RANDOMIZED COORDINATE GRADIENT DESCENT ALMOST SURELY ESCAPES STRICT SADDLE POINTS

ZIANG CHEN, YINGZHOU LI, AND ZIHAO LI

Abstract. We analyze the behavior of randomized coordinate gradient descent for nonconvex optimization, proving that under standard assumptions, the iterates almost surely escape strict saddle points. By formulating the method as a nonlinear random dynamical system and characterizing neighborhoods of critical points, we establish this result through the center-stable manifold theorem.

1. Introduction

Randomized coordinate gradient descent is a widely used optimization method in scientific computing, particularly for large-scale problems. In this paper, we analyze the escape behavior of randomized coordinate gradient descent from strict saddle points when applied to smooth, nonconvex optimization problems of the form

$$\min_{x\in\mathbb{R}^d} f(x). \tag{1.1}$$

Specifically, we study coordinate gradient descent with a randomized coordinate selection rule and a fixed step size, as detailed in Algorithm 1.

Algorithm 1 Randomized coordinate gradient descent

Initialization: $x_0 \in \mathbb{R}^d$, $t = 0$.
while not convergent **do**
  Sample a coordinate i_t uniformly random from $\{1, 2, \ldots, d\}$.
  $x_{t+1} \leftarrow x_t - \alpha e_{i_t} \partial_{i_t} f(x_t)$.
  $t \leftarrow t + 1$.
end while

The main contribution of this paper, presented in Theorem 1, establishes that under standard assumptions—namely, the smoothness of the objective function f, the boundedness of its Hessian, and the non-degeneracy of its strict saddle points—the set of all pairs (x_0, ω) converging to a strict saddle point has measure zero. Here, x_0 denotes the initial point, and ω

Date: August 12, 2025.

The work of ZC is supported in part by the National Science Foundation via grant DMS-2509011. The work of YL and ZL is supported by the National Natural Science Foundation of China (NSFC) under grant numbers 12271109, the Science and Technology Commission of Shanghai Municipality (STCSM) under grant numbers 22TQ017 and 24DP2600100, and the Shanghai Institute for Mathematics and Interdisciplinary Sciences (SIMIS) under grant number SIMIS-ID-2024-(CN). We thank Jianfeng Lu for helpful discussions.

represents the sequence of sampled coordinates. Furthermore, under the additional assumption that all critical points are isolated, the algorithm is guaranteed to converge globally to a critical point without negative Hessian eigenvalues. The proof is based on a random dynamical systems perspective, following [8], and the application of the center-stable manifold theorem that rigorously characterizes the local behavior of the algorithm near saddle points.

1.1. **Related works.** Coordinate gradient descent (CGD) is a widely used optimization technique, particularly well-suited for modern large-scale problems [33,39]. The method's popularity stems from its computational efficiency and scalability, making it applicable across diverse domains [36,37,40]. Notable applications include symmetric eigenvalue problems [3], root-finding algorithms [35], quantum circuit optimization [10], and high-dimensional statistics, where it is implemented in packages such as SparseNet [25].

Coordinate gradient descent (CGD) methods fall into two main categories: deterministic approaches that follow fixed coordinate selection rules, and randomized methods that employ stochastic sampling. Among deterministic variants, the cyclic strategy updates coordinates in a fixed periodic order, while the greedy approach (Gauss–Southwell rule) selects at each iteration the coordinate offering the steepest descent. For strongly convex objectives, cyclic CGD's convergence properties are well-established [39], while the greedy variant's behavior has been characterized in [27]. These guarantees extend to general convex functions under cyclic strategies [4,32]. In nonconvex optimization, significant attention has focused on saddle point avoidance. Work in [18,19] demonstrates that deterministic first-order methods, including cyclic CGD, almost surely escape strict saddle points. The Kurdyka–Łojasiewicz (KŁ) framework [2,5] further provides convergence guarantees for coordinate methods in nonconvex and nonsmooth settings. Extensions to variance-reduced or manifold-constrained methods are studied in [7,29].

Randomized coordinate gradient descent (RCGD) employs various sampling strategies, including uniform/non-uniform random selection and random-permutation approaches [17,38]. For convex optimization, [26] establishes sublinear convergence to the minimum in expectation ($\mathbb{E}(f(x_t))$) for general convex functions, with linear convergence under strong convexity. Further convex convergence results appear in [22, 23, 39]. The nonconvex case presents distinct challenges: unlike randomized gradient descent that escapes strict saddle points through additive Gaussian noise [11, 14, 15], RCGD's coordinate-wise stochasticity inherently limits this capability. While introducing additive perturbations might help, such modifications risk compromising the algorithm's coordinate structure, necessitating alternative analytical approaches. Relevant nonconvex convergence analyses for RCGD can be found in [8,21].

A fruitful research direction interprets iterative algorithms as discrete-time dynamical systems [18, 19, 28]. In deterministic settings, the center-stable manifold theorem [34] provides a powerful tool for analyzing system behavior near critical points, particularly through the lens of the invariant manifold theorem. This perspective naturally extends to randomized algorithms via random dynamical systems theory. Recent work by Liu and Yuan [24] demonstrates this approach, applying the invariant manifold theorem to prove saddle point avoidance for various stochastic methods. Chen et al. [8] developed quantitative finite-block analyses of random dynamical systems to establish convergence properties near saddle points.

The theory of random dynamical systems offers powerful analytical tools, including random stable, unstable, and center manifolds [1, 6, 13, 20, 30, 31]. These constructs capture sample-dependent geometric structures in the state space and, under generic conditions, typically exhibit low dimensionality. This characteristic makes them particularly well-suited for analyzing convergence probabilities to strict saddle points.

The work most closely related to ours is [8], from which our approach differs in two key aspects. First, we eliminate a technical assumption on the objective function while adopting a more practical fixed-stepsize scheme. Second, whereas [8] employs linearized finite-block analysis, our work directly applies invariant manifold theory to establish convergence properties - an approach that yields both stronger theoretical guarantees and greater robustness in the analysis.

1.2. **Main results.** We first set up the notations before stating our main theorem. For each $t \in \mathbb{N}$, denote $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ the usual probability space for the distribution $\mathcal{U}_{\{1,2,\dots,n\}}$, where $\mathcal{U}_{\{1,2,\dots,n\}}$ are the uniform distributions on the set $\{1, 2, \dots, n\}$, which is associated to the t-th iteration of Algorithm 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the product probability space of all $(\Omega_t, \Sigma_t, \mathbb{P}_t)$, $t \in \mathbb{N}$, and a sample $\omega \in \Omega$ can be represented as a sequence $(i_0, i_1, i_2, \dots)$, where $i_t \in \Omega_t$. It is clear that the iterates generated by Algorithm 1 is sample-dependent, and we would use the notation $x_t = x_t(\omega)$ to clarify the dependence if necessary.

Moreover, the dynamics of Algorithm 1 depend on the initialization $x_0 \in \mathbb{R}^d$ and the sample $\omega \in \Omega$. We set $\Theta = \mathbb{R}^d \times \Omega$ that is equipped with a product measure

$$\mu = \textsc{Leb} \times \mathbb{P},$$

where Leb is the Lebesgue measure on $\mathbb{R}^d$.

The primary objects of interest in our analysis are strict saddle points, and the collection of all strict saddle points is defined as

$$\mathrm{Crit}_s(f) := \left\{ x \in \mathbb{R}^d : \nabla f(x) = 0,\ \lambda_{\min}(\nabla^2 f(x)) < 0 \right\},$$

where $\lambda_{\min}(\nabla^2 f(x))$ is the smallest eigenvalue of the Hessian matrix $\nabla^2 f(x)$ and we use the subscript s to emphasize that it is strict.

Our main results are based on some standard and generic assumptions on the objective function f. The first assumption is that f is two times continuously differentiable with a bounded and locally Lipschitz continuous Hessian.

Assumption 1. *The function $f \in \mathcal{C}^2(\mathbb{R}^d)$ and the Hessian $\nabla^2 f$ is uniformly bounded, i.e., there exists $M > 0$ such that $\left\|\nabla^2 f(x)\right\| \le M$ for all $x \in \mathbb{R}^d$. Moreover, for every $x^* \in \mathrm{Crit}_s(f)$, there exists a neighborhood $N(x^*)$ of x^* where $\nabla^2 f$ is locally Lipschitz continuous, i.e., $\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \le L \left\|x - y\right\|$ for all $x, y \in N(x^*)$, where $L = L(x^*) > 0$ is a constant depending on x^*.*

Throughout this paper, $\|\cdot\|$ always represents the ℓ_2-norm for vectors or its induced matrix norm. Our next assumption is on the non-degeneracy of the strict saddle points.

Assumption 2. *For every $x^* \in \mathrm{Crit}_s(f)$, $\nabla^2 f(x^*)$ is non-degenerate, i.e., x^* is a non-degenerate critical point of f in the sense that any eigenvalue of $\nabla^2 f(x^*)$ is nonzero.*

The main theorem of this paper states that the set of all (x_0, ω) with $x_t(\omega)$ converging to a strict saddle point in $\mathrm{Crit}_s(f)$ is of measure zero. In particular, consider

$$\Theta(x^*) = \left\{ (x_0, \omega) \in \Theta : \lim_{t \to +\infty} x_t(\omega) = x^* \right\},$$

for each $x^* \in \mathrm{Crit}_s(f)$, and

$$\Theta(\mathrm{Crit}_s(f)) = \bigcup_{x^* \in \mathrm{Crit}_s(f)} \Theta(x^*).$$

Theorem 1. *Suppose that Assumptions 1 and 2 hold and that step $0 < \alpha < 1/M$. It holds that*

$$\mu(\Theta(\mathrm{Crit}_s(f))) = 0.$$

Moreover, with some additional but still standard assumptions, we prove that $x_t(\omega)$ always converges to a critical point without negative Hessian eigenvalues, unless (x_0, ω) is located in a μ-null set. More specifically, denote by

$$\mathrm{Crit}(f) := \{x \in \mathbb{R}^d : \nabla f(x) = 0\},$$

the set of all critical points of f, and we have the following corollary.

Corollary 2. *Suppose that Assumptions 1 and 2 hold and that stepsize α satisfies $0 < \alpha < 1/M$. Suppose in addition that every $x^* \in \mathrm{Crit}(f)$ is an isolated critical point and that the level set $L_c(f) := \{x \in \mathbb{R}^d : f(x) \le c\}$ is bounded for all $c \in \mathbb{R}$. Then there exists $\widehat{\Theta} \subseteq \Theta$ with $\mu(\Theta \setminus \widehat{\Theta}) = 0$, such that $\{x_t(\omega)\}_{t \in \mathbb{N}}$ is convergent with the limit in $\mathrm{Crit}(f) \setminus \mathrm{Crit}_s(f)$ for any $(x_0, \omega) \in \widehat{\Theta}$.*

1.3. **Organization.** The rest of this paper is devoted to the proofs of the results stated above and is organized as follows. In Section 2, we present some preliminaries on random dynamical systems, which are the foundations of our analysis. Section 3 outlines the proof with a comparison to [8]. All technical lemmas and propositions are deferred to Section 4.

2. Preliminaries on random dynamical systems

The dynamics of Algorithm 1 can be rigorously characterized using the notion of random dynamical systems. In particular, given the initilization x_0, the trajectory $\{x_t\}_{t \in \mathbb{N}}$ is fully determined by a random sample ω. Therefore, analytical tools developed for random dynamical systems would be useful for analyzing the behavior of Algorithm 1. This section briefly reviews some fundamental results in random dynamical systems. For a more detailed introduction, we refer the readers to [1, 13, 20] and references therein.

2.1. **Definition of random dynamical system.** Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathbb{T}$ denote a semigroup equipped with its Borel σ-algebra $\mathcal{B}(\mathbb{T})$, playing the role of time, with the convention that $0 \in \mathbb{T}$. Here, $(\Omega, \mathcal{F}, \mathbb{P})$ represents a general probability space, not necessary the one associated to Algorithm 1. Common choices for $\mathbb{T}$ include $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{R}$. The random dynamical system is defined as follows.

Definition 2.1 (Metric dynamical system)**.** *A metric dynamical system on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a family of maps $\{\theta(t) : \Omega \to \Omega\}_{t\in\mathbb{T}}$ satisfying that*

(i) The mapping $\mathbb{T} \times \Omega \to \Omega,\ (t, \omega) \mapsto \theta(t)\omega$ is measurable;
(ii) It holds that $\theta(0) = \mathrm{Id}_\Omega$ and $\theta(t + s) = \theta(t) \circ \theta(s),\ \forall\ s, t \in \mathbb{T}$;
(iii) $\theta(t)$ is $\mathbb{P}$-preserving for any $t \in \mathbb{T}$, where we say a map $\theta : \Omega \to \Omega$ is $\mathbb{P}$-preserving if

$$\mathbb{P}(\theta^{-1}B) = \mathbb{P}(B), \quad \forall\ B \in \mathcal{F}.$$

Definition 2.2 (Random dynamical system)**.** *Let $(X, \mathcal{F}_X)$ be a measurable space and let $\{\theta(t) : \Omega \to \Omega\}_{t\in\mathbb{T}}$ be a metric dynamical system on $(\Omega, \mathcal{F}, \mathbb{P})$. Then a random dynamical system on $(X, \mathcal{F}_X)$ over $\{\theta(t)\}_{t\in\mathbb{T}}$ is a measurable map*

$$\begin{aligned}\varphi : \mathbb{T} \times \Omega \times X \to &\quad X,\\ (t, \omega, x) &\mapsto \varphi(t, \omega, x),\end{aligned}$$

satisfying the following cocycle property: for any $\omega \in \Omega$, $x \in X$, and $s, t \in \mathbb{T}$, it holds that

$$\varphi(0, \omega, x) = x,$$

and that

$$\varphi(t + s, \omega, x) = \varphi(t, \theta(s)\omega, \varphi(s, \omega, x)). \tag{2.1}$$

The cocycle property (2.1) plays a fundamental role in the theory of random dynamical systems. Intuitively, it means that if the system evolves for time s and reaches the state $x_s = \varphi(s, \omega, x)$, then continuing the evolution is equivalent to restarting the system from x_s with a shifted random sample $\theta(s)\omega$. In other words, the metric dynamical system $\theta(s)$ maps ω to another sample controlling the dynamics starting at time s. The map $\varphi(t, \omega, \cdot)$ acts on the space X; for convenience and with a slight abuse of notation, we also denote this map by $\varphi(t, \omega)$, and write $\varphi(t, \omega)x$ in place of $\varphi(t, \omega, x)$. Under this notation, the cocycle property (2.1) takes the form:

$$\varphi(t + s, \omega) = \varphi(t, \theta(s)\omega) \circ \varphi(s, \omega).$$

In this work, we restrict our attention to the two-sided discrete-time case $\mathbb{T} = \mathbb{Z}$, which corresponds to the dynamics of Algorithm 1 and its inverse system. The inversion is included due to a technical reason stated later. We only consider the metric dynamical system $\theta(t) = \theta^t$, with $\theta : \Omega \to \Omega$ being a $\mathbb{P}$-preserving map and θ^t denotes the t-fold composition of θ. We also assume that the state space is $X = \mathbb{R}^d$ throughout this paper.

2.2. **Multiplicative ergodic theorem.** Let $A : \Omega \to \mathrm{GL}(d, \mathbb{R})$ be a measurable map. We consider a linear random dynamical system defined by

$$x_t = \Phi(t, \omega)x_0,$$

where the $\Phi(t, \omega)$ is a product of random matrices

$$\Phi(t,\omega) = \begin{cases} A(\theta^{t-1}\omega)\cdots A(\theta\omega)A(\omega), & \text{if } t > 0, \\ I, & \text{if } t = 0, \\ A(\theta^{t}\omega)^{-1}\cdots A(\theta^{-2}\omega)^{-1}A(\theta^{-1}\omega)^{-1}, & \text{if } t < 0. \end{cases}$$

In this setting, the evolution of the system is well characterized by the celebrated multiplicative ergodic theorem (also known as Oseledets' theorem), which we state as Theorem 2.3. We use the notation Φ for this linear system, reserving φ for the nonlinear dynamics.

Theorem 2.3 (Multiplicative ergodic theorem, [1, Theorem 3.4.11])**.** *Suppose that*

$$(\log \|A(\cdot)\|)_+,\ (\log \|A(\cdot)^{-1}\|)_+ \in L^1(\Omega, \mathcal{F}, \mathbb{P}), \tag{2.2}$$

where we have used the short-hand $a_+ := \max\{a, 0\}$. Then there exists an θ-invariant $\widetilde{\Omega} \in \mathcal{F}$ with $\mathbb{P}(\widetilde{\Omega}) = 1$, such that the followings hold for any $\omega \in \widetilde{\Omega}$:

(i) It holds that the limit

$$\Lambda(\omega) = \lim_{t\to+\infty} \left(\Phi(t,\omega)^\top \Phi(t,\omega)\right)^{1/2t} \tag{2.3}$$

exists and is a positive definite matrix. Here $\Phi(t,\omega)^\top$ denotes the transposition of the matrix (as $\Phi(t,\omega)$ is a linear map on X).

(ii) Suppose $\Lambda(\omega)$ has $p(\omega)$ distinct eigenvalues, which are ordered as $e^{\lambda_1(\omega)} > e^{\lambda_2(\omega)} > \cdots > e^{\lambda_{p(\omega)}(\omega)} > 0$, and let $V_i(\omega)$ be the eigenspace associated with $e^{\lambda_i(\omega)}$ with dimension $d_i(\omega)$, for $i = 1, 2, \ldots, p(\omega)$. Then the functions $p(\cdot)$, $\lambda_i(\cdot)$, and $d_i(\cdot)$, $i = 1, 2, \ldots, p(\cdot)$, are all measurable and θ-invariant on $\widetilde{\Omega}$.

(iii) There exists a splitting

$$\mathbb{R}^d = E_1(\omega) \oplus E_2(\omega) \oplus \cdots \oplus E_{p(\omega)}(\omega) \tag{2.4}$$

of $\mathbb{R}^d$ into random subspaces $E_i(\omega)$ with $\dim E_i(\omega) = d_i(\omega)$, such that

$$\lim_{t\to\pm\infty} \frac{1}{t} \log \|\Phi(t,\omega)x_0\| = \lambda_i(\omega) \iff x_0 \in E_i(\omega) \setminus \{0\}, \tag{2.5}$$

and

$$A(\omega)E_i(\omega) = E_i(\theta\omega).$$

(iv) When $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$ is ergodic, i.e., every $B \in \mathcal{F}$ with $\theta^{-1}B = B$ satisfies $\mathbb{P}(B) = 0$ or $\mathbb{P}(B) = 1$, the functions $p(\cdot)$, $\lambda_i(\cdot)$, and $d_i(\cdot)$, $i = 1, 2, \ldots, p(\cdot)$, are constant on $\widetilde{\Omega}$.

In the rest of Section 2, we slightly abuse the notation by denoting $\widetilde{\Omega}$ by Ω for simplicity. In other words, we assume that all statements in Theorem 2.3 are true for every ω, with the null set $\Omega \setminus \widetilde{\Omega}$ being removed in advance.

In Theorem 2.3, $\lambda_i(\omega)$, $i = 1, 2, \ldots, p_i(\omega)$ are the so-called Lyapunov exponents, $E_i(\omega)$, $i = 1, 2, \ldots, p_i(\omega)$ are named as the Oseledets subspaces, and the splitting (2.4) is termed Oseledets

splitting. The limiting behavior of $x_t(\omega)$ is characterized precisely by (2.5). In particular, $\|x_t(\omega)\|$ grows exponentially as $t \to +\infty$ if x_0 has nontrivial projection onto at least one $E_i(\omega)$ with $\lambda_i(\omega) > 0$, along the spliting (2.4). Otherwise, $\|x_t(\omega)\|$ either decays exponentially or has sub-exponential behavior as $t \to +\infty$. This observation motivates the following decomposition:

$$\mathbb{R}^d = E_u(\omega) \oplus E_{cs}(\omega),$$

where

$$E_u(\omega) = \bigoplus_{\lambda_i > 0} E_i(\omega), \quad \text{and} \quad E_{cs}(\omega) = \bigoplus_{\lambda_i \leq 0} E_i(\omega).$$

We call $E_u(\omega)$ the unstable Oseledets subspace and $E_{cs}(\omega)$ the center-stable Oseledets subspace. It can be seen that a necessary condition that $x_t(\omega)$ converges to 0, or even stays bounded, as $t \to +\infty$ is that $x_0(\omega) \in E_{cs}(\omega)$. Though we mainly focus on $t \to +\infty$ as it is of particular interest in the context of Algorithm 1, we remark that similar limit behavior is also true for $t \to -\infty$ since (2.5) states a two-sided limit.

2.3. **Center-stable manifold theorem.** For a nonlinear random dynamical system $\varphi(t, \omega, x)$, we say that x^* is an equilibrium if

$$\varphi(t, \omega, x^*) = x^*, \quad \forall\, t \in \mathbb{Z},\ \omega \in \Omega.$$

The behavior of the system near x^* can be approximated by its linearization at x^*. More specifically, there is a local manifold, termed the center-stable manifold, that is tangent to the center-stable Oseledets subspace associated with the linearized system, such that a necessary condition that x_t converges to x^* is that x_0 lies on the center-stable manifold. To state the theorem rigorously, we write

$$\varphi(t, \omega, x) = \Phi(t, \omega)x + F(t, \omega, x), \tag{2.6}$$

where the equilibrium is assumed to be $x^* = 0$ without loss of generality and $\Phi(t, w)$ is the linearized system given by $\Phi(t, w) = D_x\varphi(t, \omega, 0)$, and make the following assumption.

Definition 2.4 (Tempered random variable)**.** *We say that a random variable $R : \Omega \to (0, \infty)$ is tempered from above if*

$$\lim_{t \to \pm\infty} \frac{1}{t} \left(\log R(\theta^t \omega)\right)_+ = 0$$

holds almost surely, and a random variable $R : \Omega \to (0, \infty)$ is called tempered from below if $1/R$ is tempered from above. Moreover, a random variable is called tempered if it is both tempered from above and tempered from below.

Assumption 2.5. *We assume that the random dynamical system $\varphi(t, \omega, x)$ in (2.6) with an equilibrium $x^* = 0$ satisfies the following conditions:*

(i) $\varphi(t, \omega, x)$ is a $\mathcal{C}^1$ random dynamical system, and $\Phi(t, \omega) = D_x\varphi(t, \omega, 0)$ satisfies the condition (2.2).

(ii) There is a ball $U(\omega) = \{x \in \mathbb{R}^d : \|x\| < \rho_0(\omega)\}$ *where* $\rho_0 : \Omega \to (0, \infty)$ *is tempered from below such that*

$$\begin{aligned} &\sup_{x \in U(\omega)} \left\| D_x^k F(1, \omega, x) \right\| \leq B_k(\omega), \quad \forall\ 0 \leq k \leq 1,\ \omega \in \Omega, \\ &\|D_x F(1, \omega, x)\| \leq B(\omega) \|x\|, \quad \forall\ x \in U(\omega),\ \omega \in \Omega, \end{aligned} \tag{2.7}$$

where $B, B_k : \Omega \to (0, +\infty)$ *is tempered from above for* $k = 0, 1$.

Remark 2.6. *Assumption 2.5 presents a minor modification of the conditions employed in* [13, 20] *according to their proofs. This assumption aligns with* [1, Lemma 7.5.11].

Now we can state the center-stable manifold theorem.

Theorem 2.7 (Center-stable manifold theorem [1,13,20])**.** *Let* $\varphi(t, \omega, x)$ *be a random dynamical system as in* (2.6) *with an equilibrium* $x^* = 0$ *and satisfy Assumption 2.5. Then there exist a tempered variable* $\rho : \Omega \to (0, +\infty)$ *and a measurable function*

$$h^{cs} : E_{cs}(\omega) \times \Omega \to E_u(\omega),$$

such that the following are satisfied:

(i) $h^{cs}(y, \omega)$ *is measurable in* (y, ω) *and is* $\mathcal{C}^1$ *in* y *with*

$$\operatorname{Lip} h^{cs}(\cdot, \omega) < 1,\ h^{cs}(0, \omega) = 0,\ \text{and } D_y h(0, \omega) = 0.$$

(ii) For any $x \in \mathbb{R}^d$ *and any* $\omega \in \Omega$*, if*

$$\varphi(t, \omega, x) \to 0, \quad \text{as } t \to +\infty,$$

and

$$\|\varphi(t, \omega, x)\| < \rho(\theta^t \omega), \quad \forall\ t \in \mathbb{N}, \tag{2.8}$$

then

$$x \in W^{cs}(\omega) = \{y + h^{cs}(y, \omega) : y \in E_{cs}(\omega)\}.$$

The center-stable manifold theorem plays a fundamental role in our analysis. There are richer and stronger results on and related to center-stable manifolds in [13, 20], and in Theorem 2.7, we only state the results that are necessary in our proofs. In particular, Theorem 2.7 states that there exists a random manifold $W^{cs}(\omega)$, tangent to the center-stable Oseledets subspace $E_{cs}(\omega)$ and defined as the graph of $h^{cs}(\omega)$, with $\dim W^{cs}(\omega) = \dim E^{cs}(\omega)$, such that a trajectory converging to $x^* = 0$ must start at a point in $W^{cs}(\omega)$, provided that it always stays in a tempered ball in the sense of (2.8).

3. Proof sketch and discussion

We present the main proof outlines and ideas in this section, with all technical lemmas and propositions deferred to Section 4. We also make some technical discussion and comparisons with [8].

3.1. **Setup of the random dynamical system.** We rigorously define the random dynamical system associated with Algorithm 1. In this setup, the system is reversible and has two-sided discrete time $t \in \mathbb{Z}$, which is because we cannot find a standard version of the center-stable manifold theorem for one-sided time $t \in \mathbb{N}$ in the existing literature.

• *Probability space.* For each $t \in \mathbb{Z}$, denote $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ the usual probability space for the distribution $\mathcal{U}_{\{1,2,\dots,n\}}$, where $\mathcal{U}_{\{1,2,\dots,n\}}$ are the uniform distributions on the set $\{1, 2, \dots, n\}$. As mentioned in the introduction, Algorithm 1 is naturally equipped with $(\Omega, \mathcal{F}, \mathbb{P})$ that is the product space of $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ for $t \in \mathbb{N}$. We extended the probability by including the negative times, i.e., we consider $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e)$ that is the product space of $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ for $t \in \mathbb{Z}$. It is clear that for every $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \mathbb{P}^e \left(\left(\prod_{t \in \mathbb{Z}_{<0}} \Omega_t \right) \times A \right). \tag{3.1}$$

It can be seen that every sample in Ω^e is a sequence of indices $\omega^e = (\cdots, i_{-2}, i_{-1}, i_0, i_1, i_2, \cdots)$ and we denote π_t as the projection map from $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e)$ to $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ that sends ω to i_t, for every $t \in \mathbb{Z}$.

• *Metric dynamical system.* The metric dynamical system on Ω^e is $\theta(t) = \theta^t$ for $t \in \mathbb{Z}$, where $\theta : \Omega^e \to \Omega^e$ is the shifting operator defined via

$$\pi_t(\theta\omega^e) = \pi_{t+1}(\omega^e), \quad \forall\ \omega^e \in \Omega^e,\ t \in \mathbb{Z},$$

which essentially shifts every element in ω^e leftward for one position. It is clear that θ is measurable and $\mathbb{P}^e$-preserving.

• *Random dynamical system.* For any $\omega^e \in \Omega^e$, we define a (nonlinear) map on $\mathbb{R}^d$ via

$$\begin{aligned} \phi(\omega^e) : \mathbb{R}^d \to &\qquad \mathbb{R}^d, \\ x \ &\mapsto x - \alpha e_i e_i^\top \nabla f(x), \end{aligned}$$

where $i = \pi_0(\omega^e)$ is the index in ω^e at the time $t = 0$. It can be seen that $\phi(\omega^e)$ implements one iteration of Algorithm 1 with the sampled index being $i = \pi_0(\omega^e)$. Then we define $\varphi(t, \omega^e) : \mathbb{R}^d \to \mathbb{R}^d$ via

$$\varphi(t, \omega^e) = \begin{cases} \phi(\theta^{t-1}\omega^e) \circ \cdots \circ \phi(\theta\omega^e) \circ \phi(\omega^e), & \text{if } t > 0, \\ Id, & \text{if } t = 0, \\ \phi(\theta^t\omega^e)^{-1} \circ \cdots \circ \phi(\theta^{-2}\omega^e)^{-1} \circ \phi(\theta^{-1}\omega^e)^{-1}, & \text{if } t < 0, \end{cases}$$

which satisfies the cocycle property (2.1). Here, $\phi(\omega^e) : \mathbb{R}^d \to \mathbb{R}^d$ is bijective and is hence invertible if $\alpha < 1/M$, where M is the constant as in Assumption 1. We will rigorously characterize the invertibility of $\phi(\omega^e)$ in Section 4.1. Thus, $\varphi(t, \omega^e)$ defines a random dynamical system on $X = \mathbb{R}^d$ over $\{\theta^t\}_{t \in \mathbb{Z}}$, and one can see that $\varphi(t, \omega^e)$ essentially implements the first t iterations of Algorithm 1 for $t \geq 0$.

3.2. **Restatement of Theorem 1.** One can restate Theorem 1 with the extended probability space $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e)$. In particular, denote $\Theta^e = \mathbb{R}^d \times \Omega^e$ that is equipped with a product measure

$$\mu^e = \text{Leb} \times \mathbb{P}^e,$$

and then define

$$\Theta^e(x^*) = \left\{ (x_0, \omega^e) \in \Theta^e : \lim_{t \to +\infty} \varphi(t, \omega^e, x_0) = x^* \right\},$$

for each $x^* \in \text{Crit}_s(f)$, and

$$\Theta^e(\text{Crit}_s(f)) = \bigcup_{x^* \in \text{Crit}_s(f)} \Theta^e(x^*).$$

One can thus see from (3.1) that Theorem 1 can be equivalently restated as follows.

Theorem 3.1 (Restatement of Theorem 1)**.** *Suppose that Assumptions 1 and 2 hold and that $0 < \alpha < 1/M$. It holds that*

$$\mu^e(\Theta^e(\text{Crit}_s(f))) = 0.$$

Theorem 3.1 essentially states that for μ^e-almost surely (x_0, ω^e) will not converge to a strict saddle point $x^* \in \text{Crit}_s(f)$. To prove Theorem 3.1, one only needs to investigate the measure of $\Theta^e(x^*)$ for every $x^* \in \text{Crit}_s(f)$, which is stated in the following theorem.

Theorem 3.2. *Suppose that Assumptions 1 holds and that $0 < \alpha < 1/M$. For any $x^* \in \text{Crit}_s(f)$, if x^* is a non-degenerate critical point of f, i.e., all eigenvalues of $\nabla^2 f(x^*)$ are nonzero, then*

$$\mu^e(\Theta^e(x^*)) = 0.$$

In particular, the proof of Theorem 3.1 is straightforward based on Theorem 3.2.

Proof of Theorem 3.1. It follows from Assumption 2 that every $x^* \in \text{Crit}_s(f)$ is an isolated critical point of f, which is because that $\nabla f(x) = \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|) \neq 0$ for any $x \neq x^*$ in a small neighborhood of x^*. This implies that $\text{Crit}_s(f)$ is countable. Then according to Theorem 3.2, we can obtain that

$$\mu^e(\Theta^e(\text{Crit}_s(f))) = \sum_{x^* \in \text{Crit}_s(f)} \mu^e(\Theta^e(x^*)) = 0,$$

which completes the proof of Theorem 3.1. □

Corollary 2 is also an immediate consequence.

Proof of Corollary 2. Using the same arguments as in [13, Proofs of Proposition 4.11 and Proposition 4.12], it can be shown that $x_t(\omega)$ converges to a point in $\text{Crit}(f)$ as $t \to +\infty$ unless (x_0, ω) is in a μ-null set. One can further exclude $\Theta(\text{Crit}_s(f))$, that is also a μ-null set as in Theorem 1, which guarantees that the limit is in $\text{Crit}(f) \setminus \text{Crit}_s(f)$. □

3.3. **Proof of Theorem 3.2.** In this subsection, we present the main proof outline of Theorem 3.2, with proofs of some technical lemmas and propositions deferred to Section 4. We consider any $x^* \in \mathrm{Crit}_s(f)$ and assume that $x^* = 0$ without loss of generality. The linearization of $\varphi(t, \omega^e)$ at $x^* = 0$ is

$$\Phi^H(t, \omega^e) = \begin{cases} A^H(\theta^{t-1}\omega^e) \cdots A^H(\theta\omega^e) A^H(\omega^e), & \text{if } t > 0, \\ I, & \text{if } t = 0, \\ A^H(\theta^t\omega^e)^{-1} \cdots A^H(\theta^{-2}\omega^e)^{-1} A^H(\theta^{-1}\omega^e)^{-1}, & \text{if } t < 0, \end{cases}$$

where

$$A^H(\omega^e) = I - \alpha e_i e_i^\top H, \quad H = \nabla^2 f(x^*).$$

A crucial step in the proof of Theorem 3.2 is to apply the center-stable manifold theorem, i.e., Theorem 2.7, for which one needs to validate Assumption 2.5. We include the detailed validation with Assumption 1 and $\alpha < 1/M$ in Section 4.2. Moreover, we need the following two propositions.

Proposition 3.3. *Let $H = \nabla f(x^*)$ have a negative eigenvalue and $0 < \alpha < 1/\max_{1 \le i \le d} |H_{ii}|$, then the largest Lyapunov exponent of $\Phi^H(t, \omega^e)$ is positive.*

Proposition 3.4. *Suppose that Assumption 1 holds and that $0 < \alpha < 1/M$. For any $x^* \in \mathrm{Crit}_s(f)$, if x^* is a non-degenerate critical point of f, i.e., all eigenvalues of $\nabla^2 f(x^*)$ are nonzero, then there exists $\widetilde{\Theta} \subseteq \Theta = \mathbb{R}^d \times \Omega$ with $\mu(\Theta \setminus \widetilde{\Theta}) = 0$, such that for any $(x_0, \omega) \in \widetilde{\Theta}$, if $x_t(\omega) \to x^*$ as $t \to +\infty$, then $x_t(\omega) \to x^*$ exponentially as $t \to +\infty$.*

Throughout this paper, we say that a sequences y_t in $\mathbb{R}^d$ or $\mathbb{R}$ converges exponentially to y^* as $t \to +\infty$ if

$$\limsup_{t \to +\infty} \frac{1}{t} \log \|y_t - y^*\| < 0.$$

Note that our notion of exponential convergence is essentially at least exponential convergence, which includes convergence rates faster than exponential rates. The proof of Proposition 3.3 follows exactly the same line as in the proof of [8, Proposition 3.1] and is hence omitted. The proof of Proposition 3.4 is presented in Section 4.3. Proposition 3.4 states that almost every convergent trajectory has an exponential convergence rate, which eventually makes the condition (2.8) true and hence guarantees that we can identify convergent trajectories with points on the center-stable manifold. In addition, Proposition 3.3 shows that the center-stable manifold is of dimension at most $d-1$ due to the presence of the positive Lyapunov exponent, and is hence of Lebesgue measure zero. These observations lead to Theorem 3.2. We then present the proof of Theorem 3.2 based on Assumption 2.5, Proposition 3.3, and Proposition 3.4.

Proof of Theorem 3.2. Let $\widetilde{\Omega}^e \subset \Omega^e$ be θ-invariant with $\mathbb{P}^e(\widetilde{\Omega}^e) = 1$, so that all statements in Theorem 2.3 and Theorem 2.7 are true for every $\omega^e \in \widetilde{\Omega}^e$, and let $\rho : \Omega^e \to (0, +\infty)$ be the tempered random variable as in Theorem 2.7 satisfying

$$\lim_{t \to \pm\infty} \frac{1}{t} \log \rho(\theta^t \omega^e) = 0, \quad \forall\, \omega^e \in \widetilde{\Omega}^e.$$

Moreover, it follows from the Kolmogorov's zero-one law that $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e, \theta)$ is ergodic, which implies that $p(\omega^e), \lambda_i(\omega^e), d_i(\omega^e)$ are all constant over $\omega^e \in \widetilde{\Omega}^e$. We thus drop the dependence on ω^e and denote these constants by p, λ_i, d_i for simplicity. Denote $E^{cs}(\omega^e)$ and $W^{cs}(\omega^e)$ as the center-stable Oseledets subspace and the center-stable manifold, respectively. It follows from Proposition 3.3 that

$$\dim W^{cs}(\omega^e) = \dim E^{cs}(\omega^e) = d - \sum_{\lambda_i > 0} d_i \leq d-1, \quad \forall\, \omega^e \in \widetilde{\Omega}^e, \tag{3.2}$$

Define

$$\widetilde{\Theta}^e = \left(\prod_{t \in \mathbb{Z}_{<0}} \Omega_t \right) \times \widetilde{\Theta} \subseteq \Theta^e = \mathbb{R}^d \times \Omega^e,$$

where $\widetilde{\Theta} \subseteq \Theta$ is from Proposition 3.4. Consider any

$$(x_0, \omega^e) \in \Theta^e(x^*) \cap \widetilde{\Theta}^e \cap \left(\mathbb{R}^d \times \widetilde{\Omega}^e \right).$$

It follows from Proposition 3.4 that

$$\varphi(t, \omega^e, x_0) \to x^* = 0, \quad t \to +\infty,$$

exponentially, i.e.,

$$\limsup_{t \to +\infty} \frac{1}{t} \log \|\varphi(t, \omega^e, x_0)\| < 0.$$

Therefore, there exists $T(x_0, \omega^e) \in \mathbb{N}$ such that

$$\frac{1}{t} \log \|\varphi(t, \omega^e, x_0)\| < \frac{1}{t} \log \rho(\theta^t \omega^e), \quad \forall\, t \geq T(x_0, \omega^e),$$

and equivalently that,

$$\|\varphi(t, \omega^e, x_0)\| < \rho(\theta^t \omega^e), \quad \forall\, t \geq T(x_0, \omega^e).$$

We then have that

$$\varphi\left(t, \theta^{T(x_0, \omega^e)} \omega^e, \varphi\left(T(x_0, \omega^e), \omega^e, x_0\right)\right) \to x^* = 0, \quad t \to +\infty,$$

and

$$\left\| \varphi\left(t, \theta^{T(x_0, \omega^e)} \omega^e, \varphi\left(T(x_0, \omega^e), \omega^e, x_0\right)\right) \right\| < \rho\left(\theta^t \theta^{T(x_0, \omega^e)} \omega^e\right), \quad \forall\, t \in \mathbb{N}.$$

Using Theorem 2.7, we can conclude that

$$\varphi\left(T(x_0, \omega^e), \omega^e, x_0\right) \in W^{cs}\left(\theta^{T(x_0, \omega^e)} \omega^e\right),$$

which implies that

$$x_0 \in \varphi(T(x_0, \omega^e), \omega^e)^{-1} \left(W^{cs}\left(\theta^{T(x_0, \omega^e)} \omega^e\right)\right) \subseteq \bigcup_{t \in \mathbb{N}} \varphi(t, \omega^e)^{-1} \left(W^{cs}(\theta^t \omega^e)\right),$$

where the Lebesgue measure of the set on the right-hand side can be computed from (3.2) as

$$\textsc{Leb}\left(\bigcup_{t \in \mathbb{N}} \varphi(t, \omega^e)^{-1} \left(W^{cs}(\theta^t \omega^e)\right)\right) \leq \sum_{t \in \mathbb{N}} \textsc{Leb}\left(\varphi(t, \omega^e)^{-1} \left(W^{cs}(\theta^t \omega^e)\right)\right) = 0,$$

since the image of a manifold of dimension at most $d-1$ under a $\mathcal{C}^1$ map is of Lebesgue measure zero. Therefore, by Fubini's theorem, one obtains that

$$\mu^e\left(\Theta^e(x^*)\cap\widetilde{\Theta}^e\cap\left(\mathbb{R}^d\times\widetilde{\Omega}^e\right)\right)\leq\int_{\Omega^e}\textsc{Leb}\left(\bigcup_{t\in\mathbb{N}}\varphi(t,\omega^e)^{-1}\left(W^{cs}(\theta^t\omega^e)\right)\right)\mathbb{P}^e(\mathrm{d}\omega^e)=0,$$

and hence that

$$\mu^e\left(\Theta^e(x^*)\right)\leq\mu^e\left(\Theta^e(x^*)\cap\widetilde{\Theta}^e\cap\left(\mathbb{R}^d\times\widetilde{\Omega}^e\right)\right)+\mu^e\left(\Theta^e\setminus\widetilde{\Theta}^e\right)+\mu^e\left(\Theta^e\setminus(\mathbb{R}^d\times\widetilde{\Omega}^e)\right)=0,$$

where we used $\mu(\Theta\setminus\widetilde{\Theta})=0$ and $\mathbb{P}^e(\Omega^e\setminus\widetilde{\Omega}^e)=0$. The proof is completed. □

3.4. **Comparison and discussion.** As the conclusion of this section, we make a technical comparison with [8], whose main theorem is of similar style to ours, i.e., for any x_0 that is not a strict saddle point, x_t almost surely does not converge to a strict saddle point as $t\to+\infty$. The main difference is that the following assumption is required in [8].

Assumption 3.5 ([8, Assumption 3])**.** *For every $x^*\in\mathrm{Crit}_s(f)$, it holds that $\mathcal{P}_+^H(\omega)e_i\neq 0$, for every $i\in\{1,2,\dots,d\}$ and almost every $\omega\in\Omega$, where $\mathcal{P}_+^H(\omega)$ is the orthogonal projection onto $W_+^H(\omega)=\bigoplus_{\lambda_i(\omega)>0}V_i^H(\omega)$ and $V_i^H(\omega)$ is the eigenspace corresponding to the eigenvalue $\lambda_i(\omega)$ of $\Lambda^H(\omega)$ as in (2.3), $i=1,2,\dots,p(\omega)$, for the linearized system at x^* with $H=\nabla^2 f(x^*)$.*

We remark that [8] only considers sample ω corresponding to one-sided time $\mathbb{T}=\mathbb{N}$, without extending to ω^e and two-sided time $\mathbb{T}=\mathbb{Z}$. Therefore, there is no explicit notion of center-stable and unstable Oseledets subspaces, but the limiting matrix $\Lambda^H(\omega)$ and its eigenspaces $V_i^H(\omega),\ i=1,2,\dots,p(\omega)$ are still well-defined, as they are defined one-sidedly for $t\to+\infty$ in (2.3). In our notation, one can verify that $W_+^H(\omega^e)=\bigoplus_{\lambda_i>0}V_i^H(\omega^e)$ is the orthogonal complement of the center-stable Oseledets subspace in $\mathbb{R}^d$.

The main proof in [8] is that, with Assumption 3.5 and additional randomness in step-sizes, the iterate would have a non-negligible component in $W_+^H(\omega)$ with high probability, which will be amplified sufficiently and hence drive the dynamics to leave the neighborhood of $x^*\in\mathrm{Crit}_s(f)$, avoiding the convergence to x^*. However, there are two main drawbacks of this analytical framework. Firstly, additional randomness is a bit artificial, making the randomized coordinate gradient descent not in its simplest format. In fact, the sample ω in [8] is a sequence of not only the random coordinates, but also the random stepsizes. Secondly, though the techinical Assumption 3.5 can be verified generically, it excludes some practical Hessian matrices, such as $H=\nabla^2 f(x^*)=\mathrm{diag}(H_1,H_2)$ where all eigenvalues of H_1 are positive and all eigenvalues of H_2 are negative, which is already acknowledged in [8].

Overall, our present work overcomes some technical obstacles in [8], and establishes a neater and more general analytical framework for the convergence of randomized coordinate gradient descent. This general framework could be potentially applied or extended for other randomized algorithms.

4. Technical lemmas and propositions

We collect all technical lemmas and propositions in this section.

4.1. **Invertibility of $\phi(\omega^e)$.** Recall that $\phi(\omega^e) : x \mapsto x - \alpha e_i e_i^\top \nabla f(x)$, where $i = \pi_0(\omega^e)$ is the index in ω^e at the time $t = 0$, and we assume that $f \in \mathcal{C}^2(\mathbb{R}^d)$ and $\left\|\nabla^2 f(x)\right\| \leq M,\ \forall\ x \in \mathbb{R}^d$, as in Assumption 1. The Jacobian matrix of $\phi(\omega^e)$ at x is

$$J(x) = I - \alpha e_i e_i^\top \nabla^2 f(x),$$

which is always invertible by the Sherman-Morison formula if $\alpha < 1/M$. Therefore, according to the inverse function theorem, at any $x \in \mathbb{R}^d$, $\phi(\omega^e)$ is locally invertible and the inverse is also $\mathcal{C}^1$. Moreover, $\alpha < 1/M$ yields that

$$\lim_{\|x\| \to +\infty} \|\phi(\omega^e)(x)\| = +\infty,$$

which implies that $\phi(\omega^e)$ is proper, i.e., the preimage of a compact set is still compact. One can thus apply the Hadamard's global inverse function theorem [12, 16] and conclude that $\phi(\omega^e)$ is globally invertible.

4.2. **Validation of Assumption 2.5.** Suppose that Assumption 1 is satisfied and that $\alpha < 1/M$. Then $\varphi(t, \omega^e)$ is a $\mathcal{C}^1$ random dynamical system since $\phi(\omega^e) \in \mathcal{C}^1(\mathbb{R}^d)$. Using the same argument as in [8], it can be verified that $A^H(\omega^e)$ and $A^H(\omega^e)^{-1}$ are uniformly bounded in $\omega^e \in \Omega^e$ and hence that the linearized system $\Phi^H(t, \omega^e)$ satisfies the conditions of the multiplicative ergodic theorem (Theorem 2.3). Therefore, the first condition in Assumption 2.5 is satisfied.

For the second condition in Assumption 2.5, the residual of the linearization at $x^* = 0$ and $t = 1$ can be computed as

$$\begin{aligned} F(1, \omega^e, x) &= \varphi(1, \omega^e, x) - \Phi^H(1, \omega^e) = \phi(\omega^e) - A^H(\omega^e) \\ &= \left(x - \alpha e_i e_i^\top \nabla f(x)\right) - \left(x - \alpha e_i e_i^\top Hx\right) = \alpha e_i e_i^\top (Hx - \nabla f(x)), \end{aligned}$$

where $i = \pi_0(\omega^e)$. Note that $x \mapsto Hx - \nabla f(x)$ is a $\mathcal{C}^1$ map since $f \in \mathcal{C}^2(\mathbb{R}^d)$. As in assumption 1, there exists a neighborhood $N(x^*)$ of $x^* = 0$ and constants B_0, B_1, L so that

$$\begin{aligned} &\sup_{x \in N(x^*)} \left\|D_x^k F(1, \omega^e, x)\right\| \leq B_k, \quad \forall\ 0 \leq k \leq 1,\ \omega^e \in \Omega^e, \\ &\|D_x F(1, \omega^e, x)\| \leq \alpha L \|x\|, \quad \forall\ x \in N(x^*),\ \omega^e \in \Omega^e. \end{aligned}$$

This verifies (2.7) as constants are tempered random variables.

4.3. **Proof of Proposition 3.4.** This subsection proves Proposition 3.4, which only uses the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that is the product space of $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ for all $t \in \mathbb{N}$, not the extended one $(\Omega^e, \mathcal{F}^e, \mathbb{P}^e)$. We will use the filtration $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$ where $\mathcal{F}_t$ is the sigma algebra generated by $\left\{ \prod_{s=0}^{t} B_s \times \prod_{s=t+1}^{+\infty} \Omega_s : B_s \in \Sigma_s,\ s = 0, 1, \ldots, t \right\}$.

Recall that Proposition 3.4 states that for any strict saddle point $x^* \in \mathrm{Crit}_s(f)$ with non-degenerate Hessian $\nabla f(x^*)$, there exists $\widetilde{\Theta} \subset \Theta = \mathbb{R}^d \times \Omega$ with $\mu(\Theta \setminus \widetilde{\Theta}) = 0$, such that for any $(x_0, \omega) \in \widetilde{\Theta}$, $x_t(\omega) \to x^*$ as $t \to +\infty$ implies that $x_t(\omega) \to x^*$ exponentially. The proof is divided into two parts. Firstly, we construct $\Theta_1 \subset \Theta$ with $\mu(\Theta \setminus \Theta_1) = 0$, so that for any $(x_0, \omega) \in \Theta_1$, $x_t(\omega) \to x^*$ implies that $f(x_t(\omega)) \to f(x^*)$ exponentially. Secondly, we construct $\Theta_2 \subset \Theta$ with $\mu(\Theta \setminus \Theta_2) = 0$ and define $\widetilde{\Theta} = \Theta_1 \cap \Theta_2$. It can be shown that for any $(x_0, \omega) \in \widetilde{\Theta}$,

the exponential convergence of $f(x_t(\omega))$ to $f(x^*)$ implies that $x_t(\omega) \to x^*$ exponentially. These two parts are elaborated in Section 4.3.1 and Section 4.3.2, respectively. In the rest of this subsection, we always assume that $x^* = 0$ and $f(x^*) = 0$ without loss of generality.

4.3.1. *Exponential convergence of $f(x_t(\omega))$.* For any $(x_0, \omega) \in \Theta$, where $\omega = (i_0, i_1, i_2, \dots)$, and any $t \in \mathbb{N}$, we define

$$I_t(x_0, \omega) = \begin{cases} 1, & \text{if } |e_{i_t}^\top \nabla f(x_t)| \geq \frac{1}{\sqrt{d}} \|\nabla f(x_t)\|, \\ 0, & \text{otherwise,} \end{cases} \tag{4.1}$$

where $x_t = x_t(\omega)$ is generated by Algorithm 1 given (x_0, ω). We further define

$$\Theta_1 = \left\{ (x_0, \omega) \in \Theta : \liminf_{t \to +\infty} \frac{\sum_{s=0}^{t-1} I_s(x_0, \omega)}{t} \geq \frac{1}{d} \right\}. \tag{4.2}$$

Lemma 4.1. *It holds that $\mu(\Theta \setminus \Theta_1) = 0$.*

Proof. Define

$$J_t(x_0, \omega) = \begin{cases} 1, & \text{if } i_t = \min \left\{ i : |e_i^\top \nabla f(x_t)| \geq \frac{1}{\sqrt{d}} \|\nabla f(x_t)\| \right\}, \\ 0, & \text{otherwise.} \end{cases}$$

It is clear that $I_t(x_0, \omega) \geq J_t(x_0, \omega)$ and that

$$\mathbb{P}\left(J_t(x_0, \omega) = 1 \,\middle|\, \mathcal{F}_{t-1} \right) = \frac{1}{d}. \tag{4.3}$$

which is because that $\left| e_{i_t}^\top \nabla f(x_t) \right| \geq \frac{1}{\sqrt{d}} \|\nabla f(x_t)\|$ is true if the i_t-th entry of $\nabla f(x_t)$ has the largest absolute value among all entries of $\nabla f(x_t)$, and that i_t is sampled uniformly randomly from $\{1, 2, \dots, d\}$. Consider

$$\widetilde{J}_t(x_0, \omega) = J_t(x_0, \omega) - \frac{1}{d},$$

and we can have that

$$\mathbb{E}\left(\widetilde{J}_t(x_0, \omega) \,\middle|\, \mathcal{F}_{t-1} \right) = 0,$$

which indicates that $\left\{ \sum_{s=0}^{t} \widetilde{J}_s(x_0, \omega) \right\}_{t \in \mathbb{N}}$ is a martingale with respect to $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$. According to the strong law of large numbers for martingales [9], for any $x_0 \in \mathbb{R}^d$, we have for $\mathbb{P}$-almost every ω that

$$\lim_{t \to +\infty} \frac{\sum_{s=0}^{t-1} \widetilde{J}_s(x_0, \omega)}{t} = 0,$$

which implies that

$$\liminf_{t \to +\infty} \frac{\sum_{s=0}^{t-1} I_s(x_0, \omega)}{t} \geq \lim_{t \to +\infty} \frac{\sum_{s=0}^{t-1} J_s(x_0, \omega)}{t} = \lim_{t \to +\infty} \frac{\sum_{s=0}^{t-1} \widetilde{J}_s(x_0, \omega)}{t} + \frac{1}{d} = \frac{1}{d},$$

For any $x_0 \in \mathbb{R}^d$, we set

$$\Omega_1(x_0) = \left\{ \omega \in \Omega : \liminf_{t \to +\infty} \frac{\sum_{s=0}^{t-1} I_s(x_0, \omega)}{t} \geq \frac{1}{d} \right\},$$

which satisfies

$$\mathbb{P}(\Omega_1(x_0)) = 1.$$

One can thus conclude by applying the Fubini's theorem that

$$\mu(\Theta \setminus \Theta_1) = \int_{\mathbb{R}^d} \int_{\Omega} \mathbf{1}_{\Theta\setminus\Theta_1}(x_0,\omega)\mathbb{P}(\mathrm{d}\omega)\mathrm{LEB}(\mathrm{d}x_0) = \int_{\mathbb{R}^d} \mathbb{P}(\Omega \setminus \Omega_1(x_0))\mathrm{LEB}(\mathrm{d}x_0) = 0,$$

where $\mathbf{1}_{\Theta\setminus\Theta_1}(x_0,\omega)$ is the indicator function, i.e., $\mathbf{1}_{\Theta\setminus\Theta_1}(x_0,\omega) = 1$ if $(x_0,\omega) \in \Theta \setminus \Theta_1$ and $\mathbf{1}_{\Theta\setminus\Theta_1}(x_0,\omega) = 0$ otherwise. The proof is hence completed. □

Lemma 4.2. *Suppose that Assumption 1 holds and that $0 < \alpha < 1/M$. Let $x^* = 0 \in \mathrm{Crit}_s(f)$ be a strict saddle point of f with non-degenerate Hessian $\nabla^2 f(x^*)$ and let $f(x^*) = 0$. For $(x_0,\omega) \in \Theta_1$, if $x_t(\omega) \to 0$ as $t \to +\infty$, then $f(x_t(\omega)) \to 0$ exponentially as $t \to +\infty$.*

Proof. Consider any $(x_0,\omega) \in \Theta_1$ with $x_t = x_t(\omega) \to 0$. With Assumption 1 and $0 < \alpha < 1/M$, $f(x_t)$ is monotonically decreasing in t, which can be derived by Taylor's expansion at x_t,

$$\begin{aligned} f(x_{t+1}) &= f\left(x_t - \alpha e_{i_t} e_{i_t}^\top \nabla f(x_t)\right) \\ &= f(x_t) - \alpha \left(e_{i_t}^\top \nabla f(x_t)\right)^2 + \frac{1}{2}\alpha^2 \left(e_{i_t}^\top \nabla f(x_t)\right)^2 \cdot e_{i_t}^\top \nabla^2 f\left(x_t - \theta_t \alpha e_{i_t} e_{i_t}^\top \nabla f(x_t)\right) e_{i_t} \\ &\le f(x_t) - \frac{1}{2}\alpha \left(e_{i_t}^\top \nabla f(x_t)\right)^2 \le f(x_t), \end{aligned} \tag{4.4}$$

where $\theta_t \in (0,1)$. The monotonicity implies that

$$f(x_t) \ge f(0) = 0, \quad \forall\, t \in \mathbb{N}.$$

Note that $\nabla^2 f(0)$ is non-degenerate. There exist a neighborhood U_1 of 0 and a constant $\sigma > 0$ such that

$$\|\nabla f(x)\| \ge \sigma \|x\|, \quad \forall\, x \in U_1. \tag{4.5}$$

Since $x_t \to 0$, we have $x_t \in U_1$, $\forall\, t \ge T$ for some $T \in \mathbb{N}$. Therefore, for any $t \ge T$, if $I_t(x_0,\omega) = 1$, we have from (4.1), (4.4), and (4.5) that

$$f(x_{t+1}) \le f(x_t) - \frac{\alpha}{2d}\|\nabla f(x_t)\|^2 \le f(x_t) - \frac{\alpha\sigma^2}{2d}\|x_t\|^2 \le \left(1 - \frac{\alpha\sigma^2}{Md}\right) f(x_t), \tag{4.6}$$

where the last inequality follows from the Tayler expansion at 0, namely

$$f(x_t) = f(0) + \nabla f(0)^\top x_t + \frac{1}{2} x_t^\top \nabla^2 f(\theta_t' x_t) x_t = \frac{1}{2} x_t^\top \nabla^2 f(\theta_t' x_t) x_t \le \frac{M}{2}\|x_t\|^2,$$

with $\theta_t' \in (0,1)$. Applying (4.6) and the monotonicity of $f(x_t)$ repeatedly, one obtains that

$$f(x_t) \le \left(1 - \frac{\alpha\sigma^2}{Md}\right)^{\sum_{s=T}^{t-1} I_s(x_0,\omega)} f(x_T), \quad \forall\, t \ge T.$$

Therefore, one can immediately conclude the exponential convergence rate of $f(x_t) \to 0$ as the construction of Θ_1, say (4.2), suggests that

$$\liminf_{t\to+\infty} \frac{\sum_{s=T}^{t-1} I_s(x_0,\omega)}{t} = \liminf_{t\to+\infty} \frac{\sum_{s=0}^{t-1} I_s(x_0,\omega)}{t} \ge \frac{1}{d},$$

which completes the proof. □

4.3.2. *Exponential convergence of $x_t(\omega)$.* We work with the same settings as in Lemma 4.2, and let U_1, σ from (4.5). Let $\rho = \rho(\alpha, d, \sigma, M) \in (0,1)$ be another constant satisfying

$$\rho M + \frac{M\rho^2}{2} < \frac{\alpha\sigma^2}{2d}(1-\rho)^2. \tag{4.7}$$

Define

$$U_2 = U_2(\rho) = \bigcup_{x \in f^{-1}(0)} B(x, \rho \|x\|),$$

where $B(x,r)$ is the open ball in $\mathbb{R}^d$ centered at x with radius being r, and

$$S = U_1 \cap U_2 \cap f^{-1}([0, +\infty)).$$

Then we further define that

$$\Theta_2 = \big\{(x_0, \omega) \in \Theta : x_t(\omega) \in S \text{ finitely often}\big\}. \tag{4.8}$$

Lemma 4.3. *Suppose that Assumption 1 holds and that $0 < \alpha < 1/M$. Let $x^* = 0 \in \mathrm{Crit}_s(f)$ be a strict saddle point of f with non-degenerate Hessian $\nabla^2 f(x^*)$ and let $f(x^*) = 0$. It holds that $\mu(\Theta \setminus \Theta_2) = 0$.*

Proof. Fix $x_0 \in \mathbb{R}^d$. For any $k \in \mathbb{N}_+$, define a random variable $\tau_k(x_0)$ via

$$\tau_k(x_0) = \tau_k(x_0, \omega) = \begin{cases} t, & \text{if } x_t(\omega) \in S \text{ and } \#\{0 \le t' < t : x_{t'}(\omega) \in S\} = k-1, \\ +\infty, & \text{if } \#\{t' \in \mathbb{N} : x_{t'}(\omega) \in S\} < k, \end{cases}$$

i.e., τ_k is the stopping time that x_t visits S for exactly k times. By the definition of Θ_2, i.e., (4.8), it is clear that $(x_0, \omega) \in \Theta \setminus \Theta_2$ if any only if $\tau_k(x_0, \omega) < +\infty,\ \forall\ k \in \mathbb{N}$.

Suppose that $x_t \in S$. Since $x_t \in U_2$, there exists x_t' with $f(x_t') = 0$ and $x_t \in B(x_t', \rho\|x_t'\|)$. It is clear that $x_t' \neq 0$ since otherwise $B(x_t', \rho\|x_t'\|) = \emptyset$, and that $\|x_t\| \ge (1-\rho)\|x_t'\|$. Using Taylor's expansion at x_t' and Assumption 1, we have

$$\begin{aligned} f(x_t) &\le \|\nabla f(x_t')(x_t - x_t')\| + \frac{M}{2}\|x_t - x_t'\|^2 \\ &\le M\|x_t'\|\|x_t - x_t'\| + \frac{M}{2}\|x_t - x_t'\|^2 \le \left(\rho M + \frac{M\rho^2}{2}\right)\|x_t'\|^2, \end{aligned}$$

which combined with (4.6) yields that, with probability at least $\frac{1}{d}$ conditioned on x_t, it holds that

$$f(x_{t+1}) \le f(x_t) - \frac{\alpha\sigma^2}{2d}\|x_t\|^2 \le \left(\rho M + \frac{M\rho^2}{2}\right)\|x_t'\|^2 - \frac{\alpha\sigma^2}{2d}(1-\rho)^2\|x_t'\|^2 < 0,$$

where we used (4.7). This proves that as long as $x_t \in S$,

$$\mathbb{P}\big(f(x_{t+1}) < 0 \mid x_t\big) \ge \frac{1}{d},$$

which implies that

$$\mathbb{P}\big(f(x_{t+1}) \ge 0 \mid \tau_k(x_0) = t\big) \le 1 - \frac{1}{d}.$$

Note that $f(x_{t+1}) < 0$ implies that $x_{t'} \notin S$ for any $t' \geq t+1$ due to the monotonically decreasing property (4.4). For any $k \geq 1$, one can estimate that

$$\begin{aligned}
&\mathbb{P}\big(\tau_{k+1}(x_0) < +\infty \mid \tau_k(x_0) < +\infty\big) \\
&\quad = \sum_{t\in\mathbb{N}} \mathbb{P}\big(\tau_k(x_0) = t, \tau_{k+1}(x_0) < +\infty \mid \tau_k(x_0) < +\infty\big) \\
&\quad \leq \sum_{t\in\mathbb{N}} \mathbb{P}\big(\tau_k(x_0) = t, f(x_{t+1}) \geq 0 \mid \tau_k(x_0) < +\infty\big) \\
&\quad = \sum_{t\in\mathbb{N}} \mathbb{P}\big(\tau_k(x_0) = t \mid \tau_k(x_0) < +\infty\big) \cdot \mathbb{P}\big(f(x_{t+1}) \geq 0 \mid \tau_k(x_0) = t\big) \\
&\quad \leq \left(1 - \frac{1}{d}\right) \sum_{t\in\mathbb{N}} \mathbb{P}\big(\tau_k(x_0) = t \mid \tau_k(x_0) < +\infty\big) \\
&\quad = 1 - \frac{1}{d}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\big(x_t \in S, \text{ i.o.} \mid x_0\big) &\leq \mathbb{P}\big(\tau_k(x_0) < +\infty\big) \\
&\leq \mathbb{P}\big(\tau_1(x_0) < +\infty\big) \prod_{k'=1}^{k-1} \mathbb{P}\big(\tau_{k'+1}(x_0) < +\infty \mid \tau_{k'}(x_0) < +\infty\big) \\
&\leq \left(1 - \frac{1}{d}\right)^{k-1}, \quad \forall\, k \geq 1,
\end{aligned}$$

which implies that

$$\mathbb{P}\big(x_t \in S, \text{ i.o.} \mid x_0\big) = 0, \quad \forall\, x_0 \in \mathbb{R}^d.$$

One can then conclude $\mu(\Theta \setminus \Theta_2) = 0$ by integrating against $x_0 \in \mathbb{R}^d$. □

We need the following lemma characterizing some local geometric properties near x^*.

Lemma 4.4. *Suppose that Assumption 1 holds and that $0 < \alpha < 1/M$. Let $x^* = 0 \in \mathrm{Crit}_s(f)$ be a strict saddle point of f with non-degenerate Hessian $\nabla^2 f(x^*)$ and let $f(x^*) = 0$. There exists a constant $p > 0$ and a neighborhood U of x^*, such that*

$$f(x) \geq p\,\|x\|^2, \quad \forall\, x \in \big(U \cap f^{-1}([0,+\infty))\big) \setminus S. \tag{4.9}$$

With Lemma 4.4, one can directly conclude the exponential convergence of $x_t(\omega)$ to x^* from the exponential convergence of $f(x_t(\omega))$ to $f(x^*)$, assuming $x_t(\omega) \to x^*$ and $x_t(\omega) \in S$ finitely often. This leads to the proof of Proposition 3.4.

Proof of proposition 3.4. Assume $x^* = 0$ and $f(x^*) = 0$ without loss of generality and define $\widetilde{\Theta} = \Theta_1 \cap \Theta_2$. It follows from Lemma 4.1 and Lemma 4.3 that $\mu(\Theta \setminus \widetilde{\Theta}) = 0$. For any $(x_0, \omega) \in \widetilde{\Theta}$ with $x_t(\omega) \to 0$, we have from Lemma 4.2 that $f(x_t(\omega)) \to 0$ exponentially. The definition of Θ_2 in (4.8) and the monotonically decreasing property (4.4) guarantee that $x_t(\omega) \in (U \cap f^{-1}([0,+\infty))) \setminus S$ for large enough t. Then one can conclude the exponential convergence of $x_t(\omega)$ to 0 by applying (4.9). □

We prove Lemma 4.4 in the rest of this subsection, where we denote $H = \nabla^2 f(0) \in \mathbb{R}^{d\times d}$ that is symmetric with all eigenvalues being nonzero, and set

$$f^H(x) = \frac{1}{2} x^\top H x. \tag{4.10}$$

Lemma 4.5. *Let $H \in \mathbb{R}^{d\times d}$ be symmetric and consider the quadratic function f^H as in (4.10). For any $\rho^H > 0$, define*

$$U_2^H = U_2^H(\rho^H) = \bigcup_{x\in (f^H)^{-1}(0)} B(x, \rho^H \|x\|).$$

There exist constants $p_+^H > 0$ and $p_-^H < 0$ so that the followings hold for any $x \in \mathbb{R}^d \setminus U_2^H$:

(i) If $f^H(x) \geq 0$, then $f^H(x) \geq p_+^H \|x\|^2$.
(ii) If $f^H(x) \leq 0$, then $f^H(x) \leq p_-^H \|x\|^2$.

Proof. We only prove (i) since (ii) is a direct corollary of (i) for $\frac{1}{2}x^\top(-H)x$, and we assume that $(f^H)^{-1}([0,+\infty)) \setminus U_2^H$ is not empty since otherwise the result is trivial. It can be seen that $(f^H)^{-1}(0)$, $\mathbb{R}^d \setminus U_2^H$, and $(f^H)^{-1}([0,+\infty))$ are all closed under scalar multiplication, which implies for any $c, c' > 0$,

$$\frac{c}{c'} \cdot \left(\left(\partial B(0,c') \cap (f^H)^{-1}([0,+\infty))\right) \setminus U_2^H\right) = \left(\partial B(0,c) \cap (f^H)^{-1}([0,+\infty))\right) \setminus U_2^H,$$

where $\partial B(0,c')$ and $\partial B(0,c)$ are the boundaries of $B(0,c')$ and $B(0,c)$, respectively. This homogeneity property leads to that

$$\inf_{x\in(\partial B(0,c')\cap(f^H)^{-1}([0,+\infty)))\setminus U_2^H} \frac{f^H(x)}{\|x\|^2} = \inf_{x\in(\partial B(0,c)\cap(f^H)^{-1}([0,+\infty)))\setminus U_2^H} \frac{f^H(x)}{\|x\|^2}. \tag{4.11}$$

Notice that $\partial B(0,c)$ and $(f^H)^{-1}([0,+\infty))$ are both closed and that U_2^H is open. Therefore, $\left(\partial B(0,c) \cap (f^H)^{-1}([0,+\infty))\right) \setminus U_2^H$ is closed, on which $f^H(x)/\|x\|^2$ is always positive. One can thus conclude that

$$p_+^H := \inf_{x\in(\partial B(0,c)\cap(f^H)^{-1}([0,+\infty)))\setminus U_2^H} \frac{f^H(x)}{\|x\|^2} = \min_{x\in(\partial B(0,c)\cap(f^H)^{-1}([0,+\infty)))\setminus U_2^H} \frac{f^H(x)}{\|x\|^2} > 0. \tag{4.12}$$

Combining (4.11) and (4.12), we know that

$$f^H(x) \geq p_+^H \|x\|^2, \quad \forall\, x \in (f^H)^{-1}([0,+\infty)) \setminus U_2^H,$$

which proves (i). □

Lemma 4.6. *Let $x^* = 0 \in \mathrm{Crit}_s(f)$ be a strict saddle point of f with non-degenerate Hessian $H = \nabla^2 f(x^*)$ and let $f(x^*) = 0$. If H has at least one positive eigenvalue and $\rho^H < \rho/4 < 1/4$, then there exists a neighborhood U' of 0, such that $U_2^H(\rho^H) \cap U' \subseteq U_2(\rho) \cap U'$.*

Proof. Without loss of generality, we assume that $H = \mathrm{diag}(h_1, \ldots, h_{d'}, h_{d'+1}, \ldots, h_d)$, where $h_1 \geq \cdots \geq h_{d'} > 0 > h_{d'+1} \geq \cdots \geq h_d$, since otherwise one can change the coordinates via an orthogonal transformation. Define

$$c = \frac{1}{2} \min\{h_{d'}, -h_{d'+1}\} \cdot \left(2\rho^H + (\rho^H)^2\right) > 0, \tag{4.13}$$

and let $c' > 0$ depend on c and f near 0 so that

$$\left|f(x) - f^H(x)\right| < \frac{c}{4}\left\|x\right\|^2, \quad \forall\, x \in B(0, 3c'). \tag{4.14}$$

Set $U' = B(0, c')$. Consider any $x \in U_2^H(\rho^H) \cap U'$, and it suffices to show that $x \in U_2(\rho)$. There exists $x' \in (f^H)^{-1}(0) \setminus \{0\}$ so that $x \in B(x', \rho^H \left\|x'\right\|)$ by the definition of $U_2^H(\rho^H)$. One also has $x' \in B(0, 2c')$ since $\rho^H < 1/2$ and $x \in B(0, c')$. Define

$$x'_+ = x' + \rho^H P_+ x', \quad x'_- = x' + \rho^H P_- x',$$

where $P_+ = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{d\times d}$ has d' nonzero diagonal entries and $P_- = I - P_+$. Note that

$$0 = f^H(x') = f^H(P_+ x') + f^H(P_- x'),$$

which implies that

$$\begin{aligned} f^H(P_+ x') = -f^H(P_- x') = \frac{1}{2} f^H(P_+ x') - \frac{1}{2} f^H(P_- x') &\geq \frac{h_{d'}}{2}\left\|P_+ x'\right\|^2 - \frac{h_{d'+1}}{2}\left\|P_- x'\right\|^2 \\ &\geq \frac{1}{2}\min\{h_{d'}, -h_{d'+1}\}\left\|x'\right\|^2. \end{aligned}$$

One can hence obtain that

$$f^H(x'_+) = (1+\rho^H)^2 f^H(P_+ x') + f^H(P_- x') = \left(2\rho^H + (\rho^H)^2\right) f^H(P_+ x') \geq c\left\|x'\right\|^2,$$

with c being the constant in (4.13), and similarly that

$$f^H(x'_-) = f^H(P_+ x') + (1+\rho^H)^2 f^H(P_- x') = \left(2\rho^H + (\rho^H)^2\right) f^H(P_- x') \leq -c\left\|x'\right\|^2.$$

Notice also $x'_+, x'_- \in B(x', \rho^H \left\|x'\right\|) \subseteq B(0, 3c')$. It holds that

$$f^H(x'_+) \geq \frac{c}{4}\left\|x'_+\right\|^2, \quad f^H(x'_-) \leq -\frac{c}{4}\left\|x'_-\right\|^2.$$

which combined with (4.14) yields that

$$f(x'_+) > 0 > f(x'_-).$$

Therefore, there exists $x'' \in B(x', \rho^H \left\|x'\right\|)$ so that

$$f(x'') = 0.$$

We also have that

$$\left\|x - x''\right\| \leq \left\|x - x'\right\| + \left\|x' - x''\right\| < 2\rho^H \left\|x'\right\| < 4\rho^H \left\|x''\right\| < \rho\left\|x''\right\|,$$

which leads to that

$$x \in B(x'', \rho\left\|x''\right\|) \subseteq U_2(\rho).$$

The proof is thus completed. □

Proof of Lemma 4.4. Let $\rho^H \in (0, \rho/4)$ and denote $U_2^H = U_2^H(\rho^H)$. By Lemma 4.6, there exists a neighborhood U' of 0, such that $U_2^H \cap U' \subseteq U_2 \cap U'$. Then by Lemma 4.5, there exist constants $p_+^H > 0$ and $p_-^H < 0$ so that for $x \in U' \setminus U_2 \subseteq U' \setminus U_2^H$, one has either $f^H(x) \geq p_+^H \left\|x\right\|^2$ or $f^H(x) \leq p_-^H \left\|x\right\|^2$. Let $U \subseteq U_1 \cap U'$ be a neighborhood of 0 so that

$$\left|f(x) - f^H(x)\right| < p\left\|x\right\|^2, \quad \forall\, x \in U,$$

where $p = \frac{1}{2}\min\{p_+^H, -p_-^H\}$. Therefore, we have for any $x \in U \setminus U_2$ that, either $f(x) \geq p\|x\|^2$ or $f(x) \leq -p\|x\|^2$.

Consider any $x \in \big(U \cap f^{-1}([0,+\infty))\big) \setminus S$. Note that $S = U_1 \cap U_2 \cap f^{-1}([0,+\infty))$ and that $U \subseteq U_1$. We have $x \in U \setminus U_2$ and $f(x) \geq 0$, which excludes $f(x) \leq -p\|x\|^2$ and leads to $f(x) \geq p\|x\|^2$. □

References

[1] Ludwig Arnold, *Random dynamical systems*, Springer monographs in mathematics, Springer, New York, 1998.

[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** (2010), no. 2, 438–457.

[3] Zhong-Zhi Bai, *On greedy randomized coordinate updating iteration methods for solving symmetric eigenvalue problems*, Applied Numerical Mathematics **216** (2025), 76–97.

[4] Amir Beck and Luba Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM journal on Optimization **23** (2013), no. 4, 2037–2060.

[5] Radu Ioan Boţ, Minh N Dao, and Guoyin Li, *Inertial proximal block coordinate method for a class of nonsmooth sum-of-ratios optimization problems*, SIAM Journal on Optimization **33** (2023), no. 2, 361–393.

[6] Petra Boxler, *A stochastic version of center manifold theory*, Probability Theory and Related Fields **83** (1989), no. 4, 509–545.

[7] Xufeng Cai, Chaobing Song, Stephen Wright, and Jelena Diakonikolas, *Cyclic block coordinate descent with variance reduction for composite nonconvex optimization*, International conference on machine learning, 2023, pp. 3469–3494.

[8] Ziang Chen, Yingzhou Li, and Jianfeng Lu, *On the global convergence of randomized coordinate gradient descent for nonconvex optimization*, SIAM Journal on Optimization **33** (2023), no. 2, 713–738.

[9] Miklós Csörgő, *On the strong law of large numbers and the central limit theorem for martingales*, Transactions of the American Mathematical Society **131** (1968), no. 1, 259–275.

[10] Zhiyan Ding, Taehee Ko, Jiahao Yao, Lin Lin, and Xiantao Li, *Random coordinate descent: A simple alternative for optimizing parameterized quantum circuits*, Phys. Rev. Res. **6** (2024Jul), 033029.

[11] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle points—online stochastic gradient for tensor decomposition*, Conference on learning theory, 2015, pp. 797–842.

[12] Victor Guillemin and Alan Pollack, *Differential topology*, Vol. 370, American Mathematical Soc., 2010.

[13] Peng Guo and Jun Shen, *Smooth center manifolds for random dynamical systems*, Rocky Mountain Journal of Mathematics **46** (2016), no. 6, 1925 –1962.

[14] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan, *How to escape saddle points efficiently*, International conference on machine learning, 2017, pp. 1724–1732.

[15] Chi Jin, Praneeth Netrapalli, and Michael I Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, Conference on learning theory, 2018, pp. 1042–1085.

[16] Steven George Krantz and Harold R Parks, *The implicit function theorem: history, theory, and applications*, Springer Science & Business Media, 2002.

[17] Ching-Pei Lee and Stephen J Wright, *Random permutations fix a worst case for cyclic coordinate descent*, IMA Journal of Numerical Analysis **39** (2019), no. 3, 1246–1275.

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

[19] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht, *Gradient descent only converges to minimizers*, Conference on learning theory, 2016, pp. 1246–1257.

[20] Weigu Li and Kening Lu, *Sternberg theorems for random dynamical systems*, Communications on Pure and Applied Mathematics **58** (2005), no. 7, 941–988, available at `https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20083`.

[21] Yingzhou Li, Jianfeng Lu, and Zhe Wang, *Coordinatewise descent methods for leading eigenvalue problem*, SIAM Journal on Scientific Computing **41** (2019), no. 4, A2681–A2716.

[22] Ji Liu and Stephen J Wright, *Asynchronous stochastic coordinate descent: Parallelism and convergence properties*, SIAM Journal on Optimization **25** (2015), no. 1, 351–376.

[23] Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar, *An asynchronous parallel stochastic coordinate descent algorithm*, International conference on machine learning, 2014, pp. 469–477.

[24] Jun Liu and Ye Yuan, *Almost sure convergence rates analysis and saddle avoidance of stochastic gradient methods*, Journal of Machine Learning Research **25** (2024), no. 271, 1–40.

[25] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie, *Sparsenet: Coordinate descent with nonconvex penalties*, Journal of the American Statistical Association **106** (2011), no. 495, 1125–1138.

[26] Yu Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization **22** (2012), no. 2, 341–362.

[27] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke, *Coordinate descent converges faster with the gauss-southwell rule than random selection*, International conference on machine learning, 2015, pp. 1632–1641.

[28] Michael O'Neill and Stephen J Wright, *Behavior of accelerated gradient methods near critical points of nonconvex functions*, Mathematical Programming **176** (2019), 403–427.

[29] Liangzu Peng and René Vidal, *Block coordinate descent on smooth manifolds: Convergence theory and twenty-one examples*, arXiv preprint arXiv:2305.14744 (2023).

[30] David Ruelle, *Ergodic theory of differentiable dynamical systems*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques **50** (1979), no. 1, 27–58.

[31] ———, *Characteristic exponents and invariant manifolds in hilbert space*, Annals of Mathematics (1982), 243–290.

[32] Ankan Saha and Ambuj Tewari, *On the nonasymptotic convergence of cyclic coordinate descent methods*, SIAM Journal on Optimization **23** (2013), no. 1, 576–601.

[33] Hao-Jun Michael Shi, Shenyinying Tu, Yangyang Xu, and Wotao Yin, *A primer on coordinate descent algorithms*, arXiv preprint arXiv:1610.00040 (2016).

[34] Michael Shub, *Global stability of dynamical systems*, Springer Science & Business Media, 2013.

[35] Quoc Tran-Dinh and Yang Luo, *Randomized block-coordinate optimistic gradient algorithms for root-finding problems*, Mathematics of Operations Research (2025).

[36] Zhe Wang, Yingzhou Li, and Jianfeng Lu, *Coordinate descent full configuration interaction*, Journal of chemical theory and computation **15** (2019), no. 6, 3558–3569.

[37] Zhe Wang, Zhiyuan Zhang, Jianfeng Lu, and Yingzhou Li, *Coordinate descent full configuration interaction for excited states*, Journal of Chemical Theory and Computation **19** (2023), no. 21, 7731–7739.

[38] Stephen Wright and Ching-pei Lee, *Analyzing random permutations for cyclic coordinate descent*, Mathematics of computation **89** (2020), no. 325, 2217–2248.

[39] Stephen J Wright, *Coordinate descent algorithms*, Mathematical programming **151** (2015), no. 1, 3–34.

[40] Yuejia Zhang, Weiguo Gao, and Yingzhou Li, *Parallel multicoordinate descent methods for full configuration interaction*, Journal of Chemical Theory and Computation **21** (2025), no. 5, 2325–2337.

(ZC) Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Email address: ziang@mit.edu

(YL) School of Mathematical Sciences, Fudan University, Shanghai 200433, China

Email address: yingzhouli@fudan.edu.cn

(ZL) School of Mathematical Sciences, Fudan University, Shanghai 200433, China

Email address: zhli24@m.fudan.edu.cn