domination separately. The contributions and differences of our results under less restricted assumptions are the following:
First, we show in both cases that SGD remains in the gradient dominated region with high probability by only assuming local gradient domination instead of local strong convexity. Especially in the case of a local minimum x^* this is a challenging task, as we have to ensure that the SGD scheme $(X_n)_{n\in\mathbb{N}}$ remains close to x^* without exploiting convexity. We can guarantee this whenever x^* is in an isolated connected compact set of local minima $\mathcal{X}^*$. We prove convergence towards the level set of $\mathcal{X}^*$ and obtain Theorem 5.1. Second, additionally to convergence in expectation we prove almost sure convergence conditioned on the "good event". Third, due to the weaker gradient domination assumption and no convexity, one cannot expect the convergence of X_n to (local or global) minimum x^*, instead we focus on convergence of $f(X_n)$ to $f(x^*)$. In [29] they delve into the rationale behind considering this as a more robust metric.

The main result under local gradient domination in a local minimum x^* is as follows:

Theorem 5.1. *Fix some tolerance level $\delta > 0$ and let $\mathcal{X}^* \subset \mathbb{R}^d$ be an isolated compact connected set of local minima with level $l = f(x^*)$ for all $x^* \in \mathcal{X}^*$. Suppose that f satisfy the local gradient domination property in each $x^* \in \mathcal{X}^*$, f is locally G-Lipschitz continuous and satisfies Assumption 2.1 (ii). Moreover, suppose Assumption 2.4 hold true. Denote by $(X_n)_{n\in\mathbb{N}}$ the sequence generated by* (SGD) *using a step size $\gamma_n = \Theta(\frac{1}{n^\theta})$ for $\theta \in (\frac{1}{2}, 1)$ and suppose that $\gamma_n \leq \gamma_1$ for γ_1 small enough (dependent on δ). Then, the following holds:*

(i) There exist subsets $\mathcal{U}$ and $\mathcal{U}_1$ of $\mathbb{R}^d$ such that, if $X_1 \in \mathcal{U}_1$ the event $\Omega_{\mathcal{U}} = \{X_n \in \mathcal{U}$ for all $n = 1, 2, \dots\}$ has probability at least $1 - \delta$.

Moreover, there exists $\beta \in [\frac{1}{2}, 1]$ such that for any

$$\eta \in \begin{cases} \left(\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2-2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

$$(ii)\ \ |f(X_n) - l|\mathbf{1}_{\Omega_{\mathcal{U}}} \in o\left(\frac{1}{n^{1-\eta}}\right),\ a.s., \quad and \quad (iii)\ \ \mathbb{E}[|f(X_n) - l|\mathbf{1}_{\Omega_{\mathcal{U}}}] \in o\left(\frac{1}{n^{1-\eta}}\right).$$

The proof and more details on the sets $\mathcal{U}$ and $\mathcal{U}_1$ as well as on β can be found in appendix F.

The main result concerning local gradient domination in f^* is presented below and does not necessitate the existence of a local minimum or any stationary point. It is worth noting that the definition of local gradient domination in f^* guarantees the gradient domination property for any x with $f(x)$ close to f^*. Consequently, this definition ensures that functions satisfying this property cannot possess local minima or saddle points within this region.

Theorem 5.2. *Fix some tolerance level $\delta > 0$. Suppose f satisfies the local gradient domination property in f^* from Definition 2.2 with $\beta \in [\frac{1}{2}, 1]$ and $\mathcal{B}_r^* \subseteq \mathbb{R}^d$. Moreover, suppose within $\mathcal{B}_r^*$ f is G-Lipschitz continuous, Assumption 2.1 (i) and Assumption 2.4 hold true. Denote by $(X_n)_{n\in\mathbb{N}}$ the sequence generated by* (SGD) *using a step size $\gamma_n = \Theta(\frac{1}{n^\theta})$ for $\theta \in (\frac{1}{2}, 1)$ and suppose that $\gamma_n \leq \gamma_1$ for γ_1 small enough (dependent on δ). Then, the following holds:*

(i) There exist subsets $\mathcal{U}$ and $\mathcal{U}_1$ of $\mathbb{R}^d$ such that, if $X_1 \in \mathcal{U}_1$ the event $\Omega_{\mathcal{U}} = \{X_n \in \mathcal{U}$ for all $n = 1, 2, \dots\}$ has probability at least $1 - \delta$.

Moreover, for any

$$\eta \in \begin{cases} \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2 - 2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

$$(ii)\ \ (f(X_n) - f^*)\mathbf{1}_{\Omega_{\mathcal{U}}} \in o\left(\frac{1}{n^{1-\eta}}\right),\ a.s., \quad and \quad (iii)\ \ \mathbb{E}[(f(X_n) - f^*)\mathbf{1}_{\Omega_{\mathcal{U}}}] \in o\left(\frac{1}{n^{1-\eta}}\right).$$

6 Application in the training of neural networks

In supervised learning one aims to approximate an unknown model $\varphi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$ by a parametrized function $g_w : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$ with parameter $w \in \mathbb{R}^{d_w}$. Given a family of training data $((Z^{(m)}, Y^{(m)}))_{m\in\mathbb{N}}$ generated as i.i.d. samples from an unknown distribution $\mu_{(Z,Y)}$ one usually chooses the parameter $w \in \mathbb{R}^{d_w}$ by solving

$$\min_{w\in\mathbb{R}^{d_w}} \ \mathbb{E}_{\mu_{(Z,Y)}}[\Phi(g_w(Z), Y)],$$

where $\Phi : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}_+$ is a user specific data discrepancy. One popular choice of parametrizations are DNNs. We define a neural network of depth $L \in \mathbb{N}$ by the recursion

$$z_0 := z, \quad z_\ell = \sigma^{\otimes d_\ell}(A_\ell z_{\ell-1} + b_\ell),\ \ell = 1, \dots, L-1, \quad g_w(z) := A_L z_{L-1} + b_L\,.$$

The weights $((A_\ell, b_\ell))_{\ell=1}^L$ of the DNN are collected in $w \in \mathcal{W} := \times_{\ell=1}^L(\mathbb{R}^{d_\ell \times d_{\ell-1}} \times \mathbb{R}^{d_\ell}) \simeq \mathbb{R}^{d_w}$, and $\sigma^{\otimes d} : \mathbb{R}^d \to \mathbb{R}^d$ describes the component-wise application of the activation function $\sigma : \mathbb{R} \to \mathbb{R}$.

Provided that σ and Φ are analytic, and (Z, Y) are compactly supported $\mathbb{R}^{d_z} \times \mathbb{R}^{d_y}$-valued random variables, then $f^{\mathrm{DNN}} : \mathbb{R}^{d_w} \to \mathbb{R}_+$ defined by $w \mapsto \mathbb{E}_{\mu_{(Z,Y)}}[\Phi(g_w(Z), Y)]$ is analytic [8, Thm. 5.2] and therefore satisfies local gradient domination in any stationary point w_* (see Definition 2.2) [30].

In our notation, the stochastic first order oracle takes the form

$$V(w, (Z, Y)) = \nabla_w f^{\mathrm{DNN}}(w) + (\nabla_w \Phi(g_w(Z), Y) - \nabla_w f^{\mathrm{DNN}}(w))\,,$$

where we denote $\zeta = (Z, Y)$ and the iterative SGD then reads as

$$W_{n+1} = W_n - \gamma_n \nabla_w \Phi(g_{W_n}(Z_{n+1}), Y_{n+1})$$

with $\zeta_n = (Z_n, Y_n)$ independent and identical distributed. The iterative scheme of SHB can be written similarly. Note that this scenario also includes the empirical risk minimization of $\frac{1}{M}\sum_{m=1}^M \Phi(g_w(z^{(m)}), y^{(m)})$ when $\zeta = (Z, Y) \sim \frac{1}{M}\sum_{m=1}^M \delta_{(z^{(m)}, y^{(m)})}$, see Appendix A.2 for more details. The following local convergence is a direct consequence of Theorem 5.1.

Corollary 6.1. *Let $\delta > 0$. Denote by $(W_n)_{n\in\mathbb{N}}$ the sequence generated by SGD with $w \mapsto \nabla_w f^{\mathrm{DNN}}(w)$ as objective function, step size $\gamma_n \in \Theta(n^{-\theta})$ for $\theta \in (\frac{1}{2}, 1)$, and assume that f^{DNN} is analytic. Let $\mathcal{W}^*$ be an isolated compact set of local minima with level $l = f^{\mathrm{DNN}}(w^*)$ for all $w^* \in \mathcal{W}^*$ and suppose Assumption 2.4 is satisfied within $\mathcal{W}^*$. Suppose that $\gamma_n \leq \gamma_1$ for sufficiently small γ_1 (depending on δ), then there exist two subsets $\mathcal{U}, \mathcal{U}_1$ of $\mathbb{R}^{d_w}$ such that $W_1 \in \mathcal{U}_1$ implies that the event $\Omega_{\mathcal{U}} = \{W_n \in \mathcal{U}, \text{ for all } n \geq 1\}$ has probability at least $1-\delta$. Moreover, there exists $\beta \in [\frac{1}{2}, 1]$ such that for any*

$$\eta \in \begin{cases} \left(\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2-2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that $|f^{\mathrm{DNN}}(W_n) - l|\mathbf{1}_\Omega \in o\big(n^{\eta-1}\big)$ almost surely and in expectation.

In words: If the iterates of SGD reach a certain area around a local minimum, they are likely to become trapped in that region with high probability, provided that the step size is sufficiently small. This results shows that, under very general conditions, SGD converges to local minima and furthermore quantifies the convergence speed.

Remark 6.2. One may similarly apply Theorem 5.2 in the training of DNNs to derive convergence towards a global minimum with high probability provided that the initial loss $f^{\mathrm{DNN}}(X_1)$ and initial step size γ_1 are sufficiently small.

7 Application in Reinforcement Learning

Let $(\mathcal{S}, \mathcal{A}, \rho, r, p)$ be a discounted MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$ and discount factor $\rho \in [0, 1)$. Further, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ is the positive expected reward function and $p(s'|s, a)$ describes the transition probability from state s to s' under action a. As in [34] we assume that the rewards are bounded in $[0, 1]$. Consider the stationary tabular softmax policy for parameter $w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, i.e.

$$\pi_w(a|s) = \frac{\exp(w(s,a))}{\sum_{a'\in\mathcal{A}_s} \exp(w(s,a'))}, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

In the following we consider entropy regularized PG jointly with vanilla PG by setting $\lambda = 0$ when no regularisation is considered. Then, for an initial state distribution μ, the value function under the softmax parametrization is given by

$$V_\lambda^{\pi_w}(\mu) = \mathbb{E}_\mu^{\pi_w}\Big[\sum_{t=0}^{\infty} \rho^t r(S_t, A_t)\Big] - \lambda \mathbb{E}_\mu^{\pi_w}\Big[\sum_{t=0}^{\infty} \rho^t \log(\pi_w(A_t|S_t)\Big]$$

and denote by $V_\lambda^*(\mu)$ the global optimum and by π^* the optimal policy. If $\lambda > 0$ there exists a continuum of optimal parameters w^*, such that $V_\lambda^{\pi_{w^*}}(\mu) = V_\lambda^*(\mu)$. If $\lambda = 0$ no such parameters exists.

In order to maximize the objective we use stochastic gradient ascent. For $\lambda = 0$ this is called stochastic policy gradient method or REINFORCE. Note that e.g., [57] provide a stochastic first order oracle which meets the conditions required in Assumption 2.4. For $\lambda > 0$, a stochastic gradient estimator that satisfies Assumption 2.4 with $A = B = 0$ is presented in [10, Eq. (4)]. In both cases a non-uniform PL-inequality holds [34, Lem. 8, Lem. 15]: For every $w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ it holds that

$$\|\nabla_w V_\lambda^{\pi_w}(\mu)\|_2 \geq c_\lambda(w)^x \left[V_\lambda^*(\mu) - V_\lambda^{\pi_w}(\mu)\right]^x,$$

with $x = 1$ if $\lambda = 0$ and $x = \frac{1}{2}$ if $\lambda > 0$ and

$$c_\lambda(w) = \begin{cases} \frac{\min_{s\in\mathcal{S}} \pi_w(a^*(s)|s)}{\sqrt{|\mathcal{S}|}(1-\rho)} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}, & \lambda = 0, \\ \frac{2\lambda}{|\mathcal{S}|(1-\rho)} \min_s \mu(s) \min_{s,a} \pi_w(a|s)^2 \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}, & \lambda > 0. \end{cases} \tag{6}$$

Here $a^*(s)$ denotes the best possible action in state s. W.l.o.g. we assume that $a^*(s)$ is unique, otherwise one can simply consider the maximum over all possible best actions, i.e. replace $\min_{s\in\mathcal{S}} \pi_w(a^*(s)|s)$ with $\min_{s\in\mathcal{S}} \min_{a \text{ is optimal action in } s} \pi_w(a|s)$. We prove that this implies a local gradient domination property with $\beta = 1$ (weak PL) for $\lambda = 0$ and $\beta = \frac{1}{2}$ (strong PL) for $\lambda > 0$.

Proposition 7.1. *There exists $r, c > 0$ such that for all $w \in \mathcal{B}^*_{r,\lambda} = \{w : V^*_\lambda(\mu) - V^{\pi_w}_\lambda(\mu) \leq r\}$ it holds that $c_\lambda(w) \geq c$.*

As the objective function $w \mapsto V^{\pi_w}_\lambda(\mu)$ is smooth and Lipschitz on $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ ([55, Lem. E.1] for $\lambda = 0$ and [10] for $\lambda > 0$), all assumptions in Theorem 5.2 are satisfied and we obtain the following result.

Corollary 7.2. *Let $\delta > 0$. Denote by $(W_n)_{n\in\mathbb{N}}$ the sequence generated by SGD with $w \mapsto -V^{\pi_w}_\lambda(\mu)$ as objective function, step size $\gamma_n \in \Theta(n^{-\theta})$ for $\theta \in (\frac{1}{2}, 1)$ and suppose $\gamma_n \leq \gamma_1$ for sufficiently small γ_1 (depending on δ). Then, there exist two subsets $\mathcal{U}, \mathcal{U}_1$ of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that $W_1 \in \mathcal{U}_1$ implies that the event $\Omega_{\mathcal{U}} = \{W_n \in \mathcal{U}, \text{ for all } n \geq 1\}$ has probability at least $1 - \delta$. Moreover, for any*

$$\eta \in \begin{cases} (\max\{2-2\theta, \theta\}, 1), & \text{if } \lambda = 0, \\ (2-2\theta, 1), & \text{if } \lambda > 0 \end{cases}$$

it holds that $(V^*_\lambda(\mu) - V^{\pi_{W_n}}_\lambda(\mu))\mathbf{1}_\Omega \in o\big(n^{\eta-1}\big)$ almost surely and in expectation.

In words: If the (regularized) stochastic policy gradient algorithm is started close enough to the optimum a nearly $o(n^{-\frac{1}{3}})$ ($o(n^{-1})$ respectively) almost sure rate of convergence can be obtained by choosing $\theta = \frac{2}{3}$ (θ close to 1 respectively) This is in contrast to $o(n^{-1})$ (linear convergence) known in (regularized) policy gradient with access to exact gradients.

Remark 7.3. Note that r and c in Proposition 7.1 can be explicitly chosen (see Remark G.1). Hence, one can choose the neighbourhoods $\mathcal{U}$ and $\mathcal{U}_1$ w.r.t. r as in (41) and Lemma F.10 in Appendix F to find an explicit neighbourhood $\mathcal{U}_1$ as condition for initialization.

References

[1] Jing An and Jianfeng Lu. Convergence of stochastic gradient descent under a local lojasiewicz condition for deep neural networks. *arXiv Preprint*, arXiv:2304.09221, 2024.

[2] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[3] Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of SGD in non-convex over-parametrizedlearning. *arXiv Preprint*, arXiv:1811.02564, 2018.

[4] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[5] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[6] Sourav Chatterjee. Convergence of gradient descent for deep neural networks, 2022.

[7] Emilie Chouzenoux, Jean-Baptiste Fest, and Audrey Repetti. A Kurdyka-Lojasiewicz property for stochastic optimization algorithms in a non-convex setting. *arXiv Preprint*, arXiv:2302.06447, 2023.

[8] Steffen Dereich and Sebastian Kassing. Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes. *arXiv Preprint*, arXiv:2102.09385, 2021.

[9] Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv Preprint*, arXiv:2110.10117, 2022.

[10] Yuhao Ding, Junzi Zhang, and Javad Lavaei. Local analysis of entropy-regularized stochastic soft-max policy gradient methods. In *2023 European Control Conference (ECC)*, pages 1–8. IEEE, 2023.

[11] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9827–9869. PMLR, 23–29 Jul 2023.

[12] Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global Kurdyka-Lojasiewicz inequality. In *Advances in Neural Information Processing Systems*, volume 35, pages 15836–15848. Curran Associates, Inc., 2022.

[13] Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21(136):1–48, 2020.

[14] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1965–2058. PMLR, 15–19 Aug 2021.

[15] Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. In *Advances in Neural Information Processing Systems*, volume 34, pages 7937–7949. Curran Associates, Inc., 2021.

[16] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461 – 529, 2018.

[17] Benjamin Gess and Sebastian Kassing. Convergence rates for momentum stochastic gradient descent with noise of machine learning type. *arXiv Preprint*, arXiv:2302.03550, 2023.

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

[19] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing.

[20] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. Survey Certification.

[21] Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. *arXiv Preprint*, arXiv:2310.02671, 2023.

[22] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257. PMLR, 23–26 Jun 2016.

[23] Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4394–4400, 2020.

[24] Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[25] Yuqing Liang, Jinlan Liu, and Dongpo Xu. Stochastic momentum methods for non-convex learning without bounded assumptions. *Neural Networks*, 165:830–845, 2023.

[26] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[27] Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2963–2983. PMLR, 02–05 Jul 2022.

[28] Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21884–21914. PMLR, 23–29 Jul 2023.

[29] Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. *arXiv Preprint*, arXiv:2312.08531, 2023.

[30] Stanislaw Lojasiewicz. Ensembles semi-analytiques. *Lectures Notes IHES (Bures-sur-Yvette)*, 1965.

[31] Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv Preprint*, arXiv:2006.05610, 2021.

[32] Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6630–6639. PMLR, 13–18 Jul 2020.

[33] Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7555–7564. PMLR, 18–24 Jul 2021.

[34] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6820–6829. PMLR, 13–18 Jul 2020.

[35] Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 1117–1128. Curran Associates, Inc., 2020.

[36] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[37] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[38] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Springer New York, NY, 2004.

[39] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $\mathrm{O}(k^2)$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[40] Lam M. Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takác, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20:176:1–176:49, 2018.

[41] Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv Preprint*, arXiv:2302.05437, 2023.

[42] Antonio Orvieto, Jonas Kohler, and Aurelien Lucchi. The role of memory in stochastic optimization. In *Uncertainty in Artificial Intelligence*, pages 356–366. PMLR, 2020.

[43] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[44] Boris T Polyak. *Introduction to Optimization*. Translations series in mathematics and engineering. Optimization Software, Publications Division, 1987.

[45] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

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

[47] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[48] Kevin Scaman, Cedric Malherbe, and Ludovic Dos Santos. Convergence rates of non-convex stochastic gradient descent under a generic Lojasiewicz condition and local smoothness. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19310–19327. PMLR, 17–23 Jul 2022.

[49] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3935–3971. PMLR, 15–19 Aug 2021.

[50] Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems*, volume 34, pages 18866–18877. Curran Associates, Inc., 2021.

[51] Ashia C Wilson, Lester Mackey, and Andre Wibisono. Accelerating rescaled gradient descent: Fast optimization of smooth functions. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[52] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part i: Discrete time analysis. *Journal of Nonlinear Science*, 33(3):45, 2023.

[53] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2955–2961. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[54] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv Preprint*, arXiv:1604.03257, 2016.

[55] Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.

[56] Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing polyak-Łojasiewicz functions. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2948–2968. PMLR, 12–15 Jul 2023.

[57] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

[58] Beitong Zhou, Jun Liu, Weigao Sun, Ruijuan Chen, Claire Tomlin, and Ye Yuan. pbsgd: Powered stochastic gradient descent methods for accelerated non-convex optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3258–3266. International Joint Conferences on Artificial Intelligence Organization, 7 2020.

[59] Yi Zhou, Zhe Wang, and Yingbin Liang. Convergence of cubic regularization for nonconvex optimization under KL property. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

A Supplementary background

A.1 Literature Review and Classification of our Contribution

The roots of stochastic gradient methods trace back to Robbins & Monro [47]. Since then, various variants of SGD have been established as fundamental algorithms for optimizing complex models in the realm of machine learning. We refer to [5] for a detailed overview.

We start the review with the literature deriving convergence rates in expectation for SGD. Under the assumptions of smoothness and (strong) convexity [44, 36, 40, 50, 28] studied convergence rates towards global optima. Moreover, many articles additionally analyze the non-convex case and prove convergence rates for the gradient norm towards zero [18, 24, 28, 41].

Notably, several other results regarding convergence of SGD towards global optima have been established under the gradient domination setting [19]. The authors in [3] demonstrate exponential convergence rates in expectation in the overparameterized setting under strong gradient domination. See also [31] and [26], where convergence rates of order $O(\frac{1}{n})$ are shown for neural networks using the (strong) gradient domination property. High-probability bounds on the approximation error are provided in [48] under a generalized gradient domination property, the so-called Separable-Łojasiewicz assumption, fulfilled by smooth neural networks. In [23] also strong gradient domination is assumed, where the smoothness assumption is weakened through α-Hölder continuity, achieving a rate of $O(\frac{1}{n^\alpha})$ in expectation. The (ABC) condition is introduced in [20] and where $O(\frac{1}{n})$ convergence is shown under strong gradient domination. Furthermore, the authors in [12] and [14] consider generalizations of gradient domination that include our definition as a special case. They derive convergence rates in expectation which we encompass with our result and extend to almost sure convergence (see also the discussion behind Theorem 4.1).

All the results mentioned so far consider convergence in expectation or high-probability bounds, although originally, motivated by [46], research commenced with the quest for almost sure convergence rates for gradient methods. In recent years, the authors in [49] and, building upon it, the authors in [27] derive almost sure convergence rates towards global optima under strong convexity. In [49] almost sure convergence rates for SHB are also analysed but under the assumption of convexity and in [27] SHB is studied under (strong) convexity and in a non-convex setting. Returning the attention back to SGD with respect to gradient domination also some almost sure convergence results have been established. As an extension to the PL-type gradient domination, in [7] the so-called KL property is assumed, which contains gradient domination as a special case. The authors demonstrate almost sure convergence to a critical point, though without a rate. To conclude, to the best of our knowledge the derived almost sure convergence rate under gradient domination in Theorem 4.1 is novel.

Next, we want to provide further insights to the literature regarding SHB. In the realm of momentum methods, Polyak's Heavy-Ball Method (HBM) [43] and Nesterov's accelerated gradient method [39] stand out as a foundational contribution. The authors of [16] provide a detailed description of the stochastic formulation of HBM and establish almost sure convergence but without giving a rate. In [54, 42, 53, 32, 58] convergence rates in expectation are shown in (strongly) convex and non-convex settings, where the non-convex analysis covers convergence of the norm of the gradient. Convergence of momentum methods under the strong gradient domination property and linear convergence due to an overparametrized machine learning setting is shown in [17]. In [25] the authors determine $O(\frac{1}{n})$ convergence rate for SHB under strong gradient domination. Our main result for SHB presented in Theorem 4.2 describes almost sure convergence and convergence in expectation under global gradient domination. Both result are quantified with a given rate of convergence.

Finally, we aim to differentiate the present article from existing results on the convergence of SGD under the assumption of local gradient domination. In [8] almost sure convergence of SGD to a stationary point under the local gradient domination (for x^*) is demonstrated, provided that the process (X_n) remains local, albeit without a rate. A local analysis of SGD towards minima without any gradient domination assumption is presented in [13]. Instead, a rank assumption is imposed on the Hessian, and mini-batches, along with resampling, are leveraged to ensure convergence to the global optimum with high probability. The resulting rate does not converge to zero and requires an increasing batch size. Finally, under the global Lipschitz assumption on the objective, it can be verified that SGD almost surely converges to a stationary point as demonstrated in [35]. In the same work the authors derive a local convergence analysis under local strong convexity. Our analysis in

Section 5 builds upon [35] and generalizes their results to the local gradient domination property. For the application in the training of DNNs, it is worth noting that local convergence of SGD has been analyzed under stronger variants of gradient domination [52, 1]. Due to the stronger form of gradient domination, specific sub-classes of DNNs need to be considered to verify these assumptions whereas our result is only constrained to analytic activation functions. Under the machine learning noise conditions in [52], convergence toward zero loss with high probability is shown, provided that the initial loss is sufficiently small. In contrast, the authors in [1] demonstrate convergence towards zero loss under initialization in a local (strong) Łojasiewicz region. Indeed, one can construct DNNs satisfying the latter condition [6].

A.2 Example: Expected risk minimization

In order to give more insights into the considered setting of our stochastic first order oracle we formulate a specific one based on expected risk minimization. In expected risk minimization we are interested in minimizing an objective function of the form

$$f(x) = \mathbb{E}[F(x,\zeta)] = \int_\Omega F(x,\zeta(\omega))\, d\mathbb{P}(\omega)$$

where $F : \mathbb{R}^d \times M \to \mathbb{R}$ is $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{M}/\mathcal{B}(\mathbb{R})$-measurable. In our notation the stochastic first order oracle then takes the form

$$V(x,\zeta) = \nabla f(x) + (\nabla_x F(x,\zeta) - \nabla f(x)) = \nabla_x F(x,\zeta)$$

and the iterative update of SGD reads as

$$X_{n+1} = X_n - \gamma_n \nabla_x F(X_n, \zeta_{n+1})$$

with a sequence of independent and identically distributed (ζ_n). The iterative scheme of SHB can be written in similar way. Note that this scenario also includes empirical risk minimization where the objective function takes a finite sum form, with $\zeta \sim \mathcal{U}(\{1,\dots,N\})$,

$$f(x) = \frac{1}{N}\sum_{i=1}^N F(x,i) = \mathbb{E}[F(x,\zeta)]\,.$$

A.3 Auxiliary steps for convergence of SGD

Exemplifying the SGD method, we now illustrate the typical steps of the convergence analysis for first-order optimization methods. First, the smoothness of the function f is exploited by applying the descent inequality (2) to the iteration scheme and then applying conditional expectations,

$$\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] \le f(X_n) - \gamma_n \|\nabla f(X_n)\|^2 + \frac{L\gamma_n^2}{2}\mathbb{E}[\|V_{n+1}(X_n)\|^2 \mid \mathcal{F}_n].$$

Next, f^* is subtracted on both sides and the variance term of the stochastic gradient is controlled through the (ABC) condition:

$$\begin{aligned}&\mathbb{E}[f(X_{n+1}) - f^* \mid \mathcal{F}_n]\\ &\le \Big(1 + \frac{LA\gamma_n^2}{2}\Big)(f(X_n) - f^*) - \Big(\gamma_n - \frac{BL\gamma_n^2}{2}\Big)\|\nabla f(X_n)\|^2 + \frac{LC\gamma_n^2}{2}.\end{aligned} \tag{7}$$

Without further assumptions this inequality can now be used to show that the gradient $\nabla f(X_n)$ converges almost surely to zero. In order to obtain convergence towards a global optimum additional assumptions are needed. For instance, it is sufficient to incorporate the global gradient domination property defined in Definition 2.2 which yields an iterative inequality of the form

$$\begin{aligned}&\mathbb{E}[f(X_{n+1}) - f^* \mid \mathcal{F}_n]\\ &\le \Big(1 + \frac{LA\gamma_n^2}{2}\Big)(f(X_n) - f^*) - \Big(\gamma_n - \frac{BL\gamma_n^2}{2}\Big)c^2(f(X_n) - f^*)^{2\beta} + \frac{LC\gamma_n^2}{2}.\end{aligned} \tag{8}$$

A.4 Auxiliary Convergence Theorems

In the following section, we provide two specific convergence theorems used to prove almost sure convergence (Lemma A.2) as well as convergence in expectation (Lemma A.3). The former one is a direct consequence of the well-known Robbins-Siegmund theorem, provided here for completeness.

Theorem A.1 (Theorem 1 in [46]). *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space, $(Z_n)_{n\in\mathbb{N}}$, $(A_n)_{n\in\mathbb{N}}$, $(B_n)_{n\in\mathbb{N}}$ and $(C_n)_{n\in\mathbb{N}}$ be non-negative and adapted stochastic processes with*

$$\sum_{n=1}^{\infty} A_n < \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} B_n < \infty$$

almost surely. Suppose that for each $n \in \mathbb{N}$ the recursion

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \leq (1 + A_n)Z_n + B_n - C_n$$

is satisfied, then (i) there exists an almost surely finite random variable Z_∞ such that $Z_n \to Z_\infty$ almost surely as $n \to \infty$ and (ii) $\sum_{n=1}^{\infty} C_n < \infty$ almost surely.

Lemma A.2. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space, $(Y_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ and $(r_n)_{n\in\mathbb{N}}$ be non-negative and adapted stochastic processes with*

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} b_n < \infty \quad \textit{and} \quad r_n > 0$$

almost surely. Suppose that for each $n \in \mathbb{N}$ the recursion

$$\mathbb{E}[r_{n+1}Y_{n+1} \mid \mathcal{F}_n] \leq (1 - a_n)r_nY_n + b_n$$

is satisfied, then we have $r_nY_n \to 0$ almost surely as $n \to \infty$.

Proof. We define $Z_n := r_nY_n$, $B_n := b_n$ and $C_n := a_nr_nY_n$ such that

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \leq Z_n - C_n + B_n$$

for $n \in \mathbb{N}$. Using Theorem A.1 we observe that there exists Z_∞ almost surely finite such that $Z_n = r_nY_n \to Z_\infty$ almost surely as $n \to \infty$. Moreover, we obtain that

$$\sum_{n=1}^{\infty} C_n = \sum_{n=1}^{\infty} a_nr_nY_n < \infty$$

almost surely, which yields that

$$\liminf_{n\to\infty} r_nY_n = 0$$

almost surely, since $\sum_{n=1}^{\infty} a_n = \infty$ almost surely. Since limit inferior and limit coincide for converging sequences, the assertion follows:

$$Z_\infty = \lim_{n\to\infty} r_nY_n = \liminf_{n\to\infty} r_nY_n = 0$$

almost surely. □

The following Lemma will be applied to prove convergence in expectation.

Lemma A.3. *Let $(w_n)_{n\in\mathbb{N}}$ be a non-negative sequence, such that $w_{n+1} \leq (1 - a_n)w_n + b_n$, where $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are non-negative sequences satisfying*

$$\sum_{n=1}^{\infty} a_n = \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} b_n < \infty.$$

Then, $\lim_{n\to\infty} w_n = 0$.

Proof. W.l.o.g we assume that $w_{n+1} = (1 - a_n)w_n + b_n$, otherwise we could just increase a_n or decrease b_n which would have no effect on the summation tests. We obtain

$$-w_1 \leq w_n - w_1 = \sum_{k=1}^{n-1}(w_{k+1} - w_k) = \sum_{k=1}^{n-1} b_k - \sum_{k=1}^{n-1} w_ka_k.$$

Since $w_n - w_1$ is bounded below and $\sum_{k=1}^{\infty} b_k < \infty$, we deduce that $\sum_{k=1}^{n} w_ka_k$ is bounded. Since all summands are positive, the infinite sum converges. Thus, as a difference of two converging series also $(w_n)_{n\in\mathbb{N}}$ converges. Finally, the convergence of $\sum_{k=1}^{\infty} w_ka_k$ implies $\liminf_{n\to\infty} w_n = 0$ which, by the convergence of $(w_n)_{n\in\mathbb{N}}$, implies $\lim_n w_n = \liminf_n w_n = 0$. □

B Proof of Lemma 3.1

Proof. In the following, we treat both cases $\beta = \frac{1}{2}$ and $\beta \in (\frac{1}{2}, 1]$ separately.
$\underline{\beta = \frac{1}{2}}$: In this case, the inequality reduces to

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 + c_1\gamma_n^2 - c_2\gamma_n)Y_n + c_3\gamma_n^2.$$

By the choice of γ_n, there exists some $N > 0$ and $\tilde{c}_1 > 0$ such that $c_2\gamma_n - c_1\gamma_n^2 \geq \tilde{c}_1\gamma_n$ for all $n \geq N$. Hence, for all $n \geq N$

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \tilde{c}_1\gamma_n)Y_n + c_3\gamma_n^2$$

such that the claim follows by [27, Lem. 1].

$\underline{\beta \in (\frac{1}{2}, 1]}$: The proof uses the elementary inequality

$$(n+1)^{1-\eta} \leq n^{1-\eta} + (1-\eta)n^{-\eta}, \tag{9}$$

which was also applied and proved in [27, Lem. 1]. The aim is to apply the Robbins-Siegmund corollary, Lemma A.2, in order to derive the almost sure convergence rate. Let $1 \leq q < 2$ be arbitrary for now. The key step of the proof is the following computation

$$\begin{aligned}
\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] &\leq (1 + c_1\gamma_n^2)Y_n - c_2\gamma_n Y_n^{2\beta} + c_3\gamma_n^2 \\
&= (1 + c_1\gamma_n^2)Y_n - c_2\gamma_n^q Y_n + c_2\gamma_n^q Y_n - c_2\gamma_n Y_n^{2\beta} + c_3\gamma_n^2 \\
&= (1 + c_1\gamma_n^2 - c_2\gamma_n^q)Y_n + c_2\gamma_n\left(\gamma_n^{q-1}Y_n - Y_n^{2\beta}\right) + c_3\gamma_n^2.
\end{aligned} \tag{10}$$

Similar to the case $\beta = \frac{1}{2}$ there exists some $N > 0$ and $\tilde{c}_1 > 0$ such that $c_2\gamma_n^q - c_1\gamma_n^2 \geq \tilde{c}_1\gamma_n^q$ for all $n \geq N$. Hence, for all $n \geq N$ we obtain the iterative inequality of the form

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \tilde{c}_1\gamma_n^q)Y_n + c_2\gamma_n\left(\gamma_n^{q-1}Y_n - Y_n^{2\beta}\right) + c_3\gamma_n^2. \tag{11}$$

The function $x \mapsto ax - bx^{2\beta}$ takes it maximum at $\bar{x} = \left(\frac{a}{2b\beta}\right)^{\frac{1}{2\beta-1}}$ such that

$$\begin{aligned}
\gamma_n(\gamma_n^{q-1}Y_n - Y_n^{2\beta}) &\leq \frac{\gamma_n^{q+\frac{q-1}{2\beta-1}}}{(2\beta)^{\frac{1}{2\beta-1}}} - \frac{\gamma_n^{1+\frac{(q-1)2\beta}{2\beta-1}}}{(2\beta)^{\frac{2\beta}{2\beta-1}}} \\
&= \frac{1}{(2\beta)^{\frac{1}{2\beta-1}}}\gamma_n^{\frac{2q\beta-1}{2\beta-1}} - \frac{1}{(2\beta)^{\frac{2\beta}{2\beta-1}}}\gamma_n^{\frac{2q\beta-1}{2\beta-1}} \\
&= (2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta})\gamma_n^{\frac{2q\beta-1}{2\beta-1}}
\end{aligned} \tag{12}$$

holds almost surely. We define $\tilde{c}_2 = c_2(2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta}) \in (0, \infty)$ for $\beta \in (\frac{1}{2}, 1)$ and proceed with

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \tilde{c}_1\gamma_n^q)Y_n + \tilde{c}_2\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + c_3\gamma_n^2. \tag{13}$$

Next, we apply the elementary inequality (9) and choose q such that $\frac{1}{2} < \theta \leq \frac{1}{q} \leq 1$. Moreover by the choice of γ_n, there exists some $c_4 > 0$ such that $\tilde{c}_1\gamma_n^q \geq \frac{c_4}{t^{q\theta}}$ for all $n \geq N$. It follows that for all $n \geq N$

$$\begin{aligned}
&\mathbb{E}[(n+1)^{1-\eta}Y_{n+1} \mid \mathcal{F}_n] \\
&\leq (n+1)^{1-\eta}(1 - \tilde{c}_1\gamma_n^q)Y_n + (n+1)^{1-\eta}\tilde{c}_2\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}c_3\gamma_n^2 \\
&\leq (n^{1-\eta} + (1-\eta)n^{-\eta})(1 - \frac{c_4}{n^{q\theta}})Y_n + (n+1)^{1-\eta}\tilde{c}_2\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}c_3\gamma_n^2 \\
&= \left(1 + \frac{1-\eta}{n} - \frac{c_4}{n^{q\theta}} - \frac{c_4(1-\eta)}{n^{q\theta+1}}\right)n^{1-\eta}Y_n + (n+1)^{1-\eta}\tilde{c}_2\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}c_3\gamma_n^2.
\end{aligned}$$

We set $\tilde{c}_3 = \max\{\tilde{c}_2, c_3\}$ such that for all $n \geq N$

$$\begin{aligned}&\mathbb{E}[(n+1)^{1-\eta} Y_{n+1} \mid \mathcal{F}_n] \\ &\leq \left(1 + \frac{1-\eta}{n} - \frac{c_4}{n^{q\theta}} - \frac{c_4(1-\eta)}{n^{q\theta+1}}\right) n^{1-\eta} Y_n + \tilde{c}_3 (n+1)^{1-\eta} (\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2).\end{aligned}$$

Observe that $q\theta \leq 1$ by condition $\theta \leq \frac{1}{q}$. Hence, there exists $\tilde{c}_4 > 0$ and $\tilde{N} > N$for sufficiently large $\tilde{N} \geq N$ such that for all $n \geq \tilde{N}$ we have

$$\mathbb{E}[(n+1)^{1-\eta} Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \tilde{c}_4 \frac{1}{n^{q\theta}}) n^{1-\eta} Y_n + c_3 (n+1)^{1-\eta} (\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2) \tag{14}$$

In order to apply Robbins-Siegmund, more precisely Lemma A.2, we are going to verify the following three sufficient conditions:

$$\sum_{n=\tilde{N}}^{\infty} \frac{1}{n^{q\theta}} = \infty, \tag{15}$$

$$\sum_{n=\tilde{N}}^{\infty} n^{1-\eta-2\theta} < \infty, \tag{16}$$

$$\sum_{n=\tilde{N}}^{\infty} n^{1-\eta-\frac{\theta(2\beta q-1)}{2\beta-1}} < \infty. \tag{17}$$

Then, $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$ almost surely.

The first condition (15) is obviously satisfied, since we assume $\theta \leq \frac{1}{q}$. For the second condition (16) we may choose $\theta > 1 - \frac{\eta}{2}$ such that $1 - \eta - 2\theta < -1$. The third condition (17) gives $1 - \eta - \frac{\theta(2\beta q-1)}{2\beta-1} < -1$ which leads to the condition $\theta > \frac{(2-\eta)(2\beta-1)}{2\beta q-1}$. Hence, all together we obtain the sufficient condition

$$\theta \in \left(\max\left\{\frac{(2-\eta)(2\beta-1)}{2\beta q-1}, 1 - \frac{\eta}{2}\right\}, \frac{1}{q}\right].$$

In the following, we consider the two cases separately that correspond to the maximum being either $1 - \frac{\eta}{2}$ or $\frac{(2-\eta)(2\beta-1)}{2\beta q-1}$. The first case occurs precisely for $\frac{1}{q} \leq \frac{2\beta}{4\beta-1}$, the latter one for $\frac{1}{q} \geq \frac{2\beta}{4\beta-1}$.
Firstly, let $\frac{1}{q} \leq \frac{2\beta}{4\beta-1}$. In this situation the sufficient condition on θ simplifies to

$$\theta \in \left(1 - \frac{\eta}{2}, \frac{1}{q}\right].$$

The interval is non-empty for $\frac{1}{q} > \frac{2-\eta}{2}$, which requires $\eta \in (\frac{4\beta-2}{4\beta-1}, 1)$.
Secondly, let $\frac{1}{q} \geq \frac{2\beta}{4\beta-1}$. In this situation the sufficient condition on θ simplifies to

$$\theta \in \left(\frac{(2-\eta)(2\beta-1)}{2\beta q-1}, \frac{1}{q}\right],$$

the interval is non-empty for $\frac{1}{q} < 2\beta\eta - 2\beta + 2 - \eta$. Hence, $\frac{1}{q} \in (\frac{2\beta}{4\beta-1}, 2\beta\eta - 2\beta + 2 - \eta)$ which requires the condition $\eta \in (\frac{4\beta-2}{4\beta-1}, 1)$.

Either case yields sufficient conditions on θ and η (depending on the auxiliary variable q) under which $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$ holds almost surely. We will now utilize the free variable q to prove the claim.

- Let $\theta \in (\frac{1}{2}, \frac{2\beta}{4\beta-1})$: We set $q = \frac{4\beta-1}{2\beta}$ and use the first case. The assumption $\eta > 2 - 2\theta = \max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}$ implies $\theta \in \left(1 - \frac{\eta}{2}, \frac{1}{q}\right]$. (Note that $\eta > \frac{4\beta-2}{4\beta-1}$ is automatically fulfilled by $2 - 2\theta > \frac{4\beta-2}{4\beta-1}$ for this choice of θ.)

- Let $\theta \in [\frac{2\beta}{4\beta-1}, 1)$: By assumption we have $\eta > \frac{\theta+2\beta-2}{2\beta-1} = \max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}$. We choose some $\frac{1}{q} \in (\theta, 2\beta\eta - 2\beta + 2 - \eta)$ and use the second case. (Note that $\eta > \frac{4\beta-2}{4\beta-1}$ again is automatically fulfilled by $\frac{\theta+2\beta-2}{2\beta-1} > \frac{4\beta-2}{4\beta-1}$ for this choice of θ.)

All in all we have proved that $\theta \in (\frac{1}{2}, 2)$ implies $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$ almost surely for all $\eta \in (\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1)$. □

C Proof of Theorem 4.1

Proof. Recall, in Section 2 we derived Equation (5),

$$\begin{aligned}&\mathbb{E}[f(X_{n+1}) - f^* \mid \mathcal{F}_n]\\ &\leq \Big(1 + \frac{LA\gamma_n^2}{2}\Big)(f(X_n) - f^*) - \Big(\gamma_n - \frac{BL\gamma_n^2}{2}\Big)c^2(f(x) - f^*)^{2\beta} + \frac{LC\gamma_n^2}{2},\end{aligned}$$

which will be the basis of the proof of Theorem 4.1.

We treat again both cases for $\beta = \frac{1}{2}$ and $\beta \in (\frac{1}{2}, 1]$ separately:

$\underline{\beta = \frac{1}{2}}$: In this case, Equation (5) results in the super-martingale inequality

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 + \frac{LA\gamma_n^2}{2} - \gamma_n c^2 + \frac{BLc^2\gamma_n^2}{2}\Big)Y_n + \frac{LC\gamma_n^2}{2},$$

with $Y_n = f(X_n) - f^*$. By the choice of γ_n there exists $N > 0$ and a constant $\tilde{c} > 0$ such that $\gamma_n c^2 - \frac{LA\gamma_n^2}{2} - \frac{BLc^2\gamma_n^2}{2} \geq \tilde{c}\gamma_n$ for all $n \geq N$. Thus,

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 - \tilde{c}\gamma_n\Big)Y_n + \frac{LC\gamma_n^2}{2},$$

for all $n \geq N$. Then, claim (i) follows by applying Lemma 3.1 with $c_1 = 0, c_2 = \tilde{c}, c_3 = \frac{LC}{2}$ and $\beta = \frac{1}{2}$.

To prove claim (ii) we multiply $(n+1)^{1-\eta}$ on both sides and take the expectation. It follows that

$$\begin{aligned}\mathbb{E}[(n+1)^{(1-\eta)}Y_{n+1}] &\leq (1 - \tilde{c}\gamma_n)(n+1)^{1-\eta}\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\gamma_n^2\\ &\leq (1 - \tilde{c}\gamma_n)(n^{1-\eta} + (1-\eta)n^{-\eta})\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\gamma_n^2\\ &= \left(1 - \tilde{c}\gamma_n + \frac{1-\eta}{n} - \frac{\tilde{c}(1-\eta)\gamma_n}{n}\right)n^{1-\eta}\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\gamma_n^2.\end{aligned}$$

As $\theta_n \in \Theta(\frac{1}{n^\theta})$ we obtain that $\tilde{c}\gamma_n$ is the dominating term. Hence, there exists a constant $\tilde{c}_1 > 0$ and $\tilde{N} > N$ such that $\tilde{c}\gamma_n - \frac{1-\eta}{n} + \frac{\tilde{c}(1-\eta)\gamma_n}{n} \geq \tilde{c}_1\gamma_n$ for all $n \geq \tilde{N}$. Thus, for all $n \geq \tilde{N}$

$$\mathbb{E}[(n+1)^{(1-\eta)}Y_{n+1}] \leq (1 - \tilde{c}_1\gamma_n)\, n^{1-\eta}\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\gamma_n^2.$$

We apply Lemma A.3 with $w_n = n^{1-\eta}\mathbb{E}[Y_n]$, $a_n = \tilde{c}_1\gamma_n$ and $b_n = (n+1)^{1-\eta}\gamma_n^2$ and obtain that $n^{1-\eta}\mathbb{E}[Y_n] \to 0$ for $n \to \infty$ which yields claim (ii). Note that $\sum_n b_n < \infty$ as $1 - \eta < 2\theta - 1$ for $\eta \in (2 - 2\theta, 1)$.

$\underline{\beta \in (\frac{1}{2}, 1]}$: In this case, Equation (5) results in the super-martingale inequality

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 + \frac{LA\gamma_n^2}{2}\Big)Y_n - \Big(\gamma_n - \frac{BL\gamma_n^2}{2}\Big)c^2 Y_n^{2\beta} + \frac{LC\gamma_n^2}{2},$$

with $Y_n = f(X_n) - f^*$. By the choice of γ_n there exists $c_2 > 0$ and $N_1 > 0$ such that $c^2\gamma_n - \frac{BLc^2\gamma_n^2}{2} \geq c_2\gamma_n$ for all $n \geq N_1$,

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 + \frac{LA\gamma_n^2}{2}\Big)Y_n - c_2\gamma_n Y_n^{2\beta} + \frac{LC\gamma_n^2}{2}\,.$$

We deduce claim (i) from Lemma 3.1 with $c_1 = \frac{LA}{2}, c_2 = c_2, c_3 = \frac{LC}{2}$ and $\beta \in (\frac{1}{2}, 1]$.

For claim (ii) we firstly proceed as in the proof of Lemma 3.1. Therefore, one can choose the auxiliary parameter $1 < q \leq \frac{1}{\theta}$ and find constants $c_4, c_3, \tilde{N}_1 > 0$ such that for all $n \geq \tilde{N}_1$ by Equation(14) we have

$$\mathbb{E}[(n+1)^{1-\eta} Y_{n+1} \mid \mathcal{F}_n] \leq n^{1-\eta} Y_n - c_4 \frac{1}{n^{q\theta}} n^{1-\eta} Y_n + c_3 (n+1)^{1-\eta} (\gamma_n^{\frac{2\beta q - 1}{2\beta - 1}} + \gamma_n^2)\,.$$

Next, we take the expectation to obtain

$$\mathbb{E}[(n+1)^{1-\eta} Y_{n+1}] \leq (1 - c_4 \frac{1}{n^{q\theta}}) \mathbb{E}[n^{1-\eta} Y_n] + c_3 (n+1)^{1-\eta} (\gamma_n^{\frac{2\beta q - 1}{2\beta - 1}} + \gamma_n^2)$$

for all $n \geq \tilde{N}_1$, implying that $w_n = \mathbb{E}[n^{1-\eta} Y_n] \to 0$ as $n \to \infty$ by Lemma A.3. Note that we have chosen θ, η and q as in Lemma 3.1, such that $\sum_n \frac{1}{n^{q\theta}} = \infty$, $\sum_n (n+1)^{1-\eta} \gamma_n^{\frac{2\beta q - 1}{2\beta - 1}} < \infty$, and $\sum_n (n+1)^{1-\eta} \gamma_n^2 < \infty$ (see (15), (16) and (17)). Therefore, the assumptions of Lemma A.3 are met.

□

D Proof of Theorem 4.2

Recall the definition of the iteration scheme (SHB). Using the following definitions

$$Z_n = X_n + \frac{\nu}{1-\nu} W_n, \quad W_n = X_n - X_{n-1}, \tag{18}$$

one can derive the iterative evolution

$$W_{n+1} = \nu W_n - \gamma_n V(X_n) \tag{19}$$

$$Z_{n+1} = Z_n - \frac{\gamma_n}{1-\nu} V(X_n). \tag{20}$$

We will utilize these auxiliary variables in the proof.

Proof. The proof begins as in the proof of Theorem 2 in [27]. Using only L-smoothness and assumption (ABC), they show that for any $c_3 \in (0, \frac{1}{1-\nu})$, $\lambda \in (\nu, 1)$ there exist constants $c_1, c_2, c_4 > 0$ such that choosing the step size $\gamma_n \sim \frac{1}{n^\theta}$, for some $\theta \in (\frac{1}{2}, 1)$ results in [27, Equation (21)]

$$\begin{aligned} &\mathbb{E}[f(Z_{n+1}) - f^* + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\leq (1 + c_1 \gamma_n^2)(f(Z_n) - f^*) + (\lambda + c_2 \gamma_n^2) \|W_n\|^2 - c_3 \gamma_n \|\nabla f(Z_n)\|^2 + c_4 \gamma_n^2 \end{aligned} \tag{21}$$

for all $n \geq N$ and some $N > 0$ sufficiently large. Next, we apply the global gradient domination property for any $\beta \in [\frac{1}{2}, 1]$ to derive

$$\begin{aligned} &\mathbb{E}[f(Z_{n+1}) - f^* + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\leq (1 + c_1 \gamma_n^2)(f(Z_n) - f^*) - c c_3 \gamma_n (f(Z_n) - f^*)^{2\beta} + (\lambda + c_2 \gamma_n^2) \|W_n\|^2 + c_4 \gamma_n^2. \end{aligned} \tag{22}$$

For the remaining proof, we denote $Q_n := f(Z_n) - f^*$. Similar as before, we treat both cases for $\beta = \frac{1}{2}$ and $\beta \in (\frac{1}{2}, 1]$ separately:

$\underline{\beta = \frac{1}{2}}$: Instead of μ-strong convexity we use the gradient domination inequality $\|\nabla f(x)\|^2 \geq c(f^* - f(x))$, as the same inequality is implied by strong convexity using $c = \mu$. Then, Claim (i) follows using the same proof as [27, Thm. 2b)]. Note that the inequality

$$\frac{1}{2L} \|\nabla f(x)\|^2 \leq f(x) - f^*, \tag{23}$$

used in the last step only requires the L-smoothness assumption [38, Sec. 1.2.3].

For Claim (ii) we consider Equation (22) which simplifies for $\beta = \frac{1}{2}$ to

$$\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \leq (1 + c_1 \gamma_n^2 - c c_3 \gamma_n) Q_n + (\lambda + c_2 \gamma_n^2) \|W_n\|^2 + c_4 \gamma_n^2.$$

By the choice of γ_n there exists $N > 0$ and $\tilde{c}_1, \tilde{c}_2 > 0$, such that $cc_3\gamma_n - c_1\gamma_n^2 \geq \tilde{c}_1\gamma_n$ and $\lambda + c_2\gamma_n^2 \leq \tilde{c}_2\gamma_n$ for all $n \geq N$. Hence, for $n \geq N$

$$\begin{aligned}\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] &\leq (1 - \tilde{c}_1\gamma_n)Q_n + (1 - \tilde{c}_2\gamma_n)\|W_n\|^2 + c_4\gamma_n^2 \\ &\leq (1 - \min\{\tilde{c}_1, \tilde{c}_2\}) \left(Q_n + \|W_n\|^2\right) + c_4\gamma_n^2.\end{aligned}$$

Let $c_5 = \min\{\tilde{c}_1, \tilde{c}_2\}$, multiply by $(n+1)^{1-\eta}$ on both sides and use Equation (9) to obtain for $n \geq N$

$$\begin{aligned}&\mathbb{E}[(n+1)^{1-\eta} \left(Q_{n+1} + \|W_{n+1}\|^2\right) \mid \mathcal{F}_n] \\ &\leq (n+1)^{1-\eta}(1 - c_5\gamma_n) \left(Q_n + \|W_n\|^2\right) + c_4\gamma_n^2(n+1)^{1-\eta} \\ &\leq (n^{1-\eta} + (1-\eta)n^{-\eta})(1 - c_5) \left(Q_n + \|W_n\|^2\right) + c_4\gamma_n^2(n+1)^{1-\eta} \\ &= \left(1 - c_5\gamma_n + \frac{1-\eta}{n} - \frac{c_5(1-\eta)\gamma_n}{n}\right) n^{1-\eta} \left(Q_n + \|W_n\|^2\right) + c_4\gamma_n^2(n+1)^{1-\eta}.\end{aligned}$$

Taking expectation and using that there exists $\tilde{c}_5 > 0$ and $\tilde{N} > N$ such that $c_5\gamma_n - \frac{1-\eta}{n} + \frac{c_5(1-\eta)\gamma_n}{n} \geq \tilde{c}_5\gamma_n$, we have for all $n \geq \tilde{N}$

$$\begin{aligned}&\mathbb{E}[(n+1)^{1-\eta} \left(Q_{n+1} + \|W_{n+1}\|^2\right)] \\ &\leq (1 - \tilde{c}_5\gamma_n)\, \mathbb{E}\left[n^{1-\eta} \left(Q_n + \|W_n\|^2\right)\right] + c_4\gamma_n^2(n+1)^{1-\eta}.\end{aligned}$$

Note that $\sum_n \gamma_n^2(n+1)^{1-\eta} < \infty$ because $\eta \in (2 - 2\theta, 1)$ implies $1 - \eta < 2\theta - 1$. We can apply Lemma A.3 which yields that $\mathbb{E}\left[n^{1-\eta} \left(Q_n + \|W_n\|^2\right)\right] \to 0$. Hence, $\mathbb{E}[\left(Q_n + \|W_n\|^2\right)] \in o(\frac{1}{n^{1-\eta}})$.
To finish the proof, one can derive

$$f(X_n) - f^* \leq Q_n + \frac{1}{2}\|\nabla f(X_n)\|^2 + \frac{\nu^2 + L\nu^2}{1(1-\nu)^2}\|W_n\|^2 \tag{24}$$

see [27, Equation (19)] for more details. Using inequality (23), we get almost surely

$$\left(1 - \frac{1}{4L}\right) f(X_n) - f^* \leq Q_n + \frac{\nu^2 + L\nu^2}{1(1-\nu)^2}\|W_n\|^2. \tag{25}$$

implying that $\mathbb{E}[(f(X_n) - f^*)] \in o(\frac{1}{n^{1-\eta}})$ which proves Claim (ii).

<u>$\beta \in (\frac{1}{2}, 1]$</u>: For Claim (i), note that in Equation (22) $\lambda < 1$, such that

$$\begin{aligned}&\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\leq (1 + c_1\gamma_n^2)Q_n + (1 + c_2\gamma_n^2)\|W_n\|^2 + cc_3\gamma_n Q_n^{2\beta} + c_4\gamma_n^2 \\ &\leq (1 + \max\{c_1, c_2\}\gamma_n^2)(Q_n + \|W_n\|^2) + cc_3\gamma_n(Q_n + \|W_n\|^2)^{2\beta} + c_4\gamma_n^2.\end{aligned}$$

By Lemma 3.1 we obtain that $Q_n + \|W_n\|^2 = f(Z_n) - f^* + \|W_n\|^2 \in o\left(\frac{1}{n^{1-\eta}}\right)$ for all $\eta \in \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right)$. We apply the inequality in (25) to conclude that also $f(X_n) - f^* \in o\left(\frac{1}{n^{1-\eta}}\right)$ for all $\eta \in \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right)$. This proves Claim (i).

For Claim (ii), we again use the q-trick from Lemma 3.1 in Equation (22). For $1 < q < \frac{1}{\theta} < 2$ we have that

$$\begin{aligned}&\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\leq (1 + c_1\gamma_n^2 - cc_3\gamma_n^q)Q_n + cc_3\gamma_n \left(\gamma_n^{q-1}Q_n - Q_n^{2\beta}\right) + (\lambda + c_2\gamma_n^2)\|W_n\|^2 + c_4\gamma_n^2.\end{aligned}$$

Now with (12) in Lemma 3.1 there exists $\tilde{c}_3 \geq 0$ such that

$$\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \leq (1 + c_1\gamma_n^2 - cc_3\gamma_n^q)Q_n + \tilde{c}_3\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + (\lambda + c_2\gamma_n^2)\|W_n\|^2 + c_4\gamma_n^2.$$

By the choice of γ_n there exists $\tilde{c}_1 > 0$ and $N > 0$ such that $c_1\gamma_n^2 - cc_3\gamma_n^q \geq \tilde{c}_1\gamma_n^q$ and $\lambda + c_2\gamma_n^2 \leq \tilde{c}_1\gamma_n^q$ for all $n \geq N$. Thus, for all $n \geq N$,

$$\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \leq (1 - \tilde{c}_1\gamma_n^q)(Q_n + \|W_n\|^2) + \max\{\tilde{c}_3, c_4\} \left(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2\right).$$

For $\max\{\tilde{c}_3, c_4\} =: \tilde{c}_2$, we multiply on both sides with $(n+1)^{1-\eta}$ and take the expectation to obtain for $n \geq N$

$$\begin{aligned}
&\mathbb{E}[(n+1)^{1-\eta}\left(Q_{n+1} + \|W_{n+1}\|^2\right)] \\
&\leq (n+1)^{1-\eta}(1-\tilde{c}_1\gamma_n^q)\mathbb{E}[(Q_n + \|W_n\|^2)] + \tilde{c}_2(n+1)^{1-\eta}\left(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2\right) \\
&\leq (n^{1-\eta} + (1-\eta)n^{-\eta})(1-\tilde{c}_1\gamma_n^q)\mathbb{E}[(Q_n + \|W_n\|^2)] + \tilde{c}_2(n+1)^{1-\eta}\left(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2\right) \\
&= \left(1 - \tilde{c}_1\gamma_n^q + \frac{1-\eta}{n} - \frac{\tilde{c}_1(1-\eta)\gamma_n^q}{n}\right)\mathbb{E}[n^{1-\eta}(Q_n + \|W_n\|^2)] \\
&\quad + \tilde{c}_2(n+1)^{1-\eta}\left(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2\right).
\end{aligned}$$

Next, there exists $\tilde{N} > N$ and $\tilde{c}_5 > 0$ such that for all $n \geq \tilde{N}$

$$\begin{aligned}
&\mathbb{E}[(n+1)^{1-\eta}\left(Q_{n+1} + \|W_{n+1}\|^2\right)] \\
&\leq (1-\tilde{c}_5\gamma_n^q)\,\mathbb{E}[n^{1-\eta}(Q_n + \|W_n\|^2)] + \tilde{c}_2(n+1)^{1-\eta}\left(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2\right).
\end{aligned}$$

From the proof of Lemma 3.1, we choose the auxiliary parameter q such that $\sum_n (n+1)^{1-\eta}\left(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2\right) < \infty$ (see (16) and (17)). By applying again Lemma A.3 we obtain $\mathbb{E}[n^{1-\eta}(Q_n + \|W_n\|^2)] \to 0$, i.e. $\mathbb{E}[Q_n + \|W_n\|^2] \in o(\frac{1}{n^{1-\eta}})$. Finally, Claim (ii) follows again by Equation (25). □

E Numerical experiment - Toy example

We have implemented the same toy example similar to [12] to test our theoretical findings. In our implementation, we consider both SGD and SHB applied to the objective function $f_p(x) = |x|^p$, where $x \in \mathbb{R}$, for various choices of $p \geq 2$. It is straightforward to verify that f_p satisfies the global gradient domination with parameter $\beta(p) = \frac{p-1}{p}$. It is noteworthy that for $p = 2$, the f_p obviously satisfies the PL condition with $\beta = \frac{1}{2}$, whereas for increasing $p \to \infty$, we move towards $\beta(p) \to 1$. We have used the step size schedule $\Theta(n^{-\frac{2\beta(p)}{4\beta(p)-1}})$ discussed in Table 1 and observed the almost sure convergence rates $n^{-\frac{1}{4\beta(p)-1}}$ as suggested by Theorem 4.1 and Theorem 4.2. Note that our derived rates are arbitrarily close to the sharp upper bound known in expectation [12].

Details of the implementation: Both algorithms have been implemented by hand using `MATLAB`. We have initialized both SGD and SHB with the initial state $X_1 \sim \frac{1}{2}\mathcal{U}([1.5, 2.5]) + \frac{1}{2}\mathcal{U}([-2.5, 1.5])$ to force initials which are not close to the actual minimum $x^* = 0$. The initial step sizes $\gamma_1(\beta)$ for both algorithms are chosen as

$$\gamma_1(0.5) = 0.2,\ \gamma_1(0.67) = 0.13,\ \gamma_1(0.83) = 0.004,\ \gamma_1(0.92) = 10^{-6}$$

through which we counteract the decreasing smoothness for $\beta \to 1$. The momentum parameter for SHB is fixed for all β as $\nu = 0.5$. The exact gradients ∇f_p are perturbed by independent additive noise following a standard normal distribution $\mathcal{N}(0, 1)$.

F Proof of Section 5

F.1 Proof of Theorem 5.1

Suppose that the assumptions of Theorem 5.1 hold throughout this section.

In contrast to the global gradient domination analysis we may assume w.l.o.g. the uniform second moment bounds, i.e. $A = B = 0$, instead of the more general (ABC) condition. Choosing $A, B > 0$ would imply the bounded variance assumption of the gradient estimator. Note therefore, that the

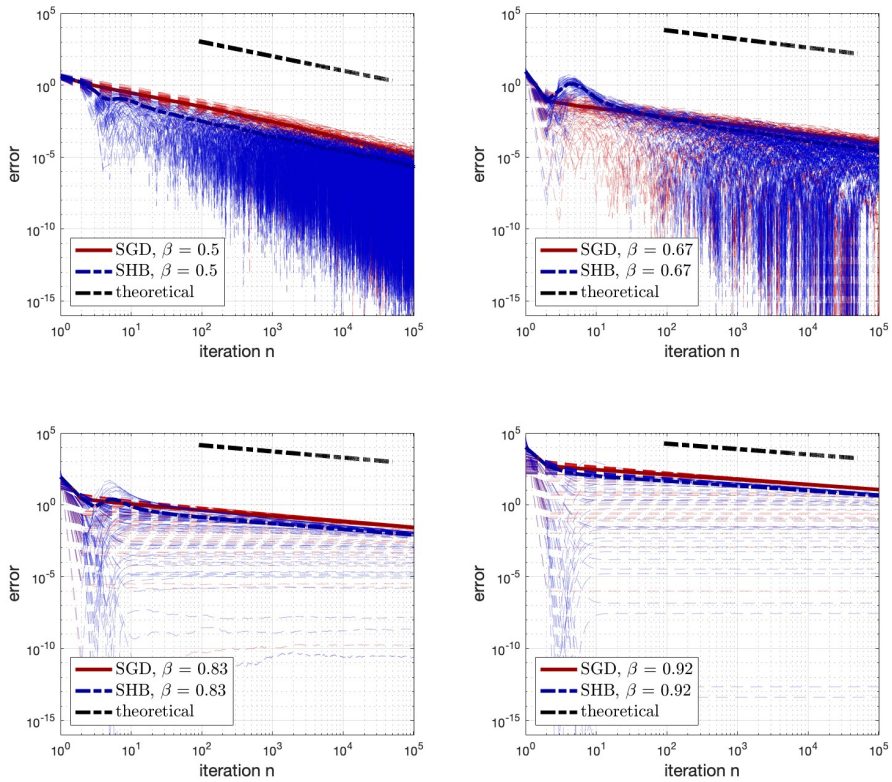

Figure 1: Pathwise error $(f_p(X_n))_{n=1,\dots,N}$ of SGD and SHB for various choices of $\beta \in \{0.5, 0.67, 0.83, 0.92\}$. For each setting we have simulated 100 runs of length , $N = 10^5$. The black dash-dotted line corresponds to the theoretical rate $n^{-\frac{1}{4\beta-1}}$.

first term $A(f(x) - f(x^*))$ and the second term $B\|\nabla f(x)\|^2$ are both locally bounded by the local Lipschitz assumptions on f and ∇f.

Note that every isolated local minimum $\{x^*\}$ is a special case of an isolated compact connected set of local minima. In this case it holds that $\beta = \beta_{x^*}$. If $\mathcal{X}^*$ contains more then one point, we can unify the gradient domination property in a neighbourhood of $\mathcal{X}^*$ due to compactness. The set $\mathcal{X}^*$ has to be connected to assure that all local minima are on the same level l.

The outline of the proof is structured as follows:

- First, we unify the gradient domination property around the set of local minima $\mathcal{X}^*$ and obtain a radius r such that the unified gradient domination property is fulfilled in all open balls with radius r around $x^* \in \mathcal{X}^*$ (Lemma F.1).
- Based on this we construct sets $\mathcal{U}, \mathcal{U}_1 \subseteq \mathbb{R}^d$ and the events $\Omega_n \in \Omega$ (see (26), (27) and (28)), such that $\Omega_{\mathcal{U}} = \bigcap_n \Omega_n$ occurs with high probability. To be precise, $\mathcal{U}_1$ and $\mathcal{U}$ are neighborhoods of $\mathcal{X}^*$ constructed such that the gradient domination property holds within this region, and when starting in $\mathcal{U}_1$ the gradient trajectory does remain in $\mathcal{U}$ for all gradient steps with high probability. Then, Ω_n describes the event that $X_k \in \mathcal{U}$ for all $k \leq n$.
- All following Lemmata before the proof of Theorem 5.1 are devoted to show that $\mathbb{P}(\Omega_n) \geq 1 - \delta$ for all $n \in \mathbb{N}$. This then proves Claim (i) of the Theorem. Claim (ii) and (iii) will be shown directly in the proof of Theorem 5.1 at the end of this subsection.
- In order to show $\mathbb{P}(\Omega_n) \geq 1 - \delta$ we construct set C_n and E_n defined in (29) and (36) such that $E_n \cap C_n \subset \Omega_{n+1}$ (Lemma F.6) while Lemma F.5 is used to prove this claim.

- The sets E_n are such that $f(X_n)$ remains close to f^*. We exploit the unified gradient domination property to construct the sets E_n (Lemma F.4) and derive a recursive inequality in Lemma F.6 c) to prove that this event occurs with high probability (Lemma F.7).
- The sets C_n are such that X_{n+1} remains close to X_n and we exploit the finite variance assumption to show that these events occur with high probability (Lemma F.8).

We denote by

$$\tilde{\mathcal{B}}_r(x) = \{y \in \mathbb{R}^d \,:\, ||x-y|| < r\}$$

the open ball with radius $r > 0$ around $x \in \mathbb{R}^d$ and by

$$\mathcal{B}_r(x) = \{y \in \mathbb{R}^d \,:\, ||x-y|| \leq r\}$$

the closed ball with radius $r > 0$ around $x \in \mathbb{R}^d$.

In the following Lemma we unify the gradient domination property around the set of local minima $\mathcal{X}^* \subset \mathbb{R}^d$.

Lemma F.1. . *There exists $r > 0$, $\beta \in [\frac{1}{2}, 1]$ and $c > 0$, such that for all $x \in \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_r(x^*)$ it holds that*

$$f(x) > l \text{ for } x \notin \mathcal{X}^* \quad \text{and} \quad ||\nabla f(x)|| \geq c(f(x)-l)^\beta \,.$$

Proof. By the local gradient domination property, for every $x^* \in \mathcal{X}^*$ there exist $r_{x^*} > 0$, $\beta_{x^*} \in [\frac{1}{2}, 1]$ and $c_{x^*} > 0$ such that

$$||\nabla f(x)|| \geq c_{x^*} |f(x) - l|^{\beta_{x^*}}, \quad \forall x \in \mathcal{B}_{r_{x^*}}(x^*).$$

Moreover, w.l.o.g we can assume that $f(x) > l$ for all $x \in \mathcal{B}_{r_{x^*}}(x^*) \setminus \mathcal{X}^*$, as $\mathcal{X}^*$ is an isolated compact connected set of local minima (otherwise choose r_{x^*} small enough).

By the compactness of $\mathcal{X}^*$ we can find a finite subset $\mathcal{Y}^* \subset \mathcal{X}^*$, such that

$$\tilde{\mathcal{U}} := \bigcup_{y^* \in \mathcal{Y}^*} \tilde{\mathcal{B}}_{r_{y^*}}(y^*) \supset \mathcal{X}^*.$$

Then, we define $\beta = \max_{y^* \in \mathcal{Y}^*} \beta_{y^*}$ and $c = \min_{y^* \in \mathcal{Y}^*} c_{y^*}$. For any $x \in \tilde{\mathcal{U}}$ there exits $y^* \in \mathcal{Y}^*$ such that

$$||\nabla f(x)|| \geq c_{y^*}(f(x)-l)^{\beta_{y^*}} \geq c(f(x)-l)^\beta.$$

Thus, there exists an open neighbourhood $\tilde{\mathcal{U}}$ of $\mathcal{X}^*$ and $\beta \in [\frac{1}{2}, 1]$, $c > 0$, such that for all $x \in \tilde{\mathcal{U}}$ it holds that

$$f(x) > l \text{ for } x \notin \mathcal{X}^* \quad \text{and} \quad ||\nabla f(x)|| \geq c(f(x)-l)^\beta.$$

As $\tilde{\mathcal{U}}$ is open by definition and $\mathcal{X}^* \subset \tilde{\mathcal{U}}$, we can find a radius $r > 0$, such that $\bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_r(x^*) \subseteq \tilde{\mathcal{U}}$. This proves the claim. □

Remark F.2. It is noteworthy that the unified gradient domination property obtained in the previous Lemma does not require an absolute value, as $f(x) \geq l$ for all $x \in \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_r(x^*)$. This is crucial to obtain the recursive inequalities in Lemma F.4 and we will exploit this also in the proof of Theorem 5.1 to obtain the convergence rates.

In the following let $\mathbf{r} > 0$, $c > 0$ and $\beta \in [\frac{1}{2}, 1]$ chosen as in the previous Lemma, such that the unified gradient domination property holds for all $x \in \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_{\mathbf{r}}(x^*)$. Further define

$$s = \inf \left\{ f(x) - l \,:\, x \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*) \right\}.$$

Lemma F.3. *It holds that $s > 0$.*

Proof. If $s = 0$, then there exists a sequence $(x_n) \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*)$ with $f(x_n) \to l$ for $n \to \infty$. By definition of the set and compactness (boundedness) of $\mathcal{X}^*$, the sequence x_n is bounded:

$$||x_n|| \leq \frac{3\mathbf{r}}{4} + \sup_{x^* \in \mathcal{X}^*} ||x^*|| < \infty.$$

Hence, there is a convergent sub-sequence (x_{n_k}) with $x_{n_k} \to x$ for $k \to \infty$ and by continuity of f it holds that $f(x) = l$. Further, it holds for all $x^* \in \mathcal{X}^*$ that $||x_n - x^*|| \geq \frac{\mathbf{r}}{2}$ for all $n \in \mathbb{N}$ such that $\inf_{x^* \in \mathcal{X}^*} ||x - x^*|| \geq \frac{\mathbf{r}}{2}$.

On the other hand, by construction we have that $x \in \overline{\bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*)} \subset \overline{\bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*)} \subset \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_{\mathbf{r}}(x^*)$. And as $f(y) > l$ for all $y \in \tilde{\mathcal{B}}_{\mathbf{r}}(x^*) \setminus \mathcal{X}^*$ we deduce from $f(x) = l$ that $x \in \mathcal{X}^*$. This is a contradiction to $\inf_{x^* \in \mathcal{X}^*} ||x - x^*|| \geq \frac{\mathbf{r}}{2}$. □

We choose $\epsilon > 0$, such that $2\epsilon + \sqrt{\epsilon} < s$. We define the sets

$$\mathcal{U}_1 = \{x \in \mathbb{R}^d : \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}, f(x) - l \leq \frac{\epsilon}{2}\} \tag{26}$$

$$\mathcal{U} = \{x \in \mathbb{R}^d : \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}\} \tag{27}$$

which are subsets of $\mathbb{R}^d$ and the decreasing sequence of events

$$\Omega_n = \{X_k \in \mathcal{U} \text{ for all } k \leq n\} \tag{28}$$

$$C_n = \{||X_{k+1} - X_k|| \leq \frac{\mathbf{r}}{4} \text{ for all } k \leq n\}, \tag{29}$$

and $C_0 = \Omega$, which are measurable sets in $(\Omega, \mathcal{F}, \mathbb{P})$.

In order to prove Theorem 5.1 we will show that Ω_n has probability at least $1 - \delta$ for all $n \in \mathbb{N}$. To do this, we construct another sequence of events $(\hat{E}_n)$ with $\hat{E}_n \subset \Omega_n$ which occurs with probability at least $1 - \delta$ for any $n \in \mathbb{N}$.

Therefore, we fix the notation $D_n := f(X_n) - l$ and $\tilde{D}_n := D_n \mathbf{1}_{\Omega_n}$[1] and prove the following (recursive) inequalities.

Lemma F.4. *If $\beta = \frac{1}{2}$, then it holds that*

$$\tilde{D}_{n+1} \leq (1 - \gamma_n c^2)\tilde{D}_n + \gamma_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2} \mathbf{1}_{\Omega_n} \|V_{n+1}(X_n)\|^2, \tag{30}$$

$$\leq \tilde{D}_1 \prod_{k=1}^{n} (1 - \gamma_k c^2) + \sum_{k=1}^{n} \gamma_k \xi_k \mathbf{1}_{\Omega_k} + \frac{L}{2} \sum_{k=1}^{n} \gamma_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_k}. \tag{31}$$

If $\beta \in (\frac{1}{2}, 1]$, for any $1 \leq q < 2$, it holds that

$$\tilde{D}_{n+1} \leq (1 - \gamma_n^q c^2)\tilde{D}_n + (2\beta)^{-\frac{1}{2\beta - 1}} (1 - \frac{1}{2\beta}) c^2 \gamma_n^{\frac{2\beta q - 1}{2\beta - 1}} + \gamma_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2} \|V_{n+1}(X_n)\|^2 \tag{32}$$

$$\leq \tilde{D}_1 \prod_{k=1}^{n} (1 - \gamma_k^q c^2) + \tilde{c} \sum_{k=1}^{n} \gamma_k^{\frac{2\beta q - 1}{2\beta - 1}} + \sum_{k=1}^{n} \gamma_k \xi_k \mathbf{1}_{\Omega_k} + \frac{L}{2} \sum_{k=1}^{n} \gamma_k^2 \|V_{k+1}(X_k)\|^2, \tag{33}$$

for $\tilde{c} = (2\beta)^{-\frac{1}{2\beta - 1}} (1 - \frac{1}{2\beta}) c^2$.

Proof. From L-smoothness we can deduce that

$$\begin{aligned} D_{n+1} &\leq D_n - \gamma_n \langle \nabla f(X_n), V_{n+1}(X_n) \rangle + \frac{L\gamma_n^2}{2} \|V_{n+1}(X_n)\|^2 \\ &= D_n - \gamma_n \|\nabla f(X_n)\|^2 - \gamma_n \langle \nabla f(X_n), Z(X_n, \zeta_{n+1}) \rangle + \frac{L\gamma_n^2}{2} \|V_{n+1}(X_n)\|^2 \\ &= D_n - \gamma_n \|\nabla f(X_n)\|^2 + \gamma_n \xi_{n+1} + \frac{L\gamma_n^2}{2} \|V_{n+1}(X_n)\|^2 \end{aligned}$$

for $Z(X_n, \zeta_{n+1})$ from Assumption 2.4 and $\xi_{n+1} = -\langle \nabla f(X_n), Z(X_n, \zeta_{n+1}) \rangle$.

We separate the two cases of β:

[1] $\mathbf{1}_{\mathcal{A}}$ denoted the indicator function for a measurable set $\mathcal{A}$ in $(\Omega, \mathcal{F}, \mathbb{P})$, i.e. $\mathbf{1}_{\mathcal{A}}(\omega) = 1$ if $\omega \in \mathcal{A}$ and $\mathbf{1}_{\mathcal{A}}(\omega) = 0$ if $\omega \notin \mathcal{A}$.

$\underline{\beta = \frac{1}{2}}$: Iterating this inequality and using $\mathbf{1}_{\Omega_{n+1}} \leq \mathbf{1}_{\Omega_n}$ it follows that

$$\begin{aligned}
\tilde{D}_{n+1} &\leq D_{n+1}\mathbf{1}_{\Omega_n} \\
&\leq D_n\mathbf{1}_{\Omega_n} - \gamma_n\mathbf{1}_{\Omega_n}\|\nabla f(X_n)\|^2 + \gamma_n\mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\gamma_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\
&\leq D_n\mathbf{1}_{\Omega_n} - \gamma_n c^2(f(X_n) - l)\mathbf{1}_{\Omega_n} + \gamma_n\mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\gamma_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\
&= (1-\gamma_n c^2)\tilde{D}_n + \gamma_n\xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2, \\
&\leq \tilde{D}_1\prod_{k=1}^{n}(1-\gamma_k c^2) + \sum_{k=1}^{n}\gamma_k\xi_{k+1}\mathbf{1}_{\Omega_k} + \frac{L}{2}\sum_{k=1}^{n}\gamma_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_k},
\end{aligned} \tag{34}$$

where we used that the unified gradient domination property holds for all X_k, $k \leq n$ on the event Ω_n.

$\underline{\beta \in (\frac{1}{2}, 1]}$: Similarly, the unified gradient domination property yields the claimed inequality for any $1 \leq q < 2$:

$$\begin{aligned}
\tilde{D}_{n+1} &\leq D_{n+1}\mathbf{1}_{\Omega_n} \\
&\leq D_n\mathbf{1}_{\Omega_n} - \gamma_n\mathbf{1}_{\Omega_n}\|\nabla f(X_n)\|^2 + \gamma_n\mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\gamma_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\
&\leq D_n\mathbf{1}_{\Omega_n} - \gamma_n c^2(f(X_n) - l)^{2\beta}\mathbf{1}_{\Omega_n} + \gamma_n\mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\gamma_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\
&= \tilde{D}_n - \gamma_n c^2\tilde{D}_n^{2\beta} + \gamma_n\xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2, \\
&= (1-\gamma_n^q c^2)\tilde{D}_n + \gamma_n c^2(\gamma_n^{1-q}\tilde{D}_n - \tilde{D}_n^{2\beta}) + \gamma_n\xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n} \\
&\leq (1-\gamma_n^q c^2)\tilde{D}_n + (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n\xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n} \\
&\leq \tilde{D}_1\prod_{k=1}^{n}(1-\gamma_k^q c^2) + \tilde{c}\sum_{k=1}^{n}\gamma_k^{\frac{2\beta q-1}{2\beta-1}} + \sum_{k=1}^{n}\gamma_k\xi_{k+1}\mathbf{1}_{\Omega_k} + \frac{L}{2}\sum_{k=1}^{n}\gamma_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_k},
\end{aligned} \tag{35}$$

for $\tilde{c} = (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2$ from the function trick (12) which we applied in the forth inequality. We also used that the unified gradient domination property holds for all X_k, $k \leq n$ on the event Ω_n . □

For $\beta \in (\frac{1}{2}, 1]$ we know from the proof of Lemma 3.1 that we can choose the auxiliary parameter q from the previous lemma in such a way, that $\sum_{n=1}^{\infty} n^{1-\eta}\gamma_n^{\frac{2\beta q-1}{2\beta-1}}$ is convergent for all $\eta \in (\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1)$ (Condition (iii) to apply Lemma A.2). As $\eta < 1$, it follows that $\sum_{n=1}^{\infty}\gamma_n^{\frac{2\beta q-1}{2\beta-1}} < \infty$ holds true for all these choices of q. Now define

$$M_n = \sum_{k=1}^{n}\gamma_k\xi_{k+1}\mathbf{1}_{\Omega_k} \quad \text{and} \quad S_n = \frac{L}{2}\sum_{k=1}^{n}\gamma_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_k}.$$

Then, (M_n) is a $(\mathcal{F}_{n-1})$-martingale with zero mean and (S_n) is a $(\mathcal{F}_{n-1})$-sub-martingale by Assumption 2.4. Note that by the choice of γ_n we have that $\sum_n \gamma_n^2 < \infty$ and hence $\mathbb{E}[S_n] < \infty$ for all $n \in \mathbb{N}$.

Next, define $R_n = M_n^2 + S_n$ for every $n \in \mathbb{N}$. Moreover, let

$$E_n = \{R_k < \epsilon \text{ for all } k \leq n\}. \tag{36}$$

which is an $\mathcal{F}_n$-measurable event on $(\Omega, \mathcal{F}, \mathbb{P})$. We define $R_0 = 0$ such that $E_0 = \Omega$.

Now let $\hat{E}_n = E_n \cap C_n$, then we will first show, that $\hat{E}_n$ fulfills the property $\hat{E}_n \subset \Omega_{n+1}$ for all $n \in \mathbb{N}$ in Lemma F.6 and then that $\hat{E}_n$ occurs with probability at least $1-\delta$ in Lemma F.8.

To prove that $\hat{E}_n \subset \Omega_{n+1}$ we need one more auxiliary result.

Lemma F.5. *Suppose $x, y \in \mathbb{R}^d$ such that*

1. $\inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}$,
2. $f(y) - l < s$,
3. $||x - y|| \leq \frac{\mathbf{r}}{4}$.

Then it holds that $\inf_{x^* \in \mathcal{X}^*} ||y - x^*|| < \frac{\mathbf{r}}{2}$.

Proof. By triangle inequality we have that $\inf_{x^* \in \mathcal{X}^*} ||y - x^*|| \leq \frac{3\mathbf{r}}{4}$, i.e there exists $x^* \in \mathcal{X}^*$ such that $||y - x^*|| \leq \frac{3\mathbf{r}}{4}$. Suppose now, that $\inf_{x^* \in \mathcal{X}^*} ||y - x^*|| \geq \frac{\mathbf{r}}{2}$, this means that $y \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*)$. By the definition of $s = \inf \left\{ f(z) - l \,:\, z \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \tilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*) \right\}$ this contradicts the second assumption $f(y) - l < s$. □

We deduce the following relations on the constructed sets:

Lemma F.6. *For $\beta \in (\frac{1}{2}, 1]$ let $\gamma_n \leq \gamma_1$ be sufficiently small such that $\sum_{n=1}^{\infty} \gamma_n^{\frac{2\beta q - 1}{2\beta - 1}} < \frac{\epsilon}{2\tilde{c}}$, and for $\beta = \frac{1}{2}$ let $\gamma_1 > 0$ be arbitrary. Furthermore, assume that the initial $X_1 \in \mathcal{U}_1$ almost surely. Then,*

a) $E_{n+1} \subset E_n$, $\hat{E}_{n+1} \subset \hat{E}_n$ and $\Omega_{n+1} \subset \Omega_n$

b) $\hat{E}_n \subset \Omega_{n+1}$

c) Define the events $\tilde{E}_n = E_{n-1} \setminus E_n = E_{n-1} \cup \{R_n \geq \epsilon\}$. Then, for $\tilde{R}_n = R_n \mathbf{1}_{E_{n-1}}$, there exists a $\tilde{C} > 0$ such that

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_{n-1}] + \gamma_n^2 [G^2 C^2 + G^2 + C] - \epsilon \mathbb{P}(\tilde{E}_{n-1}).$$

Proof. a) Follows by definition of the events.

b) Note that $\hat{E}_0 = \Omega = \Omega_1$ because

$$X_1 \in \mathcal{U}_1 = \{x : \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}, f(x) - l \leq \frac{\epsilon}{2}\} \subset \{x : \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}\} = \Omega_1$$

almost surely. We prove the assertion by induction. Let $\omega \in \hat{E}_n$. Since $\hat{E}_n \subset \hat{E}_{n-1} \subset \Omega_n$ by induction assumption, we have $\omega \in \Omega_n$ and thus $\omega \in \Omega_k$ for all $k \leq n$. We will apply Lemma F.5 with $x = X_n(\omega)$ and $y = X_{n+1}(\omega)$. By definition it holds that $\omega \in \hat{E}_n$ implies condition 3. and $\omega \in \Omega_n$ implies condition 1. of Lemma F.5. It remains to show condition 2., then it follows that $\inf_{x^* \in \mathcal{X}^*} ||X_{n+1}(\omega) - x^*|| < \frac{\mathbf{r}}{2}$, i.e. $X_{n+1}(\omega) \in \mathcal{U}$ and by $\omega \in \Omega_n$ we deduce $\omega \in \Omega_{n+1}$.

To Prove condition 2. we separate both cases for β:

<u>$\beta = \frac{1}{2}$</u>: The inequality (30) and the induction hypothesis yield

$$\begin{aligned}
D_{n+1}(\omega) &= D_{n+1}(\omega) \mathbf{1}_{\Omega_n}(\omega) \\
&\leq D_1(\omega) \prod_{k=1}^{n} (1 - \gamma_k c) + \sum_{k=1}^{n} \gamma_k \xi_{k+1}(\omega) \mathbf{1}_{\Omega_k}(\omega) + \frac{L}{2} \sum_{k=1}^{n} \gamma_k^2 \|V_{k+1}(X_k(\omega))\|^2 \mathbf{1}_{\Omega_k}(\omega) \\
&\leq \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\
&\leq 2\epsilon + \sqrt{\epsilon} < s.
\end{aligned}$$

$\underline{\beta \in (\frac{1}{2}, 1]}$: Similarly, we obtain from (45)

$$
\begin{aligned}
D_{n+1}(\omega) &= D_{n+1}(\omega)\mathbf{1}_{\Omega_n}(\omega) \\
&\leq D_1(\omega)\mathbf{1}_{\Omega_1}(\omega)\prod_{k=1}^{n}(1-\gamma_k^q c) + \tilde{c}\sum_{k=1}^{n}\gamma_k^{\frac{2\beta q-1}{2\beta-1}} + \sum_{k=1}^{n}\gamma_k\xi_{k+1}(\omega) + \frac{L}{2}\sum_{k=1}^{n}\gamma_k^2\|V_{k+1}(X_k(\omega))\|^2 \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\
&\leq 2\epsilon + \sqrt{\epsilon} < s.
\end{aligned}
$$

We used in both cases that that $\prod_{k=1}^{n}(1-\gamma_k^{(q)}c) \leq 1$ and the choice of ϵ such that $2\epsilon + \sqrt{\epsilon} < s$. This proves that condition 2. in Lemma F.5 is also satisfied which concludes the induction.

c) By definition it holds that $E_n = E_{n-1} \setminus (E_{n-1} \setminus E_n) = E_{n-1} \setminus \tilde{E}_n$. Then we have

$$
\begin{aligned}
\tilde{R}_n &= R_n\mathbf{1}_{E_{n-1}} \\
&= R_{n-1}\mathbf{1}_{E_{n-1}} + (R_n - R_{n-1})\mathbf{1}_{E_{n-1}} \\
&= R_{n-1}\mathbf{1}_{E_{n-2}} - R_{n-1}\mathbf{1}_{\tilde{E}_{n-1}} + (R_n - R_{n-1})\mathbf{1}_{E_{n-1}} \\
&= \tilde{R}_{n-1} - R_{n-1}\mathbf{1}_{\tilde{E}_{n-1}} + (R_n - R_{n-1})\mathbf{1}_{E_{n-1}}
\end{aligned}
$$

and for the last term

$$
\begin{aligned}
R_n - R_{n-1} &= M_n^2 - M_{n-1}^2 + S_n - S_{n-1} \\
&= \gamma_n^2\xi_{n+1}^2\mathbf{1}_{\Omega_n} + 2\gamma_n\xi_{n+1}\mathbf{1}_{\Omega_n}M_{n-1}^2 + \gamma_n^2\frac{L}{2}\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n}.
\end{aligned}
$$

We treat each of the summands on the RHS seperately. It follows from the G-Lipschitz continuity and bounded variance assumption in Theorem 5.1, that

$$
\begin{aligned}
\mathbb{E}[\xi_{n+1}^2\mathbf{1}_{\Omega_n}] &= \mathbb{E}[\langle\nabla f(X_n), V_{n+1}(X_n) - \nabla f(X_n)\rangle^2\mathbf{1}_{\Omega_n}] \\
&\leq \mathbb{E}[\|\nabla f(X_n)\|^2(\|V_{n+1}(X_n)\|^2 + 1)\mathbf{1}_{\Omega_n}] \leq G^2(C^2+1), \\
\mathbb{E}[\xi_{n+1}M_{n-1}^2\mathbf{1}_{\Omega_n}] &= \mathbb{E}[\mathbb{E}[\xi_{n+1}|\mathcal{F}_n]M_{n-1}^2\mathbf{1}_{\Omega_n}] = 0, \\
\mathbb{E}[\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n}] &\leq C.
\end{aligned} \tag{37}
$$

For the term $R_{n-1}\mathbf{1}_{\tilde{E}_{n-1}}$ we have

$$
\mathbb{E}[R_{n-1}\mathbf{1}_{\tilde{E}_{n-1}}] \geq \epsilon\mathbb{P}(\tilde{E}_{n-1}).
$$

Putting all together we obtain the claim

$$
\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_{n-1}] + \gamma_n^2[G^2C^2 + G^2 + C] - \epsilon\mathbb{P}(\tilde{E}_{n-1}).
$$

□

Lemma F.7. *Let $\delta > 0$ be a tolerance level and $\gamma_n \leq \gamma_1$ be sufficiently small such that $\sum_{n=1}^{\infty}\gamma_n^2 < \frac{\delta\epsilon}{2(G^2C^2+G^2+C)}$ and the condition in Lemma F.6 is fulfilled. Then, we have*

$$
\mathbb{P}(E_n) \geq 1 - \frac{\delta}{2}.
$$

Proof. The proof is along the lines of the proof of Proposition D2 in [35]. For completeness we repeat the arguments. First, observe that

$$
\mathbb{P}(\tilde{E}_{n-1}) = \mathbb{P}(E_{n-1}\backslash E_n) = \mathbb{P}(E_{n-1}\cap\{R_n \geq \epsilon\}) = \mathbb{E}[\mathbf{1}_{E_{n-1}}\mathbf{1}_{\{R_n>\epsilon\}}] \leq \mathbb{E}[\mathbf{1}_{E_{n-1}}\frac{R_n}{\epsilon}] = \frac{\mathbb{E}[\tilde{R}_n]}{\epsilon}.
$$

On the other hand it follows from Lemma F.10 that

$$
\epsilon\mathbb{P}(\tilde{E}_n) \leq \mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_0] + [G^2C^2 + G^2 + C]\sum_{k=1}^{n}\gamma_k^2 - \epsilon\sum_{k=0}^{n}\mathbb{P}(\tilde{E}_{k-1}). \tag{38}
$$

Rearranging everything yields

$$\sum_{k=0}^{n} \mathbb{P}(\tilde{E}_k) \leq \frac{[G^2C^2 + G^2 + C]\Gamma}{\epsilon}$$

with $\Gamma = \sum_{n=1}^{\infty} \gamma_n^2$. By the assumption on the step size $\frac{[G^2C^2+G^2+C]\Gamma}{\epsilon} < \frac{\delta}{2}$ and moreover since the events $\tilde{E}_n$ are disjoint we obtain

$$\mathbb{P}(\bigcup_{k=0}^{n} \tilde{E}_k) = \sum_{k=0}^{n} \mathbb{P}(\tilde{E}_k) \leq \frac{\delta}{2} \tag{39}$$

implying that

$$\mathbb{P}(E_n) = \mathbb{P}(\bigcap_{k=0}^{n} \tilde{E}_k^c) \geq 1 - \frac{\delta}{2}\,. \tag{40}$$

□

Lemma F.8. *Let $\delta > 0$ be a tolerance level and $\gamma_n \leq \gamma_1$ be sufficiently small such that the condition in Lemma F.6 and Lemma F.7 are fulfilled. Moreover, we suppose γ_1 small enough such that $\frac{4C}{\mathbf{r}^2} \sum_{k=1}^{n} \gamma_k^2 \leq \frac{\delta}{2}$. Then, we have*

$$\mathbb{P}(\hat{E}_n) \geq 1 - \delta.$$

Proof. By Lemma F.8, we have $\mathbb{P}(E_n) \geq 1 - \frac{\delta}{2}$. Moreover, by the additional step size assumption and Markov's inequality we deduce that

$$\begin{aligned}
\mathbb{P}(C_n) &= \mathbb{P}(\forall k \leq n \,:\, ||X_{k+1} - X_k|| \leq \frac{\mathbf{r}}{2}) \\
&\geq 1 - \sum_{k=1}^{n} \mathbb{P}(||X_{k+1} - X_k|| > \frac{\mathbf{r}}{2}) \\
&= 1 - \sum_{k=1}^{n} \mathbb{P}(||V_{k+1}(X_k)|| > \frac{\mathbf{r}}{2\gamma_k}) \\
&\geq 1 - \sum_{k=1}^{n} \mathbb{E}[||V_{k+1}(X_k)||^2] \frac{4\gamma_k^2}{\mathbf{r}^2} \\
&\geq 1 - \frac{4C}{\mathbf{r}^2} \sum_{k=1}^{n} \gamma_k^2 \\
&\geq 1 - \frac{\delta}{2}.
\end{aligned}$$

Together we obtain that $\mathbb{P}(\hat{E}_n) = 1 - \mathbb{P}(\hat{E}_n^c) \geq 1 - (\mathbb{P}(E_n^c) + \mathbb{P}(C_n^c)) \geq 1 - \delta$. □

Finally, we are ready to prove the main result in the local setting for the set of local minima $\mathcal{X}^*$.

Proof of Theorem 5.1. (i): Recall the definitions of $\mathcal{U}_1$ and $\mathcal{U}$ above. Then it holds that

$$\Omega_{\mathcal{U}} = \bigcap_{n=1}^{\infty} \Omega_n.$$

Hence, using Lemma F.8 we obtain

$$\mathbb{P}(\Omega_{\mathcal{U}}) = \inf_n \mathbb{P}(\Omega_n) \geq \inf_n \mathbb{P}(\hat{E}_n) \geq 1 - \delta.$$

(ii): We prove that $\tilde{D}_n = D_n \mathbf{1}_{\Omega_n} \in o(1/n^{1-\eta})$, then the claim follows since $\mathbf{1}_{\Omega_{\mathcal{U}}} \leq \mathbf{1}_{\Omega_n}$ almost surely.

From the proof of Lemma F.4 we have

$$\tilde{D}_{n+1} \leq \tilde{D}_n - \gamma_n c \tilde{D}_n^{2\beta} + \gamma_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2} \|V_n\|^2 \mathbf{1}_{\Omega_n}.$$

Hence, taking the conditional expectation gives

$$\begin{aligned}\mathbb{E}[\tilde{D}_{n+1}|\mathcal{F}_n] &\leq \tilde{D}_n - \gamma_n c \tilde{D}_n^{2\beta} + \gamma_n \mathbb{E}[\xi_{n+1}|\mathcal{F}_n]\mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\mathbb{E}[\|V_{n+1}(X_n)\|^2|\mathcal{F}_n]\mathbf{1}_{\Omega_n} \\ &\leq \tilde{D}_n - \gamma_n c \tilde{D}_n^{2\beta} + LC\gamma_n^2,\end{aligned}$$

where we have used that D_n and $\mathbf{1}_{\Omega_n}$ are $\mathcal{F}_n$-measurable and $E[\|V_{n+1}(X_n)\|^2|\mathcal{F}_n] \leq C$ from (ABC) with $A = B = 0$. By our step size choice we can apply Lemma 3.1 to obtain Claim (ii).

(iii): In the following, we again separate between the two cases of β.

$\underline{\beta = \frac{1}{2}}$: We have from Lemma F.4 Equation (45) that

$$\tilde{D}_{n+1} \leq (1 - \gamma_n c^2)\tilde{D}_n + \gamma_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\|V_{n+1}(X_n)\|^2.$$

Taking expectations and multiplying by $(n+1)^{1-\eta}$ leads to

$$\begin{aligned}&\mathbb{E}[\tilde{D}_{n+1}(n+1)^{1-\eta}] \\ &\leq (n+1)^{1-\eta}(1-\gamma_n c^2)\mathbb{E}[\tilde{D}_n] + (n+1)^{1-\eta}\frac{LC\gamma_n^2}{2} \\ &\leq \left(n^{1-\eta} + (1-\eta)n^{-\eta}\right)(1-\gamma_n c^2)\mathbb{E}[\tilde{D}_n] + (n+1)^{1-\eta}\frac{LC\gamma_n^2}{2} \\ &= \left(n^{1-\eta} + (1-\eta)n^{-\eta} - n^{1-\eta}\gamma_n c^2 - (1-\eta)n^{-\eta}\gamma_n c_{x^*}^2\right)\mathbb{E}[\tilde{D}_n] + (n+1)^{1-\eta}\gamma_n^2\frac{LC}{2} \\ &= \left(1 + \frac{1-\eta}{n} - \gamma_n c^2 - \frac{(1-\eta)\gamma_n c^2}{n}\right) n^{1-\eta}\mathbb{E}[\tilde{D}_n] + (n+1)^{1-\eta}\gamma_n^2\frac{LC}{2},\end{aligned}$$

where we used (37) in the first inequality. By our choice of γ_n there exists $\tilde{c} > 0$ and $N > 0$ such that $\gamma_n c^2 - \frac{1-\eta}{n} + \frac{(1-\eta)\gamma_n c^2}{n} \geq \tilde{c}\gamma_n$ for all $n \geq N$. Thus, for all $n \geq N$

$$w_{n+1} \leq (1 - \tilde{c}\gamma_n)\, w_n + (n+1)^{1-\eta}\gamma_n^2\frac{LC}{2},$$

where $w_n = \mathbb{E}[n^{1-\eta}\tilde{D}_n]$. Define $a_n = \tilde{c}\gamma_n$ and $b_n = (n+1)^{1-\eta}\gamma_n^2\frac{LC}{2}$. Since $\gamma_n = \Theta(\frac{1}{n^\theta})$, we have $\sum_n a_n = \tilde{c}\sum_n \gamma_n = \infty$ and

$$\sum_n b_n = \frac{LC}{2}\sum_n (n+1)^{1-\eta}\gamma_n^2 < \infty,$$

by (16) in Lemma 3.1 Hence, we apply Lemma A.3 to prove that $\lim_{n\to\infty} w_n = 0$. By the definition of w_n we have verified that $\mathbb{E}[(f(X_n) - l)\mathbf{1}_{\Omega_{\mathcal{U}}}] \leq \mathbb{E}[\tilde{D}_n] \in o(\frac{1}{n^{1-\eta}})$

$\underline{\beta \in (\frac{1}{2}, 1]}$: From Lemma F.4 Equation (45) we have

$$\tilde{D}_{n+1} \leq (1 - \gamma_n^q c^2)\tilde{D}_n + \tilde{c}\gamma_n^{\frac{2\beta q - 1}{2\beta - 1}} + \gamma_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\|V_{n+1}(X_n)\|^2,$$

for $\tilde{c} = (2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta})c^2$. Next we multiply with $(n+1)^{1-\eta}$ and use (9) to obtain

$$\begin{aligned}&\mathbb{E}[\tilde{D}_{n+1}(n+1)^{1-\eta}] \\ &\leq (n+1)^{1-\eta}(1-\gamma_n^q c^2)\mathbb{E}[\tilde{D}_n] + (n+1)^{1-\eta}\tilde{c}\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}LC\gamma_n^2 \\ &\leq \left(n^{1-\eta} + (1-\eta)n^{-\eta}\right)\left(1-\gamma_n^q c^2\right)\mathbb{E}[\tilde{D}_n] + c_1(n+1)^{1-\eta}(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2) \\ &= \left(n^{1-\eta} + (1-\eta)n^{-\eta} - \gamma_n^q c^2 n^{1-\eta} - (1-\eta)\gamma_n^q c^2 n^{-\eta}\right)\mathbb{E}[\tilde{D}_n] \\ &\quad + c_1(n+1)^{1-\eta}(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2) \\ &= \mathbb{E}[\tilde{D}_n n^{1-\eta}]\left(1 + \frac{1-\eta}{n} - \gamma_n^q c^2 - \frac{(1-\eta)\gamma_n^q c^2}{n}\right) + c_1(n+1)^{1-\eta}(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2),\end{aligned}$$

for some $c_1 > 0$. By our choice of γ_n and as $q \geq 1$, there exists a $c_2 > 0$ and $N > 0$ such that $\gamma_n^q c^2 - \frac{1-\eta}{n} + \frac{(1-\eta)\gamma_n^q c^2}{n} \geq c_2 \gamma_n^q$ for all $n \geq N$. Thus, for $n \geq N$

$$\mathbb{E}[\tilde{D}_{n+1}(n+1)^{1-\eta}] \leq \mathbb{E}[\tilde{D}_n n^{1-\eta}]\,(1 - c_2\gamma_n^q) + c_1(n+1)^{1-\eta}(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2).$$

Define $w_n = \mathbb{E}[\tilde{D}_n n^{1-\eta}]$, $a_n = c_2\gamma_n^q$ and $b_n = c_1(n+1)^{1-\eta}(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2)$. We will again apply Lemma A.3. By the step size choice $\gamma_n = \Theta(\frac{1}{n^\theta})$ we have $\sum_n a_n = c_2 \sum_n \gamma_n^q = \infty$, because $q \leq \frac{1}{\theta}$. Further,

$$\sum_n b_n = c_1 \sum_n (n+1)^{1-\eta}(\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n^2) < \infty,$$

because we choose the auxiliary parameter q as in the proof of Lemma 3.1 where we showed in (16) and (17) that

$$\sum_{n=N}^{\infty} n^{1-\eta-2\theta} < \infty \quad \text{and} \quad \sum_{n=N}^{\infty} n^{1-\eta-\frac{\theta(2\beta q-1)}{2\beta-1}} < \infty$$

All together we deduce that w_n vanishes at infinity. Again, by the definition of w_n we have that $\mathbb{E}[(f(X_n) - l)\mathbf{1}_{\Omega_{\mathcal{U}}}] \leq \mathbb{E}[\tilde{D}_n] \in o(\frac{1}{n^{1-\eta}})$ □

F.2 Proof of Theorem 5.2

Suppose throughout this section that the assumptions in Theorem 5.2 are satisfied.

The proof will be similar to the previous section. Instead of assuring that (X_n) remains close to the set where we could guarantee the unified gradient domination property, it is now sufficient that $f(X_n)$ remains close to f^* by the different definition of gradient domination definition in f^*. This will simplify the proof. Moreover, we may again assume w.l.o.g. the uniform second moment bounds, i.e. $A = B = 0$, instead of the more general (ABC) condition by the same argument as above but on the level sets.

Recall the notation

$$\mathcal{B}_r^* = \{x \in \mathbb{R}^d \,:\, f(x) - f^* \leq r\}.$$

and let $r > 0$ be the radius of the gradient domination property in f^*, then there exists $\epsilon > 0$, such that $2\epsilon + \sqrt{\epsilon} < r$, i.e

$$\mathcal{U} := \mathcal{B}_{2\epsilon+\sqrt{\epsilon}}^* \subset \mathcal{B}_r^*. \tag{41}$$

Moreover, we define the set

$$\mathcal{U}_1 := \mathcal{B}_{\frac{\epsilon}{2}}^* \tag{42}$$

and the measurable subsets

$$\Omega_n = \{X_k \in \mathcal{U}, \text{ for all } k \leq n\}$$

in $(\Omega, \mathcal{F}, \mathbb{P})$.

The proof of Theorem 5.2 is again based on a series of auxiliary lemmas. The goal of these is to prove that with high probability we do not leave the gradient dominated region, i.e. Claim (i) in Theorem 5.2.

In the following, we fix the notation $D_n := f(X_n) - f^*$ and $\tilde{D}_n := D_n \mathbf{1}_{\Omega_n}$ and obtain the parallel result to Lemma F.4.

Lemma F.9. *If $\beta = \frac{1}{2}$, it holds that*

$$\tilde{D}_{n+1} \leq (1 - \gamma_n c^2)\tilde{D}_n + \gamma_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2, \tag{43}$$

$$\leq \tilde{D}_1 \prod_{k=1}^{n}(1 - \gamma_k c^2) + \sum_{k=1}^{n} \gamma_k \xi_k \mathbf{1}_{\Omega_k} + \frac{L}{2}\sum_{k=1}^{n} \gamma_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_k}. \tag{44}$$

If $\beta \in (\frac{1}{2}, 1]$, for any $1 \leq q \leq 2$, it holds that

$$\tilde{D}_{n+1} \leq (1-\gamma_n^q c^2)\tilde{D}_n + (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2\gamma_n^{\frac{2\beta q-1}{2\beta-1}} + \gamma_n\xi_n\mathbf{1}_{\Omega_n} + \frac{L\gamma_n^2}{2}\|V_{n+1}(X_n)\|^2 \quad (45)$$

$$\leq \tilde{D}_1\prod_{k=1}^n(1-\gamma_k^q c^2) + \tilde{c}\sum_{k=1}^n\gamma_k^{\frac{2\beta q-1}{2\beta-1}} + \sum_{k=1}^n\gamma_k\xi_k\mathbf{1}_{\Omega_k} + \frac{L}{2}\sum_{k=1}^n\gamma_k^2\|V_{k+1}(X_k)\|^2, \quad (46)$$

for $\tilde{c} = (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2$.

Proof. The proof follows line for line as in Lemma F.4 by replacing l with f^* and taking the different definition of $\tilde{D}_n$ and Ω_n into account. □

We continue as in the previous section:
For $\beta > \frac{1}{2}$ we know from the proof of Lemma 3.1 that we can choose the auxiliary parameter q from the previous lemma in such a way, that $\sum_{n=1}^\infty n^{1-\eta}\gamma_n^{\frac{2\beta q-1}{2\beta-1}}$ is convergent for all $\eta \in (\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1)$ (Condition (iii) to apply Lemma A.2). As $\eta < 1$, it follows that $\sum_{n=1}^\infty \gamma_n^{\frac{2\beta q-1}{2\beta-1}} < \infty$ holds true for all these choices of q. Now define

$$M_n = \sum_{k=1}^n \gamma_k\xi_{k+1}\mathbf{1}_{\Omega_k} \quad \text{and} \quad S_n = \frac{L}{2}\sum_{k=1}^n\gamma_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_k}.$$

Then, $(M_n)_{n\in\mathbb{N}}$ is a $(\mathcal{F}_{n-1})$-martingale with zero mean and $(S_n)_{n\in\mathbb{N}}$ is a $(\mathcal{F}_{n-1})$-sub-martingale by Assumption 2.4. Note that by the choice of γ_n we have that $\sum_n \gamma_n^2 < \infty$ and hence $\mathbb{E}[S_n] < \infty$ for all $n \in \mathbb{N}$. Next, define $R_n = M_n^2 + S_n$ for every $n \in \mathbb{N}$. Moreover, let

$$E_n = \{R_k < \epsilon \text{ for all } k \leq n\}.$$

which is an $\mathcal{F}_n$-measurable event on $(\Omega, \mathcal{F}, \mathbb{P})$. We define $R_0 = 0$ such that $E_0 = \Omega$.

With these definitions we can directly prove a parallel result to Lemma F.6 without the auxiliary result in Lemma F.5.

Lemma F.10. *For $\beta \in (\frac{1}{2}, 1]$ let $\gamma_n \leq \gamma_1$ be sufficiently small such that $\sum_{n=1}^\infty \gamma_n^{\frac{2\beta q-1}{2\beta-1}} < \frac{\epsilon}{2\tilde{c}}$, and for $\beta = \frac{1}{2}$ let $\gamma_1 > 0$ be arbitrary. Furthermore, assume that the initial $X_1 \in \mathcal{U}_1 = \{x : f(x) - f(x^*) \leq \frac{\epsilon}{2}\}$ almost surely. Then,*

a) $E_{n+1} \subset E_n$ and $\Omega_{n+1} \subset \Omega_n$

b) $E_n \subset \Omega_{n+1}$

c) Define the events $\tilde{E}_n = E_{n-1} \setminus E_n = E_{n-1} \cup \{R_n \geq \epsilon\}$. Then, for $\tilde{R}_n = R_n\mathbf{1}_{E_{n-1}}$, there exists a $\tilde{C} > 0$ such that

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_{n-1}] + \gamma_n^2[G^2C^2 + G^2 + C] - \epsilon\mathbb{P}(\tilde{E}_{n-1}).$$

Proof. a) Follows by definition.

b) Note that $E_0 = \Omega = \Omega_1$ because $X_1 \in \mathcal{U}_1 = \Omega_1$ almost surely. We prove the assertion by induction. Let $\omega \in E_n$. Since $E_n \subset E_{n-1} \subset \Omega_n$ by induction assumption, we have $\omega \in \Omega_n$ and thus $\omega \in \Omega_k$ for all $k \leq n$. It remains to show that $X_{n+1}(\omega) \in \mathcal{U}$ to prove that $\omega \in \Omega_{n+1}$. We separate both cases for β:

$\underline{\beta = \frac{1}{2}}$: The inequality (43) and the induction hypothesis yield

$$\begin{aligned} D_{n+1}(\omega) &\leq D_1(\omega)\prod_{k=1}^n(1-\gamma_k c) + \sum_{k=1}^n\gamma_k\xi_{k+1}(\omega)\mathbf{1}_{\Omega_k}(\omega) + \frac{L}{2}\sum_{k=1}^n\gamma_k^2\|V_{k+1}(X_k(\omega))\|^2 \\ &\leq \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\ &\leq 2\epsilon + \sqrt{\epsilon}. \end{aligned}$$

Hence, $X_{n+1}(\omega) \in \mathcal{U}$ by definition of $\mathcal{U}$.

$\underline{\beta \in (\frac{1}{2}, 1]}$: Similarly, we obtain from (45)

$$\begin{aligned} D_{n+1}(\omega) &\leq D_1(\omega) \prod_{k=1}^{n} (1 - \gamma_k^q c) + \tilde{c} \sum_{k=1}^{n} \gamma_k^{\frac{2\beta q - 1}{2\beta - 1}} + \sum_{k=1}^{n} \gamma_k \xi_{k+1}(\omega) \mathbf{1}_{\Omega_k}(\omega) + \frac{L}{2} \sum_{k=0}^{n} \gamma_k^2 \|V_{k+1}(X_k(\omega))\|^2 \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\ &\leq 2\epsilon + \sqrt{\epsilon}, \end{aligned}$$

where we used that $\prod_{k=1}^{n}(1 - \gamma_k^q c^*) < 1$. Hence, it holds again that $X_{n+1}(\omega) \in \mathcal{U}$.

This prove that $\omega \in \Omega_{n+1}$ and closes the induction.

c) Follows line by line as in Lemma F.6 part c). □

Lemma F.11. *Let $\delta > 0$ be a tolerance level and $\gamma_n \leq \gamma_1$ be sufficiently small such that $\sum_{n=1}^{\infty} \gamma_n^2 < \frac{\delta\epsilon}{2(G^2C^2+G^2+C)}$ and the condition in Lemma F.10 is fulfilled. Then, we have*

$$\mathbb{P}(E_n) \geq 1 - \delta.$$

Proof. Line by line as in Lemma F.7. □

Finally, we are ready to prove the main result in the local setting for f^*.

Proof of Theorem 5.2. (i): Recall the definition of $\mathcal{U}_1$ and $\mathcal{U}$ above. Then it holds that

$$\Omega_{\mathcal{U}} = \bigcap_{n=1}^{\infty} \Omega_n.$$

Hence, using Lemma F.11 we obtain

$$\mathbb{P}(\Omega_{\mathcal{U}}) = \inf_n \mathbb{P}(\Omega_n) \geq \inf_n \mathbb{P}(E_n) \geq 1 - \delta.$$

The proof of Claim (ii) and (iii) follows line by line as in the proof of Theorem 5.1 by replacing l with f^* and taking the different definitions of D_n, $\tilde{D}_n$ and Ω_n into account.

□

G Proofs of Section 7

Proof of Proposition 7.1. We consider the cases $\lambda = 0$ and $\lambda > 0$ separately.
Case $\lambda = 0$:
Define the optimal reward gap in every state $s \in \mathcal{S}$ by

$$\Delta^*(s) = Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) > 0,$$

where $a^*(s)$ denotes the best possible action in state s and $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the optimal Q-function defined by $Q^*(s, a) = \mathbb{E}_\mu^{\pi^*}[\sum_{t=0}^{\infty} \rho^{-t} r(S_t, A_t) | A_0 = a]$. W.l.o.g. we assume that $a^*(s)$ is unique. Similary let $Q^{\pi_w}(s, a) = \mathbb{E}_\mu^{\pi_w}[\sum_{t=0}^{\infty} \rho^{-t} r(S_t, A_t) | A_0 = a]$ be the Q-function for policy π_w.

For any $0 < \alpha < 1$ choose $r = \min_{s \in \mathcal{S}} \mu(s) \min_{s \in \mathcal{S}} \Delta^*(s)(1 - \alpha)$ and assume that $w \in \mathcal{B}_r^*$, i.e. $V^*(\mu) - V^{\pi_w}(\mu) \leq r$. Then, we have for every $s \in \mathcal{S}$ that

$$V^*(\delta_s) - V^{\pi_w}(\delta_s) \leq \frac{r}{\min_{s \in \mathcal{S}} \mu(s)}.$$

It follows for every $s \in \mathcal{S}$ that

$$\begin{aligned}
\frac{r}{\min_{s\in\mathcal{S}} \mu(s)} &\geq V^*(\delta_s) - V^{\pi_w}(\delta_s) \\
&= Q^*(s, a^*(s)) - \sum_{a\in\mathcal{A}_s} \pi_w(a|s) Q^{\pi_w}(s,a) \\
&\geq \sum_{a\in\mathcal{A}_s} \pi_w(a|s)(Q^*(s,a^*(s)) - Q^*(s,a)) \\
&= \sum_{a\neq a^*(s)} \pi_w(a|s)(Q^*(s,a^*(s)) - Q^*(s,a)) \\
&\geq (1 - \pi_w(a^*(s)|s) \min_s \Delta^*(s).
\end{aligned}$$

Rearranging results in

$$\pi_w(a^*(s)|s) \geq 1 - \frac{r}{\min_{s\in\mathcal{S}} \mu(s) \min_{s\in\mathcal{S}} \Delta^*(s)} = \alpha.$$

Hence, for all $w \in \mathcal{B}_r^*$ we can bound $c(w)$ by

$$c(w) \geq \frac{\alpha}{\sqrt{|\mathcal{S}|}(1-\rho)} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} > 0.$$

Thus, setting $c = \frac{\alpha}{\sqrt{|\mathcal{S}|}(1-\rho)} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}$ proves the claim.

Case $\lambda > 0$:
For any $\alpha \in (0,1)$ choose $r = \alpha^2 \exp\left(\frac{-1}{(1-\rho)\lambda}\right)^2 \frac{\lambda \min_s \mu(s)}{2\ln 2}$ and assume that $w \in \mathcal{B}_{r,\lambda}^*$. By [10, Lem. 12] we have

$$\begin{aligned}
|\pi_w(a|s) - \pi^*(a|s)| &\leq \sqrt{\frac{2(V_\lambda^*(\mu) - V_\lambda^{\pi_w}(\mu))\ln 2}{\lambda \min_s \mu(s)}} \\
&\leq \sqrt{\frac{2r\ln 2}{\lambda \min_s \mu(s)}} \\
&= \alpha \exp\left(\frac{-1}{(1-\rho)\lambda}\right) \\
&\leq \alpha \min_{s,a} \pi^*(a|s).
\end{aligned}$$

where the last inequality is due to [37, Thm. 1]. It follows directly that

$$\min_{s,a} \pi_w(s,a) \geq (1-\alpha) \min_{s,a} \pi^*(s,a) > 0.$$

Hence, we can bound $c(w)$ uniformly for all $w \in \mathcal{B}_{r,\lambda}^*$ by

$$c(w) \geq \frac{2\lambda}{|\mathcal{S}|(1-\rho)} \min_s \mu(s)(1-\alpha)^2 \min_{s,a} \pi^*(a|s)^2 \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}. \tag{47}$$

Thus, setting $c = \frac{2\lambda}{|\mathcal{S}|(1-\rho)} \min_s \mu(s)(1-\alpha)^2 \min_{s,a} \pi^*(a|s)^2 \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}$ proves the claim. □

Remark G.1. It is noteworthy, that we have multiple choices of r and c depending on $\alpha \in (0,1)$.