SIAM J. OPTIM.
Vol. 33, No. 2, pp. 713–738

ON THE GLOBAL CONVERGENCE OF RANDOMIZED COORDINATE GRADIENT DESCENT FOR NONCONVEX OPTIMIZATION*

ZIANG CHEN†, YINGZHOU LI‡, AND JIANFENG LU§

Abstract. In this work, we analyze the global convergence property of a coordinate gradient descent with random choice of coordinates and stepsizes for nonconvex optimization problems. Under generic assumptions, we prove that the algorithm iterate will almost surely escape strict saddle points of the objective function. As a result, the algorithm is guaranteed to converge to local minima if all saddle points are strict. Our proof is based on viewing the coordinate descent algorithm as a nonlinear random dynamical system and a quantitative finite block analysis of its linearization around saddle points.

Key words. randomized coordinate gradient descent, global convergence, strict saddle point, random dynamical system

MSC codes. 90C26, 37N40

DOI. 10.1137/21M1460375

1. Introduction.

In this paper, we analyze the global convergence of a coordinate gradient descent algorithm for a smooth but nonconvex optimization problem:

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

More specifically, we consider coordinate gradient descent with random coordinate selection and random stepsizes, as shown Algorithm 1.1.

Algorithm 1.1 Randomized coordinate gradient descent

Initialization: $x_0 \in \mathbb{R}^d$, $t = 0$.
while not convergent **do**
  Draw a coordinate i_t uniformly random from $\{1, 2, \ldots, d\}$.
  Draw a stepsize α_t uniformly random in $[\alpha_{\min}, \alpha_{\max}]$.
  $x_{t+1} \leftarrow x_t - \alpha_t e_{i_t} \partial_{i_t} f(x_t)$.
  $t \leftarrow t + 1$.
end while

* Received by the editors November 19, 2021; accepted for publication (in revised form) November 27, 2022; published electronically June 9, 2023.
https://doi.org/10.1137/21M1460375
Funding: This work is supported in part by the National Science Foundation via grants DMS-2012286 and CHE-2037263 and by the U.S. Department of Energy via grant DE-SC0019449. The second author is partially supported by National Natural Science Foundation of China under grant 12271109.
† Corresponding author. Department of Mathematics, Duke University, Durham, NC 27708 USA (ziang@math.duke.edu).
‡ School of Mathematical Sciences, Fudan University, Shanghai 200433, People's Republic of China (yingzhouli@fudan.edu.cn).
§ Department of Mathematics, Physics, and Chemistry, Duke University, Durham, NC 27708 USA (jianfeng@math.duke.edu).

The main result of this paper, Theorem 1, is that for any initial guess x_0 that is not a strict saddle point of f, under some mild conditions, with probability 1, Algorithm 1.1 will escape any strict saddle points, and thus, under some additional structural assumptions of f, the algorithm will globally converge to a local minimum.

In order to establish the global convergence, we view the algorithm as a random dynamical system and carry out the analysis based on the theory of random dynamical systems. This might be of separate interest; in particular, to the best of our knowledge, the theory of random dynamical system has not been utilized in analyzing randomized algorithms, while it offers a natural framework to establish long time behavior of such algorithms. Let us now briefly explain the random dynamical system view of the algorithm and our analysis; more details can be found in section 3.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space for all randomness used in the algorithm such that each $\omega \in \Omega$ is a sequence of coordinates and stepsizes. The iterate of Algorithm 1.1 can be described as a random dynamical system $x_t = \varphi(t,\omega)x_0$, where $\varphi(t,\omega) : \mathbb{R}^d \to \mathbb{R}^d$ is a nonlinear map for any given $t \in \mathbb{N}$ and $\omega \in \Omega$.

Consider an isolated stationary point x^* of the dynamical system, which corresponds to a critical point of f. Near x^*, the dynamical system can be approximated by its linearization: $x_t = \Phi(t,\omega)x_0$, where $\Phi(t,\omega) \in \mathbb{R}^{d\times d}$. The limiting behavior of the linear dynamical system can be well understood by the celebrated multiplicative ergodic theorem: Under some assumptions, the limit $\Lambda(\omega) = \lim_{t\to\infty}(\Phi(t,\omega)^\top \Phi(t,\omega))^{1/2t}$ exists almost surely. The eigenvalues of the matrix $\Lambda(\omega)$, $e^{\lambda_1(\omega)} > e^{\lambda_2(\omega)} > \cdots > e^{\lambda_{p(\omega)}(\omega)}$, characterize the long time behavior of the system. In particular, if the largest Lyapunov exponent $\lambda_1(\omega)$ is strictly positive, then if x_0 has some nontrivial component in the unstable subspace, $x_t = \Phi(t,\omega)x_0$ would exponentially diverge from x^*. More details of the preliminaries of the linear random dynamical system can be found in section 2.

Intuitively, one expects that the nonlinear dynamical system can be approximated by its linearization around a critical point x^* and would hence escape the strict saddle point, following the linearized system. However, the approximation by the linear dynamical system cannot hold for an infinite time horizon due to error accumulation. Therefore, we cannot naively conclude using the multiplicative ergodic theorem and the linear approximation. Instead, a major part of the analysis is devoted to establish a quantitative finite block analysis of the behavior of the dynamical system over a finite time interval. In particular, we will prove that when the iterate is in a neighborhood of x^*, the distance $\|x_t - x^*\|$ will be exponentially amplified for a duration T with high probability. This would then be used to prove that with probability 1, the nonlinear system will escape strict saddle points.

1.1. Related work. Coordinate gradient descent is a popular approach in optimization; see e.g., the review articles [55, 46]. Advantages of the coordinate gradient method include that compared with the full gradient descent, it allows larger stepsize [36] and enjoys faster convergence [45], and it is also friendly for parallelization [32, 42].

The convergence of coordinate gradient descent has been analyzed in several settings on the property of the objective function and on the strategy of coordinate selection. The understanding of convergence for convex problems is quite complete: For methods with cyclic choice of coordinates, the convergence has been established in [4, 45, 49], and the worst-case complexity is investigated when the objective function is convex and quadratic in [50]. For methods with random choice of coordinates, it is shown in [36] that $\mathbb{E}f(x_t)$ converges to $f^* = \min_{x\in\mathbb{R}^d} f(x)$ sublinearly in the convex case and linearly in the strongly convex case. Convergence of objective function in

high probability has also been established in [36]. We also refer the reader to [41, 33, 32, 55] for further convergence results for random coordinate selection for convex problems. More recently, convergence of methods with random permutation of coordinates (i.e., a random permutation of the d coordinates is used for every d step of the algorithm) have been analyzed, mostly for the case of quadratic objective functions [21, 38, 15, 56]. It has been an ongoing research direction to compare various coordinate selection strategies in various settings. In addition, in the nonconvex and nonsmooth setting, the convergence of coordinate/alternating descent methods can be analyzed for tame/semialgebraic functions with Kurdyka–Łojasiewicz property (see, e.g., [3, 2, 6, 7]).

For nonconvex objective functions, the global convergence analysis is less developed, as the situation becomes more complicated. Escaping strict saddle points has been a focused research topic in nonconvex optimization, motivated by applications in machine learning. It has been established that various first-order algorithms with gradient noise or added randomness to iterates would escape strict saddle points; see, e.g., [11, 24, 16, 18, 17, 14] for works in this direction.

Among previous works for escaping saddle points, perhaps the closest in spirit to our current result are [23, 39, 22, 30], where algorithms without gradient or iterate randomness are studied. It is proved in [23] that for almost every initial guess, the trajectory of the gradient descent algorithm (without any randomness) with constant stepsize would not converge to a strict saddle point. The result has been extended in [22] to a broader class of deterministic first-order algorithms, including coordinate gradient descent with cyclic choice of coordinate. The global convergence result for cyclic coordinate gradient descent is also proved in [30] under slightly more relaxed conditions. A similar convergence result is also obtained for the heavy-ball method in [39]. Let us emphasize that in the case of coordinate algorithms, it is not merely a technical question whether the algorithm can escape the strict saddle points without randomly perturbing gradients or iterates. In fact, one simply cannot employ such random perturbations, e.g., adding a random Gaussian vector to the iterate, since doing so would destroy the coordinate nature of the algorithm.

The analysis in the works [23, 22, 39, 30] is based on viewing the algorithm as a deterministic dynamical system, and applying the center-stable manifold theorem for deterministic dynamical system [47], which characterizes the local behavior near a stationary point of nonlinear dynamical systems. Such a framework obviously does not work for randomized algorithms. To some extent, our analysis can be understood as a natural generalization to the framework of random dynamical systems, which allows us to analyze the long time behavior of randomized algorithms, in particular, coordinate gradient descent with random coordinate selection.

Let us mention that various stable, unstable, and center manifold theorems have been established in the literature of random dynamical systems; see, e.g., [1, 43, 44, 8, 34]. These sample-dependent random manifolds also characterize the local behavior of random dynamical systems. However, as far as we can tell, one cannot simply apply such "off-the-shelf results" for the analysis of Algorithm 1.1. Instead, for study of the algorithm, we have to carry out a quantitative finite block analysis for the random dynamical system near the stationary points. Our proof technique is inspired by stability analysis of the Lyapunov exponent of random dynamical systems, as in [20, 10].

1.2. Organization. The rest of this paper will be organized as follows. In section 2, we review the preliminaries of random dynamical system for the convenience of the reader. Our main result is stated in section 3. The proofs can be found in section 4.

2. Preliminaries of random dynamical systems. In this section, we recall basic notions and results of random dynamical systems; for more details, we refer the reader to standard references, such as [1]. After introducing the preliminaries in this section, we will define the random dynamical system associated with Algorithm 1.1 in section 3.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathbb{T}$ be a semigroup with $\mathcal{B}(\mathbb{T})$ being its Borel σ-algebra. $\mathbb{T}$ serves as the notion of time. In the setting of Algorithm 1.1, we have $\mathbb{T} = \mathbb{N}$, corresponding to the one-sided discrete time setting. Other possible examples of $\mathbb{T}$ include $\mathbb{T} = \mathbb{Z}$, $\mathbb{T} = \mathbb{R}_{\geq 0}$, and $\mathbb{T} = \mathbb{R}$, with the assumption that $0 \in \mathbb{T}$.

Let us first define a random dynamical system. As we have mentioned in the introduction, the dynamics starting from x_0 can be determined once a sample $\omega \in \Omega$ is fixed. From the viewpoint of a random dynamical system, specifying the dynamics of x is equivalent to specifying the dynamics of ω: Suppose at time 0 that the dynamics corresponds to ω. Then to prescribe the future dynamics starting from time t, we can specify the corresponding $\theta(t)\omega \in \Omega$ for some map $\theta(t) : \Omega \to \Omega$. More precisely, we have the following definition of dynamics on Ω.

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*

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

The random dynamical system can then be defined as follows.

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*

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

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) is a key property of a random dynamical system: After time s, if we restart the system at x_s, the future dynamic corresponds to the sample $\theta(s)\omega$. Note that $\varphi(t, \omega, \cdot)$ is a map on X. With some ambiguity of notation, we also use $\varphi(t, \omega)$ to denote this map on X and write $\varphi(t, \omega)x = \varphi(t, \omega, x)$. Then the cocycle property (2.1) can be written as

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

In this work, we will focus on the one-sided discrete time $\mathbb{T} = \mathbb{N}$ and $\theta(t) = \theta^t$, where θ is $\mathbb{P}$-preserving and θ^t is the t-fold composition of θ. Suppose that $X = \mathbb{R}^d$ and $A : \Omega \to \mathrm{GL}(d, \mathbb{R})$ is measurable. Consider a linear random dynamical system defined as (we use Φ for the linear system while reserving φ for nonlinear dynamics considered later)

$$\Phi(t, \omega) = A(\theta^{t-1}\omega) \cdots A(\theta\omega)A(\omega),$$

where the right-hand side is the product of a sequences of random matrices. In this setting, the behavior of the linear system $x_t = \Phi(t, \omega)x_0$ is well understood by the celebrated multiplicative ergodic theorem, also known as the Oseledets theorem, which we recall in Theorem 2.3. Such a type of result was first established by V. I. Oseledets [37] and was further developed in many works, such as [40, 43, 52].

THEOREM 2.3 (multiplicative ergodic theorem [1, Theorem 3.4.1]). *Suppose that*

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

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

(i) *It holds that the limit*

$$\Lambda(\omega) = \lim_{t\to\infty} \left(\Phi(t, \omega)^\top \Phi(t, \omega)\right)^{1/2t} \tag{2.2}$$

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* $e^{\lambda_1(\omega)} > e^{\lambda_2(\omega)} > \cdots > e^{\lambda_{p(\omega)}(\omega)}$. *Denote* $V_i(\omega)$ *the corresponding eigenspace 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) *Set* $W_i(\omega) = \bigoplus_{j\geq i} V_j(\omega)$, $i = 1, 2, \ldots, p(\omega)$, *and* $W_{p(\omega)+1}(\omega) = \{0\}$. *Then it holds that*

$$\lim_{t\to\infty} \frac{1}{t} \log \|\Phi(t, \omega)x\| = \lambda_i(\omega), \quad \forall\ x \in W_i(\omega)\backslash W_{i+1}(\omega) \tag{2.3}$$

for $i = 1, 2, \ldots, p(\omega)$. *The maps* $V(\cdot)$ *and* $W(\cdot)$ *from* $\widetilde{\Omega}$ *to the Grassmannian manifold are measurable.*

(iv) *It holds that*

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

(v) *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 Theorem 2.3, $\lambda_1(\omega) > \lambda_2(\omega) > \cdots > \lambda_{p(\omega)}(\omega)$ are known as Lyapunov exponents, and $\{0\} \subseteq W_{p(\omega)}(\omega) \subseteq \cdots \subseteq W_1(\omega) \subseteq \mathbb{R}^d$ is the Oseledets filtration. We can see from the above theorem that both the Lyapunov exponents and the Oseledets filtration are A-forward invariant.

The Lyapunov exponents describe the asymptotic growth rate of $\|\Phi(t, \omega)x\|$ as $t \to \infty$. More specifically, (2.3) implies that when $x \in W_i(\omega)\backslash W_{i+1}(\omega)$ for any $\epsilon > 0$, there exists some $T > 0$ such that

$$e^{t(\lambda_i(\omega)-\epsilon)} \le \|\Phi(t,\omega)x\| \le e^{t(\lambda_i(\omega)+\epsilon)}$$

holds for any $t > T$. The subspaces spanned by eigenvectors of $\Lambda(\omega)$ corresponding to eigenvalues smaller than, equal to, and greater than 0 are the stable subspace, center subspace, and unstable subspace, respectively. The stable and unstable subspaces correspond to exponential convergence and exponential divergence, respectively. When starting from the center subspace, we would get some subexponential behavior.

The multiplicative ergodic theorem also generalizes to continuous time and two-sided time. We refer the interested reader to [1, Theorems 3.4.1 and 3.4.11] for details.

The stable, unstable, and center subspaces can be generalized to stable, unstable, and center manifolds when considering nonlinear systems; see, e.g., [1, 43, 44, 8, 27, 34, 31, 25, 13]. These manifolds play similar roles in characterizing the local behavior of nonlinear random dynamical systems as the subspaces for linear random dynamical systems. In particular, the Hartman–Grobman theorem establishes the topological conjugacy between a nonlinear system and its linearization [53]. There are also other conjugacy results for random dynamical systems; see, e.g., [27, 26, 28, 29].

3. Main results.

3.1. Setup of the random dynamical system. Let us first specify the random dynamical system corresponding to the Algorithm 1.1.

- *Probability space.* For each $t \in \mathbb{N}$, denote $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ the usual probability space for the distribution $\mathcal{U}_{\{1,2,\dots,n\}} \times \mathcal{U}_{[\alpha_{\min},\alpha_{\max}]}$, where $\mathcal{U}_{\{1,2,\dots,n\}}$ and $\mathcal{U}_{[\alpha_{\min},\alpha_{\max}]}$ are the uniform distributions on the set $\{1,2,\dots,n\}$ and interval $[\alpha_{\min},\alpha_{\max}]$, respectively. Let $(\Omega,\mathcal{F},\mathbb{P})$ be the product probability space of all $(\Omega_t,\Sigma_t,\mathbb{P}_t),\ t\in\mathbb{N}$. Denote π_t as the projection from $(\Omega,\mathcal{F},\mathbb{P})$ onto $(\Omega_t,\Sigma_t,\mathbb{P}_t),\ t\in\mathbb{N}$. Thus, a sample $\omega\in\Omega$ can be represented as a sequence $((i_0,\alpha_0),(i_1,\alpha_1),\dots)$, where $(i_t,\alpha_t)=\pi_t(\omega),\ t\in\mathbb{N}$. Let $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be the filtration defined by

$$\mathcal{F}_t = \sigma\left\{(B_0\times\cdots\times B_t)\times\left(\prod_{j>t}\Omega_j\right) : B_i\in\Sigma_i,\ i=0,1,\dots,t\right\}.$$

- *Metric dynamical system.* The metric dynamical system on Ω is constructed by the (left) shifting operator $\tau:\Omega\to\Omega$ defined as

$$\tau(\omega)=\tau(\pi_0(\omega),\pi_1(\omega),\cdots):=(\pi_1(\omega),\pi_2(\omega),\cdots),$$

  which is clearly measurable and $\mathbb{P}$-preserving. The metric dynamical system is then given by $\theta(t)=\tau^t$ for $t\in\mathbb{N}$.
- *Random dynamical system.* For any $\omega\in\Omega$ and $t\in\mathbb{N}$, we define $\phi(\omega)$ to be a (nonlinear) map on $\mathbb{R}^d$ as

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

  where $(i,\alpha)=\pi_0(\omega)$ is the first pair/element in the sequence ω, and we define the map $\varphi(t,\omega)$ via

$$\varphi(t,\omega)=\phi(\tau^{t-1}\omega)\circ\cdots\circ\phi(\tau\omega)\circ\phi(\omega),\quad \text{for } t\ge 1,$$

while $\varphi(0,\omega)$ is the identity operator. It is clear that $\varphi(t,\omega)$ satisfies the cocycle property (2.1) and hence defines a random dynamical system on $X = \mathbb{R}^d$ over $\{\tau^t\}_{t\in\mathbb{N}}$. The iterate of Algorithm 1.1 follows the random dynamical system as

$$x_t = \phi(\tau^{t-1}\omega)x_{t-1} = \cdots = \phi(\tau^{t-1}\omega)\circ\cdots\circ\phi(\tau\omega)\circ\phi(\omega)x_0 = \varphi(t,\omega)x_0.$$

It can be seen that $\{x_t\}_{t\in\mathbb{N}}$ is $\{\mathcal{F}_t\}$-predictable, i.e., x_t is $\mathcal{F}_{t-1}$-measurable for any $t\in\mathbb{N}_+$, since x_t is determined by samples $(i_0,\alpha_0),(i_1,\alpha_1),\ldots,(i_{t-1},\alpha_{t-1})$.

In our analysis, we will use linearization of the dynamical system $\varphi(t,\omega)$ at a critical point x^* of f. Without loss of generality, we assume $x^* = 0$; otherwise, we consider the system with state being $x - x^*$. The resulting linear system, which depends on $H = \nabla^2 f(x^*) = (H_{ij})_{1\le i,j\le d}$, is given by (here and in the following, we use the superscript H to indicate dependence on the matrix)

$$\Phi^H(t,\omega) = A^H(\tau^{t-1}\omega)\cdots A^H(\tau\omega)A^H(\omega), \tag{3.1}$$

where

$$A^H(\omega) = I - \alpha e_i e_i^\top H, \quad (i,\alpha) = \pi_0(\omega). \tag{3.2}$$

Note that $A^H(\cdot)$ is bounded in Ω. We know that $(\log\|A^H(\cdot)\|)_+$ is integrable. When $\alpha < 1/|H_{ii}|$, the matrix $A^H(\omega) = I - \alpha e_i e_i^\top H$ is invertible, and the inverse is given explicitly by applying the Sherman–Morrison formula:

$$A^H(\omega)^{-1} = \left(I - \alpha e_i e_i^\top H\right)^{-1} = I + \frac{\alpha e_i e_i^\top H}{1-\alpha H_{ii}}. \tag{3.3}$$

In particular, we have

$$\|A^H(\omega)^{-1}\| \le 1 + \frac{\alpha\|H\|}{1-\alpha|H_{ii}|}. \tag{3.4}$$

Thus, if we take the maximal stepsize $\alpha_{\max}$ such that $\alpha_{\max} < 1/\max_{1\le i\le d}|H_{ii}|$, $\|A^H(\cdot)^{-1}\|$ is bounded in Ω, and as a result, $(\log\|A^H(\cdot)^{-1}\|)_+$ is also integrable. Therefore, the assumptions of Theorem 2.3 hold. The shifting operator τ is ergodic on $(\Omega,\mathcal{F},\mathbb{P})$ by Kolmogorov's 0–1 law. Then Theorem 2.3 applies for $\theta = \tau$ with $p^H(\cdot)$, $\lambda_i^H(\cdot)$, and $d_i^H(\cdot)$ all being a.e. constant. For any $\omega\in\widetilde{\Omega}$ that is the set in Theorem 2.3 satisfying $\mathbb{P}(\widetilde{\Omega}) = 1$, we denote

$$W_+^H(\omega) = \bigoplus_{\lambda_i>0} V_i^H(\omega), \text{ and } W_-^H(\omega) = \bigoplus_{\lambda_i\le 0} V_i^H(\omega). \tag{3.5}$$

Then the following invariant property holds:

$$W_-^H(\tau\omega) = A^H(\omega)W_-^H(\omega).$$

Note that $W_-^H(\omega)$ works as a center-stable subspace. That is, for any $x\in W_-^H(\omega)$ and any $\epsilon>0$, it holds that $\|\Phi(t,\omega)x\| \le e^{t\epsilon}$ for sufficiently large t and for $x\notin W_-^H(\omega)$, $\|\Phi(t,\omega)x\|$ grows exponentially as $t\to\infty$ with rate greater than $\min_{\lambda_i>0}\lambda_i - \epsilon$ for any $\epsilon>0$.

3.2. Assumptions. In this section, we specify the assumptions of the objective function f in this paper. The first is a standard smoothness assumption of f.

Assumption 3.1. $f \in C^2(\mathbb{R}^d)$ and the Hessian $\nabla^2 f$ is uniformly bounded; i.e., there exists $M > 0$ such that $\|\nabla^2 f(x)\| \le M$ for all $x \in \mathbb{R}^d$.

An optimization algorithm is expected to converge, under some reasonable assumptions, to a critical point of f where the gradient vanishes. Our aim is to further characterize the possible limits of the algorithm iterates. For this purpose, we distinguish $\mathrm{Crit}_s(f)$, the set of all strict saddle points (including local maxima with the nondegenerate Hessian) of f,

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

where we use the subscript s to emphasize that it is strict. Due to the presence of a negative eigenvalue of Hessian, if we were considering the gradient dynamics near the critical point, the saddle point would be an unstable equilibrium. Our first result is that this instability also occurs in the linear random dynamical system $\Phi^H(t,\omega)$, where $H = \nabla^2 f(x^*)$. In other words, the dimension of $W_+^H(\omega)$ defined in (3.5) is greater than 0. While this would mainly serve as a preliminary step for our analysis of the nonlinear dynamics, the conclusion by itself might be of interest and is stated as follows. The proof will be deferred to section 4.1.

PROPOSITION 3.2. *Let H have a negative eigenvalue and $0 < \alpha_{\min} < \alpha_{\max} < 1/\max_{1\le i\le d}|H_{ii}|$. Then the largest Lyapunov exponent of $\Phi^H(t,\omega)$ is positive.*

Our goal is to generalize such results to the nonlinear dynamics near strict saddle points of f, for which we would require two additional assumptions as follows.

Assumption 3.3. For every $x^* \in \mathrm{Crit}_s(f)$, $\nabla^2 f(x^*)$ is nondegenerate; i.e., x^* is a nondegenerate critical point of f in the sense that any eigenvalue of $\nabla^2 f(x^*)$ is nonzero.

Assumption 3.3 is also a standard assumption, which, in particular, guarantees that each strict saddle point is isolated due to the nondegenerate Hessian. For each strict saddle point, Proposition 3.2 guarantees that the corresponding unstable subspace $W_+^H(\omega)$ is nontrivial (has dimension at least 1). We would in fact require a stronger technical assumption on its structure.

Assumption 3.4. 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 $W_+^H(\omega)$ with $H = \nabla^2 f(x^*)$.

We expect that Assumption 3.4 holds generically. We also show in Appendix A that Assumption 3.4 can be verified when H has no zero off-diagonal elements (and $W_+^H(\omega)$ is not trivial). However, there exist cases that Assumption 3.4 might not hold. One example is $H = \nabla^2 f(x^*) = \mathrm{diag}(H_1, H_2)$, where $H_1 \in \mathbb{R}^{d_1\times d_1}$ only has positive eigenvalues and $H_2 \in \mathbb{R}^{d_2\times d_2}$ only has negative eigenvalues, which implies that $W_-^H(\omega) = \mathrm{span}\{e_i : 1 \le i \le d_1\}$ and $W_+^H(\omega) = \mathrm{span}\{e_i : d_1 + 1 \le i \le d\}$.

We also remark that Assumption 3.1 is essential in our framework since the linearized system is defined using the Hessian matrix. Analysis of the randomized coordinate method for nonsmooth optimization problems requires new techniques and deserves future research.

3.3. Main results. Given an initial guess x_0, the behavior of the algorithm, in particular, the limit of x_t, depends on the particular sample $\omega \in \Omega$. For any

$x^* \in \mathrm{Crit}_s(f)$, we denote the set of all ω such that the algorithm starting at x_0 would converge to x^*:

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

We further define the set $\Omega(\mathrm{Crit}_s(f), x_0)$ as the union of all $\Omega(x^*, x_0)$ over $x^* \in \mathrm{Crit}_s(f)$:

$$\Omega(\mathrm{Crit}_s(f), x_0) := \bigcup_{x^*\in\mathrm{Crit}_s(f)} \Omega(x^*, x_0).$$

Thus, if $\omega \notin \Omega(\mathrm{Crit}_s(f), x_0)$, the limit $\lim_{t\to\infty} x_t$, if it exists, will not be one of the strict saddle points. Our main result in this paper proves that the set is of measure zero; i.e., for any initial guess x_0 that is not a strict saddle point, with probability 1, Algorithm 1.1 will not converge to a strict saddle point.

THEOREM 1. *Suppose that Assumptions* 3.1, 3.3, *and* 3.4 *hold and that* $0 < \alpha_{\min} < \alpha_{\max} < 1/M$. *Then for any* $x_0 \in \mathbb{R}^d \backslash \mathrm{Crit}_s(f)$, *it holds that*

$$\mathbb{P}(\Omega(\mathrm{Crit}_s(f), x_0)) = 0.$$

The intuition behind the proof of Theorem 1 is to compare the nonlinear dynamics around a strict saddle point $x^* \in \mathrm{Crit}_s(f)$ with its linearization, as the linear dynamics has nontrivial unstable subspace, thanks to Proposition 3.2. Ideally, one would hope that the nonlinear dynamics would closely follow the linear dynamics and thus leave the neighborhood of x^* eventually; the obstacle for such an argument is, however, that the approximation of the linearization is only valid for a finite time interval. Therefore, to establish the instability behavior of the nonlinear dynamics, we would need a much more refined and quantitative argument using the instability of the linear system. In particular, we would need to show that over a finite interval, with high probability, the linear system and hence the nonlinear system would drive x_t away from the strict saddle point with quantitative bounds; see Theorem 4.4 in section 4.2. Theorem 1 then follows from an argument with a similar spirit as the law of large numbers; see section 4.3.

Remark 3.5. The technical Assumption 3.4 and the randomness in stepsizes are made so that the iterate $x_t = x_{t-1} - \alpha_{t-1} e_{i_{t-1}} e_{i_{t-1}}^\top \nabla f(x_{t-1})$ would obtain some nontrivial component in the unstable subspace, which would be further amplified within a sufficiently long but finite time interval. When $\|\mathcal{P}_+^H(\tau^t\omega)e_{i_{t-1}}\|$ and $|e_{i_{t-1}}^\top \nabla f(x_{t-1})|$ are fixed and relatively large, a random α_{t-1} would keep $\|\mathcal{P}_+^H(\tau^t\omega)x_t\|$ away from 0 with high probability; see section 4.2 for more details. It is an interesting open question whether it is possible to establish similar results without such assumptions. Our conjecture is that $\mathbb{P}(\Omega_s(x_0)) = 0$ still holds for $x_0 \in \mathbb{R}^d \backslash \mathrm{Crit}_s(f)$ unless x_0 is located in a set with Lebesgue measure zero, similar to the results established in [22].

As an application of our main result (Theorem 1), we can obtain the global convergence to stationary points with no negative Hessian eigenvalues for Algorithm 1.1. More specifically, denote by

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

the set of all critical points of f. Then we have the following corollary, which will also be proved in section 4.3.

COROLLARY 3.6. *Under the same assumptions as in Theorem* 1 *and assuming further that every* $x^* \in \mathrm{Crit}(f)$ *is an isolated critical point, for any* $x_0 \in \mathbb{R}^d \backslash \mathrm{Crit}_s(f)$ *with bounded level set* $L(x_0) = \{x \in \mathbb{R}^d : f(x) \le f(x_0)\}$*, with probability* 1*,* $\{x_t\}_{t\in\mathbb{N}}$ *is convergent with limit in* $\mathrm{Crit}(f)\backslash\mathrm{Crit}_s(f)$*.*

Remark 3.7. In Corollary 3.6, if we further assume that all saddle points of f are strict, then the algorithm iterate converges to a local minimum with probability 1. Let us also mention that for many nonconvex problems, saddle points are suboptimal, while there do not exist "bad" local minima, e.g., phase retrieval [48], deep learning [19, 35], and low-rank matrix problems [12]. For these problems, convergence to local minima suffices to guarantee good performance.

Remark 3.8. In our setting, without adding noise to the gradient or iterate, we cannot hope for good convergence rates for an arbitrary initial iterate. In fact, as shown in [9], the convergence of a deterministic gradient descent algorithm to a local minimum might take an exponentially long time; we expect similar behavior for the randomized coordinate gradient descent Algorithm 1.1. Let us also remark that while we need the random stepsize as discussed in Remark 3.5, the interval $[\alpha_{\min}, \alpha_{\max}]$ could be made arbitrarily small; the result holds as long as $0 < \alpha_{\min} < \alpha_{\max} < 1/M$.

4. Proofs.

We collect all proofs in this section.

4.1. Analysis of the linearized system.

We will first study the linear dynamical system, for which we assume the objective function is given by

$$f^H(x) = \frac{1}{2} x^\top H x, \tag{4.1}$$

where H is a symmetric matrix in $\mathbb{R}^{d\times d}$ with at least one negative eigenvalue. In this case, the coordinate descent algorithm is given by

$$x_{t+1} = \left(I - \alpha_t e_{i_t} e_{i_t}^\top H\right) x_t,$$

which corresponds to the linear dynamical system $\Phi^H(t,\omega)$ with single step map $A^H(\omega)$, defined in (3.1) and (3.2), respectively.

Our main goal in this subsection is to prove Proposition 3.2 for this linear dynamical system, which states that at least one Lyapunov exponent of $\Phi^H(t,\omega)$ is positive. It suffices to show that there exists some x_0 such that $\|x_t\|$ grows exponentially to infinity, which will follow from an energy argument, similar to the proof of [22, Proposition 5]. Although we consider a randomized coordinate gradient descent algorithm instead of a cyclic one, one step, i.e., Lemma 4.3, in the proof of Proposition 4.1 follows closely the proof in [22, Appendix A]. We start from x_0 with $f^H(x_0) < 0$ and consider a finite time interval with length $m \ge d$. Proposition 4.1 establishes a quantitative decay estimate for $f^H(x_{t+m})$ compared with $f^H(x_t)$, which leads to our desired result (Proposition 3.2).

PROPOSITION 4.1. *Let* $m \ge d$ *be fixed. For the objective function* (4.1) *with* $\lambda_{\min}(H) < 0$*, suppose that* $0 < \alpha_{\min} < \alpha_{\max} < 1/\max_{1\le i\le d} |H_{ii}|$ *and that there exists* $c \in (0,1)$ *depending on* m*,* H*,* $\alpha_{\min}$*, and* $\alpha_{\max}$ *such that*

$$f^H(x_{t+m}) - f^H(x_t) \le c\, f^H(x_t)$$

holds as long as $f^H(x_t) < 0$ *and* $\{1, 2, \ldots, d\} = \{i_t, i_{t+1}, \ldots, i_{t+m-1}\}$ *(in the sense of sets).*

Remark 4.2. The condition $\{1,2,\dots,d\}=\{i_t,i_{t+1},\dots,i_{t+m-1}\}$ above is known as the "generalized Gauss–Seidel rule" in the literature of coordinate methods [51, 54].

Proof. Without loss of generality, we assume that $t=0$. Due to the choice of $\alpha_{\max}$, we have the following simple nonincreasing property for any $t'\in\mathbb{N}$:

$$\begin{aligned} f^H(x_{t'+1}) &= \frac{1}{2}x_{t'+1}^\top Hx_{t'+1} \\ &= \frac{1}{2}x_{t'}^\top\left(I-\alpha_{t'}e_{i_{t'}}e_{i_{t'}}^\top H\right)^\top H\left(I-\alpha_{t'}e_{i_{t'}}e_{i_{t'}}^\top H\right)x_{t'} \\ &= f^H(x_{t'})-\alpha_{t'}\left(e_{i_{t'}}^\top Hx_{t'}\right)^2+\frac{1}{2}\alpha_t^2e_{i_{t'}}^\top He_{i_{t'}}\left(e_{i_{t'}}^\top Hx_{t'}\right)^2 \\ &\le f^H(x_{t'})-\frac{\alpha_{t'}}{2}\left(e_{i_{t'}}^\top Hx_{t'}\right)^2. \end{aligned} \tag{4.2}$$

Write $x_0=y^*+y_0$ with $y^*\in\ker(H)$ and $y_0\in\operatorname{ran}(H)$. Let

$$y_{t'+1}=y_{t'}-\alpha_{t'}e_{i_{t'}}e_{i_{t'}}^\top Hy_{t'},\qquad t'=0,1,\dots,m-1. \tag{4.3}$$

Then $x_{t'}=y^*+y_{t'}$ holds for any $t'=0,1,\dots,m$. Using (4.2), to give an upper bound for $f^H(x_{t+m})-f^H(x_t)$, we would like a nontrivial lower bound for $\alpha_{t'}(e_{i_{t'}}^\top Hx_{t'})^2=\alpha_{t'}(e_{i_{t'}}^\top Hy_{t'})^2$ for some $t'\in\{t,t+1,\dots,t+m-1\}$, which is guaranteed by Lemma 4.3, whose proof will be postponed.

LEMMA 4.3. *Suppose that* $\{1,2,\dots,d\}=\{i_0,i_1,\dots,i_{m-1}\}$. *For any*

$$0<\delta\le\min\left\{\frac{1}{2m},\frac{\alpha_{\min}\sigma_{\min}(H)}{2\sqrt{m}(m\alpha_{\max}\sigma_{\max}(H)+1)}\right\}, \tag{4.4}$$

where $\sigma_{\min}(H)$ *and* $\sigma_{\max}(H)$ *are the smallest and largest positive singular values of* H, *respectively, if* $0\ne y_0\in\operatorname{ran}(H)$, *then there exists* $T\in\{0,1,\dots,m-1\}$ *such that* $\alpha_T|e_{i_T}^\top Hy_T|\ge\delta\|y_T\|$, *where the sequence* y_t *is given as in* (4.3).

Assuming the lemma, there exists $T\in\{0,1,\dots,m-1\}$ such that

$$\alpha_T\left|e_{i_T}^\top Hy_T\right|\ge\delta\|y_T\|,$$

with a fixed $\delta>0$ satisfying (4.4). We can further constrain that $\frac{\delta^2}{\alpha_{\min}\sigma_{\max}(H)}<1$. Thus, we have

$$f^H(x_m)\le f^H(x_{T+1})\le f^H(x_T)-\frac{\alpha_T}{2}\left(e_{i_T}^\top Hx_T\right)^2\le f^H(x_T)-\frac{\delta^2}{2\alpha_T}\|y_T\|^2,$$

which, combined with $\sigma_{\max}(H)\|y_T\|^2\ge -y_T^\top Hy_T=-x_T^\top Hx_T=-2f^H(x_T)$, yields that

$$f^H(x_m)\le\left(1+\frac{\delta^2}{\alpha_T\sigma_{\max}(H)}\right)f^H(x_T)\le\left(1+\frac{\delta^2}{\alpha_T\sigma_{\max}(H)}\right)f^H(x_0).$$

Set $c=\frac{\delta^2}{\alpha_{\max}\sigma_{\max}(H)}$, and we get that $f^H(x_m)-f^H(x_0)\le cf^H(x_0)$.

We finish the proof by establishing Lemma 4.3 below.

Proof of Lemma 4.3. Suppose on the contrary that $\alpha_t|e_{i_t}^\top H y_t| < \delta\|y_t\|$ for any $t \in \{0,1,\dots,m-1\}$. It holds that

$$\|y_1 - y_0\| = \alpha_0 \left|e_{i_0}^\top H y_0\right| < \delta\|y_0\| < 2\delta\|y_0\|.$$

We claim that

$$\|y_t - y_0\| < 2t\delta\|y_0\| \tag{4.5}$$

for any $t = 1,2,\dots,m-1$. By induction, assume that $\|y_t - y_0\| < 2t\delta\|y_0\|$ holds for some $t \in \{1,2,\dots,m-2\}$. Then

$$\|y_{t+1} - y_t\| = \alpha_t \left|e_{i_t}^\top H y_t\right| < \delta\|y_t\| < \delta(2t\delta+1)\|y_0\| < 2\delta\|y_0\|,$$

where the last inequality uses $2t\delta < 2m\delta \le 1$. It follows that $\|y_{t+1} - y_0\| \le \|y_t - y_0\| + \|y_{t+1} - y_t\| < 2(t+1)\delta\|y_0\|$.

Using (4.5) and $\max_{1\le i\le d}\|He_i\| \le \sigma_{\max}(H)$, we have

$$\begin{aligned}\alpha_t\left|e_{i_t}^\top H y_0\right| &\le \alpha_t\left|e_{i_t}^\top H(y_t - y_0)\right| + \alpha_t\left|e_{i_t}^\top H y_t\right| \\ &< \alpha_{\max}\sigma_{\max}(H)\cdot 2t\delta\|y_0\| + 2\delta\|y_0\| \\ &< 2\delta(m\alpha_{\max}\sigma_{\max}(H)+1)\|y_0\|\end{aligned}$$

for $t = 0,1,\dots,m-1$. Since $\operatorname{span}\{e_{i_k} : k = 0,1,\dots,m-1\} = \mathbb{R}^d$, noticing that $y_0 \in \operatorname{ran} H$, we have

$$\begin{aligned}\alpha_{\min}\sigma_{\min}(H)\|y_0\| \le \alpha_{\min}\|Hy_0\| &\le \left(\sum_{t=0}^{m-1}\left(\alpha_t\left|e_{i_t}^\top H y_0\right|\right)^2\right)^{1/2} \\ &< 2\delta\sqrt{m}(m\alpha_{\max}\sigma_{\max}(H)+1)\|y_0\|,\end{aligned}$$

which contradicts the choice of δ in (4.4). ∎

We are now ready to prove Proposition 3.2, which states the existence of a positive Lyapunov exponent of the linear dynamical system.

Proof of Proposition 3.2. It suffices to show that for almost every $\omega \in \Omega$, there exist some $x_0 \in \mathbb{R}^d$, $\epsilon > 0$, and $T > 0$ such that $x_t = \Phi(t,\omega)x_0$ satisfies $\|x_t\| \ge e^{\epsilon t}$ for any $t > T$. Let x_0 be an eigenvector corresponding to a negative eigenvalue of H. Then it holds that $f^H(x_0) < 0$. Consider a fixed $m \ge d$. For any $k \in \mathbb{N}$, set $I_k = 1$ if $\{1,2,\dots,d\} = \{i_{km}, i_{km+1},\dots,i_{km+m-1}\}$ and $I_k = 0$ otherwise. We can see that the random variables $I_0, I_1, I_2, \dots$ are independent and identically distributed with $\mathbb{E}I_0 = \mathbb{P}(I_0 = 1) \in (0,1)$. By Proposition 4.1, we obtain that

$$f^H(x_{(k+1)m}) \le \begin{cases}(1+c)f^H(x_{km}) & \text{if } I_k = 1,\\ f^H(x_{km}) & \text{if } I_k = 0,\end{cases}$$

where c is the constant from Proposition 4.1. Therefore,

$$\frac{\lambda_{\min}(H)}{2}\|x_{km}\|^2 \le f^H(x_{km}) \le (1+c)^{\sum_{j=0}^{k-1} I_j} f^H(x_0),$$

which implies that

$$\|x_{km}\| \ge \left(\frac{2f^H(x_0)}{\lambda_{\min}(H)}\right)^{1/2}\cdot(1+c)^{\frac{1}{2}\sum_{j=0}^{k-1} I_j}. \tag{4.6}$$

Note that $\mathbb{E}|I_0| = \mathbb{E}I_0 < \infty$. The strong law of large number suggests that for almost every $\omega \in \Omega$, there exists some K such that for all $k \geq K$,

$$\sum_{j=0}^{k-1} I_j \geq \frac{\mathbb{E}I_0}{2} k. \tag{4.7}$$

Combining (4.6) and (4.7), we arrive at

$$\|x_{km}\| \geq \left(\frac{2f^H(x_0)}{\lambda_{\min}(H)}\right)^{1/2} \cdot (1+c)^{\frac{\mathbb{E}I_0}{4m} \cdot km}.$$

Noticing that $(1+c)^{\frac{\mathbb{E}I_0}{4m}}$ is greater than 1, $\|x_{km}\|$ grows exponentially in km, and we complete the proof. ∎

4.2. Finite block analysis. In this subsection, we study the behavior of the nonlinear dynamical system near a strict saddle point of f, which, without loss of generality, can be assumed to be $x^* = 0$. As mentioned above, in a small neighborhood of x^*, while it is not possible to control the difference between nonlinear and linear systems for infinite time, the nonlinear system can be approximated by the linear system during a finite time horizon.

The main conclusion of this subsection is the following theorem, which states that after a finite time interval with length T, the distance of the iterate from $x^* = 0$ will be amplified exponentially with high probability.

THEOREM 4.4. *Suppose that Assumptions* 3.1, 3.3, *and* 3.4 *hold and that* $0 < \alpha_{\min} < \alpha_{\max} < 1/M$. *There exists* $\epsilon_* \in (0, 1/6)$ *such that for any* $\epsilon \in (0, \epsilon_*)$, *we have* $T_* = T_*(\epsilon) \in \mathbb{N}_+$, *and for any* $T \in \mathbb{N}_+$ *with* $T \geq T_*$ *and any* $t \in \mathbb{N}$, *conditioned on* $\mathcal{F}_{t-1}$, *with probability at least* $1 - 4\epsilon$, *it holds for all* $x_t \in V$ *that*

$$\|x_{t+T}\| \geq \exp\left(\frac{6\epsilon}{1-6\epsilon}\left|\log(1 - M\alpha_{\max})\right| T\right)\|x_t\|, \tag{4.8}$$

where V *is a neighborhood of* $x^* = 0$, *depending on* ϵ, T, *and* f *near* x^*.

The lower bound (4.8) quantifies the amplification of $\|x_{t+T}\|$: While we always have $\|x_{t+T}\| \geq (1 - M\alpha_{\max})^T \|x_t\|$ (see (4.20) below), the result states that with probability at least $1 - 4\epsilon$, the amplification factor is at least the right-hand side of (4.8), which is exponentially large in T. Hence, on average, $\|x_{t+T}\|$ would be much larger than $\|x_t\|$. This would lead to the escaping of the iterate from the neighborhood of $x^* = 0$.

To prove Theorem 4.4, we would require a more quantitative characterization of the behavior of its linearization at x^*. In particular, we need a high probability estimate of the distance of the iterate from x^* after some time interval. For this purpose, conditioned on $\mathcal{F}_{t-1}$ with the iterate x_t, we will first show in Lemma 4.8 that, after some finite time, the orthogonal projection of the iterate x_{ϱ_t} onto the unstable subspace, where $t < \varrho_t \leq t + L$ for some constant L, is significant. The component in the unstable subspace would then be further amplified subsequently by $\Phi^H(S, \tau^{\varrho_t}\omega)$, where $H = \nabla^2 f(x^*)$. Here the time duration S would be chosen sufficiently large such that the distance from x_{ϱ_t+S} to x^* is exponentially amplified. Theorem 4.4 follows by setting $T = L + S$. In the second step above, we would need to control the closeness between the linear and nonlinear systems within a time horizon with length S.

Such a finite block analysis approach has been used to establish the stability of the Lyapunov exponent of random dynamical systems [20, 10], which inspired our proof technique for Theorems 4.4 and 1.

We first set the small constant ϵ in Theorem 4.4, which controls the failure probability of the amplification bound. Let $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ be the Lyapunov exponents of the linearized system at $x^* = 0$. We set

$$\lambda_+ = \min_{\lambda_i > 0} \lambda_i \quad \text{and} \quad 0 < \gamma < \frac{1}{2} \min \left\{ \min_{1 \le i < p} |\lambda_i - \lambda_{i+1}|, \lambda_+ \right\}. \tag{4.9}$$

Note that $\gamma < \lambda_+$. Let $\epsilon_* \in (0, 1/6)$ be sufficiently small such that

$$(1 - 6\epsilon)(\lambda_+ - \gamma) + 6\epsilon \cdot \log(1 - M\alpha_{\max}) > 0, \quad \forall\, \epsilon \in (0, \epsilon_*). \tag{4.10}$$

The reason for such choice will become clear later (see (4.23)). For the rest of the section, we will consider a fixed $\epsilon \in (0, \epsilon_*)$.

We now state and prove several lemmas for Theorem 4.4. First, in the following Lemma 4.5, we construct a stopping time $\varrho_t - 1$ that is bounded almost surely, and the component of the gradient $|e_{i_{\varrho_t - 1}}^\top \nabla f(x_{\varrho_t - 1})|$ is comparable with $\|\nabla f(x_{\varrho_t - 1})\|$ in amplitude with high probability.

LEMMA 4.5. *Let $0 < \mu \le \frac{1}{\sqrt{d}}$ be a fixed constant. There exists some constant $L > 0$ such that for any $t \in \mathbb{N}$, there exists a measurable $\varrho_t : \Omega \to \mathbb{N}_+$ such that $t < \varrho_t \le t + L$ and*

$$\mathbb{P}\Big(\big| e_{i_{\varrho_t - 1}}^\top \nabla f(x_{\varrho_t - 1}) \big| \ge \mu \|\nabla f(x_{\varrho_t - 1})\| \,\Big|\, \mathcal{F}_{t-1} \Big) \ge 1 - \epsilon. \tag{4.11}$$

Proof. For any $t \in \mathbb{N}$, use ℓ_0 to denote the smallest nonnegative integer ℓ such that

$$\ell_0 = \arg\min_{\ell} \Big\{ \ell \in \mathbb{N}_+ : \big| e_{i_{t+\ell-1}}^\top \nabla f(x_{t+\ell-1}) \big| \ge \mu \|\nabla f(x_{t+\ell-1})\| \Big\}.$$

It is clear that $\mathbb{P}(\ell_0 > \ell \mid \mathcal{F}_{t-1}) \le (1 - 1/d)^\ell$ since for each step, the coordinate is randomly chosen. Hence, there exists some $L > 0$ such that

$$\mathbb{P}(\ell_0 \le L \mid \mathcal{F}_{t-1}) \ge 1 - \epsilon.$$

We finish the proof by setting $\varrho_t = t + \min\{\ell_0, L\}$, which has the desired property. ∎

We now carry out the amplification part of the finite block analysis for the linearized dynamics at $x^* = 0$. To simplify expressions in the following, for $t_1 < t_2$, we introduce the shorthand notation

$$(i, \alpha)_{t_1 : t_2 - 1} = \big((i_{t_1}, \alpha_{t_1}), \ldots, (i_{t_2 - 1}, \alpha_{t_2 - 1}) \big) \in \Omega_{t_1} \times \cdots \times \Omega_{t_2 - 1}$$

and the finite time transition matrix (i.e., composition of linear maps)

$$\Phi^H\big((i, \alpha)_{t_1 : t_2 - 1} \big) = (I - \alpha_{t_2 - 1} e_{i_{t_2 - 1}} e_{i_{t_2 - 1}}^\top H) \cdots (I - \alpha_{t_1} e_{i_{t_1}} e_{i_{t_1}}^\top H).$$

Recall that $(\Omega_t, \Sigma_t, \mathbb{P}_t)$ is the probability space for $\mathcal{U}_{\{1,2,\ldots,d\}} \times \mathcal{U}_{[\alpha_{\min}, \alpha_{\max}]}$ for $t \in \mathbb{N}$. We also denote $\mathcal{P}_+^H((i, \alpha)_{t_1 : t_2 - 1})$ as the projection operator onto the subspace spanned by the right singular vectors of $\Phi^H((i, \alpha)_{t_1 : t_2 - 1})$ corresponding to d_+ largest singular

values, where $d_+ = \sum_{\lambda_i>0} d_i$ and d_i is the dimension of the ith eigenspace as in Theorem 2.3 (ii) for the linearized system at x^*.

As we mentioned in the proof sketch, we want $\Phi^H(S, \tau^{\varrho_t}\omega) = \Phi^H((i,\alpha)_{\varrho_t:\varrho_t+S-1})$ to amplify x_{ϱ_t}, for which we need to establish a nontrivial lower bound for the unstable component $\|\mathcal{P}_+^H((i,\alpha)_{\varrho_t:\varrho_t+S-1})x_{\varrho_t}\|$. This is achieved by several lemmas. We will establish three lower bounds in the following:

- $\|\mathcal{P}_+^H(\tau^{\varrho_t}\omega)e_{i_{\varrho_t-1}}\|$ using Lemma 4.6;
- $\|\mathcal{P}_+^H((i,\alpha)_{\varrho_t:\varrho_t+S-1})e_{i_{\varrho_t-1}}\|$ in Lemma 4.7; and finally the desired
- $\|\mathcal{P}_+^H((i,\alpha)_{\varrho_t:\varrho_t+S-1})x_{\varrho_t}\|$ in Lemma 4.8.

Let us first control $\|\mathcal{P}_+^H(\tau^{\varrho_t}\omega)e_{i_{\varrho_t-1}}\|$ in the following lemma, which utilizes Assumption 3.4. For simplicity of notation, in Lemmas 4.6 and 4.7, we state the results for $\|\mathcal{P}_+^H(\omega)e_i\|$ and $\|\mathcal{P}_+^H((i,\alpha)_{0:S-1})e_j\|$ instead, which is slightly more general.

LEMMA 4.6. *Under Assumption* 3.4*, there exist* $\delta > 0$ *and measurable* $\Omega_1^\epsilon \subset \widetilde{\Omega}$*, where* $\widetilde{\Omega}$ *is from Theorem* 2.3*, such that* $\mathbb{P}(\Omega_1^\epsilon) \geq 1-\epsilon$ *and*

$$\|\mathcal{P}_+^H(\omega)e_i\| \geq \delta, \quad \forall\ \omega \in \Omega_1^\epsilon,\ i \in \{1,2,\dots,d\}.$$

Proof. Assumption 3.4 implies that

$$\mathbb{P}\big(\{\omega \in \widetilde{\Omega} : \|\mathcal{P}_+^H(\omega)e_i\| > 0, \quad \forall\ i \in \{1,2,\dots,d\}\}\big) = 1.$$

Notice that

$$\big\{\omega \in \widetilde{\Omega} : \|\mathcal{P}_+^H(\omega)e_i\| > 0, \quad \forall\ i \in \{1,2,\dots,d\}\big\} = \bigcup_{n\in\mathbb{N}_+} \Big\{\omega \in \widetilde{\Omega} : \|\mathcal{P}_+^H(\omega)e_i\| \geq \frac{1}{n}, \quad \forall\ i \in \{1,2,\dots,d\}\Big\}.$$

The lemma follows from continuity of measure. ∎

We will then be able to handle $\|\mathcal{P}_+^H((i,\alpha)_{0:S-1})e_j\|$ using Lemma 4.6 and the closeness between $(\Phi^H(S,\omega)^\top\Phi^H(S,\omega))^{1/2S}$ with $\Lambda(\omega)$ as the former converges to the latter as $S \to \infty$ by Theorem 2.3. More precisely, denote the singular values of $X \in \mathbb{R}^{d\times d}$ by $s_1(X) \geq s_2(X) \geq \cdots \geq s_d(X)$. Then for $S \in \mathbb{N}_+$ sufficiently large, we have

$$\left|\frac{1}{S}\log s_j\left(\Phi^H(S,\omega)\right) - \lambda_{\mu(j)}\right| = \left|\frac{1}{S}\log s_j\left(\Phi^H\big((i,\alpha)_{0:S-1}\big)\right) - \lambda_{\mu(j)}\right| \leq \gamma \tag{4.12}$$

for every $j \in \{1,2,\dots,d\}$, where $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ are the Lyapunov exponents from Theorem 2.3, γ is given by (4.9), and the map $\mu : \{1,2,\dots,d\} \to \{1,2,\dots,p\}$ satisfies that $\mu(j) = i$ if and only if $d_1 + \cdots + d_{i-1} < j \leq d_1 + \cdots + d_i$, so μ corresponds the index for the singular values with that of the Lyapunov exponents. Moreover, the convergence also implies that

$$\|\mathcal{P}_+^H(S,\omega) - \mathcal{P}_+^H(\omega)\| \leq \frac{\delta}{2}$$

for sufficiently large S, which then leads to

$$\|\mathcal{P}_+^H(S,\omega)\, e_j\| = \|\mathcal{P}_+^H((i,\alpha)_{0:S-1})\, e_j\| \geq \frac{\delta}{2} \tag{4.13}$$

for every $j \in \{1,2,\dots,d\}$, where $\mathcal{P}_+^H(S,\omega)$ is the projection operator onto the subspace spanned by the right singular vectors of $\Phi^H(S,\omega)$ corresponding to d_+ largest singular values. Let

$$\Omega^S = \big\{(i,\alpha)_{0:S-1} \in \Omega_0 \times \cdots \times \Omega_{S-1} : (4.12) \text{ and } (4.13) \text{ hold}\big\}. \tag{4.14}$$

The following lemma states that Ω^S has high probability for sufficiently large S, where, with slight abuse of notation, we write $\mathbb{P}(\Omega^S) = \mathbb{P}(\Omega^S \times (\bigtimes_{t\geq S}\Omega_t))$.

LEMMA 4.7. *Under the same assumptions of Lemma* 4.6*, there exists some* $S_* > 0$ *such that for every* $S \in \mathbb{N}_+, S \geq S_*$*, it holds that* $\mathbb{P}(\Omega^S) \geq 1 - 2\epsilon$*.*

Proof. For *a.e.* $\omega \in \Omega$, it follows from Theorem 2.3, in particular (2.2), and standard matrix perturbation analysis (see, e.g., [5, Theorems VI.2.1 and VII.3.1]) that

$$\frac{1}{S} s_j(\Phi^H(S,\omega)) \to \lambda_{\mu(j)}, \quad S \to \infty, \tag{4.15}$$

for any $j \in \{1, 2, \ldots, d\}$ and that

$$\mathcal{P}_+^H(S,\omega) \to \mathcal{P}_+^H(\omega), \quad S \to \infty. \tag{4.16}$$

By Egorov's theorem, there exists $\Omega_2^\epsilon \subset \Omega_1^\epsilon$ with $\mathbb{P}(\Omega_2^\epsilon) \geq 1 - 2\epsilon$ such that the convergences in (4.15) and (4.16) are both uniform on Ω_2^ϵ. Here Ω_1^ϵ is as in Lemma 4.6. Therefore, for some S_* sufficiently large, we have

$$\left|\frac{1}{S}\log s_j\left(\Phi^H(S,\omega)\right) - \lambda_{\mu(j)}\right| \leq \gamma, \quad \forall\ j \in \{1,2,\ldots,d\},\ S \geq S_*,\ \omega \in \Omega_2^\epsilon,$$

and

$$\|\mathcal{P}_+^H(S,\omega) - \mathcal{P}_+^H(\omega)\| \leq \frac{\delta}{2}, \quad \forall\ S \geq S_*,\ \omega \in \Omega_2^\epsilon. \tag{4.17}$$

Combining Lemma 4.6 and (4.17), we obtain that

$$\|\mathcal{P}_+^H(S,\omega)e_i\| \geq \frac{\delta}{2}, \quad \forall\ i \in \{1,2,\ldots,d\},\ \forall\ S \geq S_*,\ \omega \in \Omega_2^\epsilon.$$

For any $S \geq S_*$, by the definition of Ω^S, it holds that

$$\Omega_2^\epsilon \subset \Omega^S \times \Big(\bigtimes_{t\geq S}\Omega_t\Big),$$

which implies the desired estimate

$$\mathbb{P}(\Omega^S) = \mathbb{P}\bigg(\Omega^S \times \Big(\bigtimes_{t\geq S}\Omega_t\Big)\bigg) \geq \mathbb{P}(\Omega_2^\epsilon) \geq 1 - 2\epsilon. \qquad \square$$

Note that $\alpha_{\varrho_t - 1} \sim \mathcal{U}_{[\alpha_{\min}, \alpha_{\max}]}$ is independent of $\mathcal{F}_{\varrho_t - 2}$, $i_{\varrho_t - 1}$, and $(i,\alpha)_{\varrho_t : \varrho_t + S - 1}$. The next lemma shows that with high probability, the choice of $\alpha_{\varrho_t - 1}$ will lead to a nontrivial orthogonal projection of x_{ϱ_t} onto the unstable subspace of $\Phi^H(S, \tau^{\varrho_t}\omega) = \Phi^H((i,\alpha)_{\varrho_t : \varrho_t + S - 1})$.

LEMMA 4.8. *For any* $S \in \mathbb{N}_+$*,* $x_{\varrho_t - 1}$*,* $i_{\varrho_t - 1}$*, and* $(i,\alpha)_{\varrho_t : \varrho_t + S - 1} \in \Omega^S$*, there exists* $I \subset [\alpha_{\min}, \alpha_{\max}]$ *with* $m(I) \geq (1-\epsilon)(\alpha_{\max} - \alpha_{\min})$*, where* $m(\cdot)$ *is the Lebesgue measure, such that for any* $\alpha_{\varrho_t - 1} \in I$*, it holds that*

$$\|\mathcal{P}_+^H\left((i,\alpha)_{\varrho_t : \varrho_t + S - 1}\right) x_{\varrho_t}\| \geq \frac{\epsilon\delta(\alpha_{\max} - \alpha_{\min})}{4} \left|e_{i_{\varrho_t - 1}}^\top \nabla f(x_{\varrho_t - 1})\right|. \tag{4.18}$$

Proof. We assume that $|e_{i_{\varrho_t-1}}^\top \nabla f(x_{\varrho_t-1})| \neq 0$; otherwise, the result is trivial. For simplicity of notation, we write

$$\begin{aligned}\mathcal{P}_+^H\left((i,\alpha)_{\varrho_t:\varrho_t+S-1}\right)x_{\varrho_t} &= \mathcal{P}_+^H\left((i,\alpha)_{\varrho_t:\varrho_t+S-1}\right)x_{\varrho_t-1}\\ &\quad - \alpha_{\varrho_t-1}e_{i_{\varrho_t-1}}^\top \nabla f(x_{\varrho_t-1})\mathcal{P}_+^H\left((i,\alpha)_{\varrho_t:\varrho_t+S-1}\right)e_{i_{\varrho_t-1}}\\ &=: y_2 - \alpha_{\varrho_t-1}y_1,\end{aligned}$$

where the last line defines y_1 and y_2. Using the shorthand notation

$$r = \frac{\epsilon\delta(\alpha_{\max}-\alpha_{\min})}{4}\left|e_{i_{\varrho_t-1}}^\top \nabla f(x_{\varrho_t-1})\right|,$$

we then observe that (4.18) holds if and only if $\alpha_{\varrho_t-1}y_1$ is not located in a ball with radius r centered at y_2.

It follows from the definition of Ω^S and (4.13) that $\|\mathcal{P}_+^H\left((i,\alpha)_{\varrho_t:\varrho_t+S-1}\right)e_{i_{\varrho_t-1}}\| \geq \frac{\delta}{2}$, which then leads to

$$\|y_1\| \geq \frac{\delta}{2}\left|e_{i_{\varrho_t-1}}^\top \nabla f(x_{\varrho_t-1})\right| = \frac{2r}{\epsilon(\alpha_{\max}-\alpha_{\min})}.$$

Thus, the set of α_{ϱ_t-1} such that $\alpha_{\varrho_t-1}y_1 \in \mathcal{B}_r(y_2)$ consists of an interval J in $\mathbb{R}$ with $\|\sup(J)\cdot y_1 - \inf(J)\cdot y_1\| \leq 2r$, as the diameter of $\mathcal{B}_r(y_2)$ is $2r$, which implies that $m(J) \leq 2r/\|y_1\| \leq \epsilon(\alpha_{\max}-\alpha_{\min})$. The lemma is proved then by setting $I = [\alpha_{\max}-\alpha_{\min}]\backslash J$. ∎

With Lemmas 4.5, 4.6, 4.7, and 4.8, we now prove Theorem 4.4, which relies on approximation of the nonlinear dynamics by linearization and the amplification from the finite block analysis for the linearized system.

Proof of Theorem 4.4. Without loss of generality, we will assume $t = 0$ in the proof to simplify notation. Since $H = \nabla^2 f(x^*)$ is nondegenerate, we can take a neighborhood U of $x^* = 0$ and some fixed $\sigma > 0$ such that

$$\|\nabla f(x)\| \geq \sigma\|x\|, \quad \forall\ x \in U. \tag{4.19}$$

Assumption 3.1 implies that

$$\|\nabla f(x)\| = \|\nabla f(x) - \nabla f(x^*)\| \leq M\|x - x^*\| = M\|x\|, \quad \forall\ x \in \mathbb{R}^d.$$

Using the above inequality and $\alpha_{\max} < 1/M$, it holds for every $\omega \in \Omega$ and $t' \in \mathbb{N}$ that

$$\begin{aligned}\|x_{t'+1}\| &= \|x_{t'} - \alpha_{t'}e_{i_{t'}}e_{i_{t'}}^\top \nabla f(x_{t'})\|\\ &\geq \|x_{t'}\| - \alpha_{t'}\|e_{i_{t'}}e_{i_{t'}}^\top\| \cdot \|\nabla f(x_{t'})\|\\ &\geq (1 - M\alpha_{\max})\|x_{t'}\|,\end{aligned} \tag{4.20}$$

and similarly that

$$\|x_{t'+1}\| \leq (1 + M\alpha_{\max})\|x_{t'}\|.$$

We thus define

$$r_- := 1 - M\alpha_{\max} \quad \text{and} \quad r_+ := 1 + M\alpha_{\max},$$

so that

$$r_-\|x_{t'}\| \le \|x_{t'+1}\| \le r_+\|x_{t'}\|. \tag{4.21}$$

We now choose the time duration S large enough in the finite block analysis to guarantee significant amplification. More specifically, we choose S so large that $S \ge S_*$ (S_* defined in Lemma 4.7) and that the following two inequalities hold:

$$\exp(S(\lambda_+ - \gamma)) \cdot \frac{\epsilon\delta\mu\sigma(r_-)^{L-1}(\alpha_{\max} - \alpha_{\min})}{8} \ge (r_+)^L \tag{4.22}$$

and

$$(1-6\epsilon)\left(S(\lambda_+ - \gamma) + \log\frac{\epsilon\delta\mu\sigma(r_-)^{2(L-1)}(\alpha_{\max} - \alpha_{\min})}{8}\right) + 6\epsilon(L+S)\log r_- > 0, \tag{4.23}$$

where L is the upper bound defined in Lemma 4.5, $\mu \le 1/\sqrt{d}$ is a fixed constant as in Lemma 4.5, δ is from Lemma 4.6, and σ is set in (4.19). Thanks to (4.10) for our choice of ϵ and that $\gamma < \lambda_+$ from (4.9), (4.22) and (4.23) are satisfied for sufficiently large S.

Next, we show that for any S sufficiently large as above, there exists a convex neighborhood $U_1 \subset U$ of $x^* = 0$ such that for any $t' \in \mathbb{N}$, any $x_{t'} \in U_1$, and any $(i,\alpha)_{t':t'+S-1}$, it holds that

$$\|x_{t'+S}\| \ge \|\Phi^H\big((i,\alpha)_{t':t'+S-1}\big)x_{t'}\| - \|x_{t'}\|. \tag{4.24}$$

We first define a convex neighborhood $U_0 \subset U$ of $x^* = 0$ such that

$$\begin{aligned}\|\big(x - \alpha e_i e_i^\top \nabla f(x)\big) - \big(I - \alpha e_i e_i^\top H\big)x\| &= \|\alpha e_i e_i^\top (\nabla f(x) - Hx)\| \\ &= \|\alpha e_i e_i^\top \int_0^1 (\nabla^2 f(\eta x) - Hx)\mathrm{d}\eta\| \le \frac{1}{S(r_+)^{S-1}}\|x\|\end{aligned}$$

for any $x \in U_0$, any $i \in \{1,2,\dots,d\}$, and any $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. Applying the inequality S times for $x_{t'} \in U_1 = (r_+)^{-(S-1)}U_0$, we have

$$\begin{aligned}&\|x_{t'+S} - \Phi^H\big((i,\alpha)_{t':t'+S-1}\big)x_{t'}\| \\ &\le \|x_{t'+S} - \left(I - \alpha_{t'+S-1}e_{i_{t'+S-1}}e_{i_{t'+S-1}}^\top H\right)x_{t'+S-1}\| \\ &\quad + \|I - \alpha_{t'+S-1}e_{i_{t'+S-1}}e_{i_{t'+S-1}}^\top H\| \cdot \|x_{t'+S-1} - \Phi^H\big((i,\alpha)_{t':t'+S-2}\big)x_{t'}\| \\ &\le \frac{1}{S(r_+)^{S-1}}\|x_{t'+S-1}\| + r_+\|x_{t'+S-1} - \Phi^H\big((i,\alpha)_{t':t'+S-2}\big)x_{t'}\| \\ &\le \frac{1}{S(r_+)^{S-1}}\left(\|x_{t'+S-1}\| + r_+\|x_{t'+S-2}\| + \cdots + (r_+)^{S-1}\|x_{t'}\|\right) \\ &\le \|x_{t'}\|\end{aligned}$$

and hence inequality (4.24).

Setting $V = (r_+)^{-(L-1)}U_1$, we then have $x_{\varrho_0-1} \in U$ for any $x_0 \in V$, which implies that $\|\nabla f(x_{\varrho_0-1})\| \ge \sigma\|x_{\varrho_0-1}\|$ as $\varrho_0 \le L$. According to Lemmas 4.5, 4.6, 4.7, and 4.8, for any given $x_0 \in V$, with probability at least $1 - 4\epsilon$, we have $(i,\alpha)_{\varrho_0:\varrho_0+S-1} \in \Omega^S$, and the following holds:

$$\left|e_{i_{\varrho_0-1}}^\top \nabla f(x_{\varrho_0-1})\right| \geq \mu \|\nabla f(x_{\varrho_0-1})\|, \tag{4.25}$$

$$\|\mathcal{P}_+^H\left((i,\alpha)_{\varrho_0:\varrho_0+S-1}\right) x_{\varrho_0}\| \geq \frac{\epsilon\delta(\alpha_{\max}-\alpha_{\min})}{4}\left|e_{i_{\varrho_0-1}}^\top \nabla f(x_{\varrho_0-1})\right|, \tag{4.26}$$

where the probability is the marginal probability on $(i,\alpha)_{0:L+S-1} \in \Omega_0 \times \cdots \times \Omega_{L+S-1}$.

Recall λ_+ and γ in (4.9) and that $d_+ = \sum_{\lambda_i>0} d_i$. It follows from (4.12) of the construction of the set Ω^S that

$$\frac{1}{S}\log s_j\left(\Phi^H\left((i,\alpha)_{\varrho_0:\varrho_0+S-1}\right)\right) \geq \lambda_{\mu(j)} - \gamma \geq \lambda_+ - \gamma, \quad \forall\ j \leq d_+.$$

This is to say that the d_+ largest singular values of $\Phi^H((i,\alpha)_{\varrho_0:\varrho_0+S-1})$ are all greater than or equal to $\exp(S(\lambda_+-\gamma))$. Therefore, it holds that

(4.27)

$$\begin{aligned}
\|\Phi^H\left((i,\alpha)_{\varrho_0:\varrho_0+S-1}\right) x_{\varrho_0}\| &\geq \|\Phi^H\left((i,\alpha)_{\varrho_0:\varrho_0+S-1}\right)\mathcal{P}_+^H\left((i,\alpha)_{\varrho_0:\varrho_0+S-1}\right) x_{\varrho_0}\| \\
&\overset{(4.26)}{\geq} \exp(S(\lambda_+-\gamma)) \cdot \frac{\epsilon\delta(\alpha_{\max}-\alpha_{\min})}{4}\left|e_{i_{\varrho_0-1}}^\top \nabla f(x_{\varrho_0-1})\right| \\
&\overset{(4.25)}{\geq} \exp(S(\lambda_+-\gamma)) \cdot \frac{\epsilon\delta\mu(\alpha_{\max}-\alpha_{\min})}{4}\|\nabla f(x_{\varrho_0-1})\|,
\end{aligned}$$

where the first inequality is because $\Phi^H((i,\alpha)_{\varrho_0:\varrho_0+S-1})\mathcal{P}_+^H((i,\alpha)_{\varrho_0:\varrho_0+S-1})x_{\varrho_0}$ and $\Phi^H((i,\alpha)_{\varrho_0:\varrho_0+S-1})(I-\mathcal{P}_+^H((i,\alpha)_{\varrho_0:\varrho_0+S-1}))x_{\varrho_0}$ are orthogonal. Combining (4.24) and (4.27), we obtain that

$$\begin{aligned}
\|x_{\varrho_0+S}\| &\geq \exp(S(\lambda_+-\gamma)) \cdot \frac{\epsilon\delta\mu(\alpha_{\max}-\alpha_{\min})}{4}\|\nabla f(x_{\varrho_0-1})\| - \|x_{\varrho_0}\| \\
&\overset{(4.21)}{\geq} \left(\exp(S(\lambda_+-\gamma)) \cdot \frac{\epsilon\delta\mu\sigma(r_-)^{L-1}(\alpha_{\max}-\alpha_{\min})}{4} - (r_+)^L\right)\cdot\|x_0\| \\
&\overset{(4.22)}{\geq} \exp(S(\lambda_+-\gamma)) \cdot \frac{\epsilon\delta\mu\sigma(r_-)^{L-1}(\alpha_{\max}-\alpha_{\min})}{8}\cdot\|x_0\|.
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
\|x_{L+S}\| &\overset{(4.21)}{\geq} (r_-)^{L-1}\|x_{\varrho_0+S}\| \\
&\geq \exp(S(\lambda_+-\gamma)) \cdot \frac{\epsilon\delta\mu\sigma(r_-)^{2(L-1)}(\alpha_{\max}-\alpha_{\min})}{8}\cdot\|x_0\|.
\end{aligned}$$

We finally arrive at (4.8) by setting $T = L+S$ and combining the above with (4.23). ∎

4.3. Proof of main results. In this section, we first prove the following theorem, which relies on the local amplification with high probability established in Theorem 4.4. The main result (Theorem 1) will follow as an immediate corollary since $\mathrm{Crit}_s(f)$ is countable and strict saddle points are isolated.

THEOREM 4.9. *Suppose that Assumptions* 3.1, 3.3, *and* 3.4 *hold and that* $0 < \alpha_{\min} < \alpha_{\max} < 1/M$. *Then for every* $x^* \in \mathrm{Crit}_s(f)$ *and every* $x_0 \in \mathbb{R}^d \backslash \{x^*\}$, *it holds that*

$$\mathbb{P}(\Omega(x^*, x_0)) = 0.$$

Proof. Without loss of generality, we assume that $x^* = 0$. Conditioned on $\mathcal{F}_{t-1}$ with $x_t \in V$, where V can be assumed to be bounded, Theorem 4.4 states that with probability at least $1 - 4\epsilon$,

$$\|x_{t+T}\| \geq A\|x_t\|,$$

where to simplify notation we denote

$$A := \exp\left(\frac{6\epsilon}{1-6\epsilon}\big|\log(1 - M\alpha_{\max})\big|\, T\right)$$

as the amplification factor appearing on the right-hand side of (4.8). Notice also that, due to (4.21), we always have

$$\|x_{t+T}\| \geq (r_-)^T \|x_t\|.$$

It suffices to show that for any $x_0 \in V \backslash \{x^*\}$, with probability 1, there exists some $t \in \mathbb{N}_+$ such that $x_t \notin V$.

Let us consider the iterates every T steps: Denote $y_t = x_{Tt}$ and $\mathcal{G}_t = \mathcal{F}_{Tt-1}$ for $t \in \mathbb{N}$. Denote stopping time

$$\rho = \inf\{t \in \mathbb{N} : y_t \notin V\}.$$

It suffices to show that $\mathbb{P}(\rho < \infty) = 1$. We define a sequence of random variables I_t as follows:

$$I_t(\omega) = \begin{cases} 1 & \text{if } \|y_{t+1}\| \geq A\|y_t\|, \\ 0 & \text{otherwise.} \end{cases}$$

By the discussion in the beginning of the proof, we have

$$\mathbb{P}(I_t = 1 \mid \mathcal{G}_t, t < \rho) \geq 1 - 4\epsilon,$$

and moreover, setting $S_t(\omega) = \sum_{0 \leq s < t} I_t(\omega)$, we have, for $t < \rho(\omega)$,

$$\frac{\|y_t\|}{\|y_0\|} \geq A^{S_t(\omega)} \cdot (r_-)^{T(t - S_t(\omega))}.$$

Denote $R := \sup_{x \in V} \|x\| < \infty$. Since $(1 - 5\epsilon)\log A + 5\epsilon T \log r_- > 0$, there exists $t_* \in \mathbb{N}$ such that

$$\left(A^{1-5\epsilon} \cdot (r_-)^{5\epsilon T}\right)^t > \frac{R}{\|y_0\|}, \quad \forall\ t \geq t_*.$$

Therefore, for any $t \geq t_*$, it holds that

$$\mathbb{P}(\rho > t) = \mathbb{P}\big(\rho > t, S_t \leq (1-5\epsilon)t\big) \leq \sum_{i \leq (1-5\epsilon)t} \binom{t}{i} (1-4\epsilon)^i (4\epsilon)^{t-i}.$$

As we will show in the next lemma, the right-hand side of the above goes to 0 as $t \to \infty$, and thus $\lim_{t\to\infty} \mathbb{P}(\rho > t) = 0$, which implies that $\mathbb{P}(\rho < \infty) = 1$. ∎

LEMMA 4.10. *For any $\epsilon \in (0, 1/4)$, it holds that*

$$\lim_{t\to\infty} \sum_{i \leq (1-5\epsilon)t} \binom{t}{i} (1-4\epsilon)^i (4\epsilon)^{t-i} = 0.$$

Proof. Let $X_0, X_1, X_2, \dots$ be a sequence of i.i.d. random variables with X_i being a Bernoulli random variable with expectation $1-4\epsilon$. Denote the average $\bar{X}_t = \frac{1}{t}\sum_{0\le s<t} X_s$. The weak law of large numbers yields that

$$\sum_{i\le(1-5\epsilon)t} \binom{t}{i}(1-4\epsilon)^i(4\epsilon)^{t-i} = \mathbb{P}\left(\bar{X}_t \le 1-5\epsilon\right) \le \mathbb{P}\left(|\bar{X}_t - \mathbb{E}X_0| \ge \epsilon\right) \to 0$$

as $t\to\infty$. ∎

The main theorem then follows directly from Theorem 4.9.

Proof of Theorem 1. Assumption 3.3 guarantees that, in a small neighborhood of x^*, the gradient $\nabla f(x) = \nabla^2 f(x^*)(x-x^*) + o(\|x-x^*\|)$ is nonvanishing as long as $x \ne x^*$, which implies that x^* is an isolated stationary point. Therefore, $\mathrm{Crit}_s(f)$ is countable. Then Theorem 1 follows directly from Theorem 4.9 and the countability of $\mathrm{Crit}_s(f)$. ∎

We now prove the global convergence, i.e., Corollary 3.6, for which we will show that Algorithm 1.1 converges to a critical point of f with some appropriate assumptions. We first show that the limit of each convergent subsequence of $\{x_t\}_{t\in\mathbb{N}}$ is a critical point of f.

PROPOSITION 4.11. *If Assumption* 3.1 *holds and* $0<\alpha_{\min}<\alpha_{\max}<1/M$*, for any* $x_0\in\mathbb{R}^d$ *with bounded level set* $L(x_0)=\{x\in\mathbb{R}^d : f(x)\le f(x_0)\}$*, with probability* 1*, every accumulation point of* $\{x_t\}_{t\in\mathbb{N}}$ *is in* $\mathrm{Crit}(f)$.

Proof. Algorithm 1.1 is always monotone since the following holds for any $t\in\mathbb{N}$ by Taylor's expansion:

$$\begin{aligned} f(x_{t+1}) &= f\left(x_t - \alpha_t e_{i_t} e_{i_t}^\top \nabla f(x_t)\right) \\ &= f(x_t) - \alpha_t \left(e_{i_t}^\top \nabla f(x_t)\right)^2 \\ &\quad + \frac{1}{2}\alpha_t^2 \left(e_{i_t}^\top \nabla f(x_t)\right)^2 \cdot e_{i_t}^\top \nabla f\left(x_t - \theta_t \alpha_t e_{i_t} e_{i_t}^\top \nabla f(x_t)\right) e_{i_t} \\ &\le f(x_t) - \frac{1}{2}\alpha_t (e_i^\top \nabla f(x_t))^2 \\ &\le f(x_t), \end{aligned} \tag{4.28}$$

where $\theta_t\in(0,1)$, which implies that the whole sequence $\{x_t\}_{t\in\mathbb{N}}$ is contained in the bounded level set $L(x_0)$.

Let us consider any $\eta>0$ and set

$$L(x_0,\eta) = \{x\in L(x_0) : \|\nabla f(x)\| \ge \eta\},$$

which is either empty or compact. We claim that with probability 1, the accumulation points of $\{x_t\}_{t\in\mathbb{N}}$ will not be located in $L(x_0,\eta)$. This is clear when $L(x_0,\eta)$ is empty, so it suffices to consider compact $L(x_0,\eta)$. Set $\mu\in(0,1/\sqrt{d}]$ as a fixed constant. For any $x\in L(x_0,\eta)$, there exists an open neighborhood U_x of x and a coordinate $i_x\in\{1,2,\dots,d\}$ such that

$$\left|e_{i_x}^\top \nabla f(y)\right| \ge \mu\|\nabla f(x)\| \ge \mu\eta, \quad \forall\ y\in U_x, \tag{4.29}$$

and that

$$\sup_{y\in U_x} f(y) - \inf_{y\in U_x} f(y) < \frac{\alpha_{\min}\mu^2\eta^2}{2}. \tag{4.30}$$

Noticing that $L(x_0,\eta)\subset\bigcup_{x\in L(x_0,\eta)}U_x$, by the compactness, there exist finitely many points, say, $x^1,x^2,\dots,x^K$, such that

$$L(x_0,\eta)\subset\bigcup_{1\le k\le K}U_{x^k}.$$

For any $k\in\{1,2,\dots,K\}$, combining (4.28), (4.29), and (4.30), we know that for any t, conditioned on $\mathcal{F}_{t-1}$ with $x_t\in U_{x^k}$, if $i_t=i_x$ (which has probability $1/d$), then $f(x_{t+1})<\inf_{y\in U_{x^k}}f(y)$, and thus $x_{t'}\notin U_{x^k}$ for all $t'>t$.

Therefore, the probability that there are infinitely many $t\in\mathbb{N}$ with $x_t\in U_{x^k}$ is zero, which implies that $\{x_t\}_{t\in\mathbb{N}}$ does not have accumulation points in U_{x^k} with probability 1. We conclude that with probability 1, $L(x_0,\eta)$ does not contain any accumulation points of $\{x_t\}_{t\in\mathbb{N}}$, as K is finite. Since this holds for any $\eta>0$, we have for $\mathbb{P}$-a.e. $\omega\in\Omega$ that $\{x_t\}_{t\in\mathbb{N}}$ has no accumulation points in $\bigcup_{n\ge1}L(x_0,1/n)$, which then leads to the desired result. ∎

Proposition 4.11 implies that any accumulation point of the algorithm iterate is a critical point. If we further assume that each critical point of f is isolated, we would conclude that the whole sequence $\{x_t\}_{t\in\mathbb{N}}$ converges and that the limit is in $\mathrm{Crit}(f)$.

PROPOSITION 4.12. *Under the assumptions of Proposition* 4.11. *If every* $x^*\in\mathrm{Crit}(f)$ *is an isolated critical point of* f*, then with probability* 1*,* x_t *converges to some critical point of* f *as* $t\to\infty$.

Proof. It follows from Proposition 4.11 that $\|\nabla f(x_t)\|$ converges to 0 as $t\to\infty$ for a.e.*a.e.* $\omega\in\Omega$. In fact, if there were a subsequence $\{x_{t_k}\}_{k\in\mathbb{N}}$ and $\epsilon>0$ with $\|\nabla f(x_{t_k})\|\ge\epsilon$, $\forall\, k\in\mathbb{N}$, then by the boundedness of $L(x_0)$, $\{x_{t_k}\}_{k\in\mathbb{N}}$ would have some accumulation point which is not a stationary point of f, which leads to a contradiction.

Moreover, $\mathrm{Crit}(f)\cap L(x_0)$ is a finite set since otherwise $\mathrm{Crit}(f)\cap L(x_0)$ would have a limiting point which would be a nonisolated critical point of f, violating the assumption.

Consider a fixed $\omega\in\Omega$ with $\lim_{t\to\infty}\|\nabla f(x_t)\|=0$. Select an open neighborhood U_{x^*} for every $x^*\in\mathrm{Crit}(f)\cap L(x_0)$ such that there exists some $\delta>0$ with

$$\mathrm{dist}(U_{x^*},U_{y^*})=\inf_{x\in U_{x^*},y\in U_{y^*}}\|x-y\|>\delta,\quad\forall\, x^*,y^*\in\mathrm{Crit}(f).$$

If $\{x_t\}_{t\in\mathbb{N}}$ has more than one accumulation point, there would be infinitely many iterates located in $L(x_0)\backslash\bigcup_{x^*\in\mathrm{Crit}(f)\cap L(x_0)}U_{x^*}$, which is compact. Therefore, $\{x_t\}_{t\in\mathbb{N}}$ would have an accumulation point in $L(x_0)\backslash\bigcup_{x^*\in\mathrm{Crit}(f)\cap L(x_0)}U_{x^*}$, which contradicts Proposition 4.11. ∎

Corollary 3.6 is now an immediate consequence.

Proof of Corollary 3.6. The result follows directly from Theorem 1 and Proposition 4.12. ∎

Appendix A. Validity of Assumption 3.4. In this appendix, we provide some justification of Assumption 3.4, which is expected to hold generically. In particular, the following proposition validates this assumption when the off-diagonal entries of H are all nonzero.

PROPOSITION A.1. *Suppose that the largest Lyapunov exponent of* $\Phi^H(t,\omega)$ *is positive. Then Assumption* 3.4 *holds as long as* $1<\alpha_{\min}<\alpha_{\max}<1/\max_{1\le i\le d}|H_{ii}|$ *and every off-diagonal entry of* H *is nonzero.*

Proof. For any element ω in Ω, we take the smallest ℓ such that $\{1,2,\dots,d\} = \{i_0, i_1, \dots, i_{\ell-1}\}$ and write

$$\omega = ((i_0,\alpha_0),\dots,(i_{\ell-1},\alpha_{\ell-1}),\omega'),$$

where $\omega' = \tau^\ell \omega \in \Omega$. We have that ℓ is finite for a.e. $\omega \in \Omega$. Note that we can view $\ell - 1$ as a stopping time, in particular, given ℓ, ω' has distribution $\mathbb{P}$ and is independent with $\mathcal{F}_{\ell-1}$.

Let $\{v_1', v_2', \dots, v_m'\}$ be a set of basis vectors for $W_-^H(\omega') = W_-^H(\tau^\ell\omega)$. Then a set of basis vectors for $W_-^H(\omega)$ is given by

$$v_j = \left(I - \alpha_0 e_{i_0} e_{i_0}^\top H\right)^{-1} \cdots \left(I - \alpha_{\ell-1} e_{i_{\ell-1}} e_{i_{\ell-1}}^\top H\right)^{-1} v_j', \quad j = 1,2,\dots,m.$$

Denote the matrices concatenated by the column vectors as $V' = (v_1'|v_2'|\cdots|v_m')$ and $V = (v_1|v_2|\cdots|v_m)$. If $e_i \in W_-^H(\omega) = \operatorname{span}\{v_1, v_2, \dots, v_m\}$, then $V_{\hat{i},:}$ is column-rank deficient since the existence of a positive Lyapunov exponent implies that $m \le d-1$. Here and for the rest of the appendix, we denote by $V_{\hat{i},:}$ the $(d-1)\times m$ matrix obtained via removing ith row of $V \in \mathbb{R}^{d\times m}$.

Therefore, as Assumption 3.4 is equivalent to that $e_i \notin W_-^H(\omega)$ holds for any $i \in \{1,2,\dots,d\}$ and almost every $\omega \in \Omega$, it suffices to show that $V_{\hat{i},:}$ has full column rank with probability 1. The key point is that given ℓ, $\alpha_0, \alpha_1, \dots, \alpha_{\ell-1}$ are independent with $i_0, i_1, \dots, i_{\ell-1}$ and $\omega' = \tau^\ell\omega$. Thus, it suffices to show that with fixed ℓ, $i_0, i_1, \dots, i_{\ell-1}$, $\omega' = \tau^\ell\omega$, and $v_1', v_2', \dots, v_m'$, the set of all $\alpha_0, \alpha_1, \dots, \alpha_{\ell-1}$ that yield the rank deficiency of $V_{\hat{i},:}$ is of measure zero, and without loss of generality, we can assume $i = 1$. Noticing that $i_0, i_1, \cdots, i_{\ell-1}$ cover all the coordinates and that every off-diagonal entry of H is nonzero, the desired result follows directly from the following Lemma A.2 applied repeatedly. ∎

LEMMA A.2. *Suppose that $X = (X_1|X_2|\cdots|X_d)^\top$ and $Y = (Y_1|Y_2|\cdots|Y_d)^\top$ are full-column-rank $d\times m$ matrices satisfying $Y = (I - \alpha e_k e_k^\top H)^{-1} X$ (we suppress in the notation the dependence of Y on k and α for simplicity). Then the following holds:*

(i) *If $X_{\hat{1},:}$ has full column rank, then for any $k = \{1,2,\dots,d\}$, $Y_{\hat{1},:}$ also has full column rank for a.e. α.*

(ii) *Suppose that $X_{\hat{1},:}$ is column-rank deficient, and let $2 \le j_1 < j_2 < \ldots < j_{m-1} \le d$ be row indices such that*

$$X_j \in \operatorname{span}\{X_{j_1}, X_{j_2}, \dots, X_{j_{m-1}}\}, \quad \forall\ j \in \{2,3,\dots,d\}.$$

If $k \in \{1, j_1, j_2, \dots, j_{m-1}\}$, then we have with probability 1 that either $Y_{\hat{1},:}$ has full column rank or $Y_{\hat{1},:}$ is column-rank deficient with

$$Y_j \in \operatorname{span}\{Y_{j_1}, Y_{j_2}, \dots, Y_{j_{m-1}}\}, \quad \forall\ j \in \{2,3,\dots,d\}.$$

If $k \notin \{1, j_1, j_2, \dots, j_{m-1}\}$ and $H_{k1} \neq 0$, then $Y_{\hat{1},:}$ has full column rank.

Proof of Lemma A.2. By (3.3), it holds that $Y_j = X_j$ for $j \neq k$ and that

$$Y_k = \frac{1}{1 - \alpha H_{kk}} \left(X_k + \alpha \sum_{j\neq k} H_{kj} X_j \right).$$

For point (i), we notice that if $k = 1$, then $Y_{\hat{1},:} = X_{\hat{1},:}$ has full column rank. If $k > 1$, then it follows from $X_1 \in \operatorname{span}\{X_2, \dots, X_d\}$ that $Y_{\hat{1},:}$ also has full column rank for a.e. α.

For point (ii), we have

$$\operatorname{span}\{X_1, X_{j_1}, X_{j_2}, \ldots, X_{j_{m-1}}\} = \mathbb{R}^m.$$

If $k \in \{1, j_1, j_2, \ldots, j_{m-1}\}$, then $\operatorname{span}\{Y_1, Y_{j_1}, Y_{j_2}, \ldots, Y_{j_{m-1}}\} = \mathbb{R}^m$ holds for a.e. α. Therefore, we obtain that $Y_{j_1}, Y_{j_2}, \ldots, Y_{j_{m-1}}$ are linearly independent, which implies that either $Y_{\hat{1},:}$ has full column rank or

$$Y_j \in \operatorname{span}\{Y_{j_1}, Y_{j_2}, \ldots, Y_{j_{m-1}}\}, \quad \forall\, j \in \{2, 3, \ldots, n\}.$$

If $k \notin \{j_1, j_2, \ldots, j_{m-1}\}$, then $Y_{\hat{1},:}$ has full column rank since $H_{k1} \neq 0$. ∎

Acknowledgments. We thank Jonathan Mattingly, Zhe Wang, and Stephen J. Wright for helpful discussions.

REFERENCES

[1] L. Arnold, *Random Dynamical Systems,* Springer Monogr. Math., Springer, New York, 1998.
[2] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.
[3] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., 137 (2013), pp. 91–129.
[4] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060.
[5] R. Bhatia, *Matrix Analysis*, Vol. 169, Springer Science & Business Media, New York, 2013.
[6] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494.
[7] R. I. Boţ, M. N. Dao, and G. Li, *Inertial Proximal Block Coordinate Method for a Class of Nonsmooth and Nonconvex Sum-of-Ratios Optimization Problems*, preprint, arXiv:2011.09782, 2020.
[8] P. Boxler, *A stochastic version of center manifold theory*, Probab. Theory Related Fields, 83 (1989), pp. 509–545.
[9] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, *Gradient descent can take exponential time to escape saddle points*, in Advances in Neural Information Processing Systems, 2017, pp. 1067–1077.
[10] G. Froyland, C. González-Tokman, and A. Quas, *Stochastic stability of Lyapunov exponents and Oseledets splittings for semi-invertible matrix cocycles*, Comm. Pure Appl. Math., 68 (2015), pp. 2052–2081.
[11] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points—Online stochastic gradient for tensor decomposition*, in Conference on Learning Theory, 2015, pp. 797–842.
[12] R. Ge, C. Jin, and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, in International Conference on Machine Learning, PMLR, 2017, pp. 1233–1242.
[13] P. Guo and J. Shen, *Smooth center manifolds for random dynamical systems*, Rocky Mountain J. Math., 46 (2016), pp. 1925–1962.
[14] X. Guo, J. Han, and W. Tang, *Perturbed Gradient Descent with Occupation Time*, preprint, arXiv:2005.04507, 2020.
[15] M. Gürbüzbalaban, A. Ozdaglar, N. D. Vanli, and S. J. Wright, *Randomness and permutations in coordinate descent methods*, Math. Program., 181 (2020), pp. 349–376.
[16] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, in International Conference on Machine Learning, 2017, pp. 1724–1732.
[17] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, *On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points*, J. ACM, 68 (2021), pp. 1–29.
[18] C. Jin, P. Netrapalli, and M. I. Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, in Conference on Learning Theory, PMLR, 2018, pp. 1042–1085.

[19] K. Kawaguchi, *Deep learning without poor local minima*, in Advances in Neural Information Processing Systems, 2016.
[20] F. Ledrappier and L.-S. Young, *Stability of Lyapunov exponents*, Ergodic Theory Dynam. Systems, 11 (1991), pp. 469–484.
[21] C.-P. Lee and S. J. Wright, *Random permutations fix a worst case for cyclic coordinate descent*, IMA J. Numer. Anal., 39 (2019), pp. 1246–1275.
[22] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order methods almost always avoid strict saddle points*, Math. Program., 176 (2019), pp. 311–337.
[23] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, *Gradient descent only converges to minimizers*, in Conference on Learning Theory, 2016, pp. 1246–1257.
[24] K. Y. Levy, *The Power of Normalization: Faster Evasion of Saddle Points*, preprint, arXiv:1611.04831, 2016.
[25] J. Li, K. Lu, and P. Bates, *Normally hyperbolic invariant manifolds for random dynamical systems: Part* I—*Persistence*, Trans. Amer. Math. Soc., (2013), pp. 5933–5966.
[26] W. Li and K. Lu, *Poincaré theorems for random dynamical systems*, Ergodic Theory Dynam. Systems, 25 (2005), 1221.
[27] W. Li and K. Lu, *Sternberg theorems for random dynamical systems*, Comm. Pure Appl. Math., 58 (2005), pp. 941–988.
[28] W. Li and K. Lu, *A siegel theorem for dynamical systems under random perturbations*, Discrete Contin. Dyn. Syst. Ser. B, 9 (2008), p. 635.
[29] W. Li and K. Lu, *Takens theorem for random dynamical systems*, Discrete Contin. Dyn. Syst. Ser. B, 21 (2016), p. 3191.
[30] Y. Li, J. Lu, and Z. Wang, *Coordinatewise descent methods for leading eigenvalue problem*, SIAM J. Sci. Comput., 41 (2019), pp. A2681–A2716.
[31] Z. Lian and K. Lu, *Lyapunov exponents and invariant manifolds for random dynamical systems in a Banach space*, Mem. Amer. Math. Soc., 206 (2010).
[32] J. Liu and S. J. Wright, *Asynchronous stochastic coordinate descent: Parallelism and convergence properties*, SIAM J. Optim., 25 (2015), pp. 351–376.
[33] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar, *An asynchronous parallel stochastic coordinate descent algorithm*, in International Conference on Machine Learning, 2014, pp. 469–477.
[34] P.-D. Liu and M. Qian, *Smooth Ergodic Theory of Random Dynamical Systems*, Springer, New York, 2006.
[35] H. Lu and K. Kawaguchi, *Depth Creates No Bad Local Minima*, preprint, arXiv:1702.08580, 2017.
[36] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362.
[37] V. I. Oseledets, *A multiplicative ergodic theorem. Liapunov characteristic numbers for dynamical systems*, Trans. Moscow Math. Soc., 19 (1968), pp. 197–231.
[38] P. Oswald and W. Zhou, *Random reordering in SOR-type methods*, Numer. Math., 135 (2017), pp. 1207–1220.
[39] M. O'Neill and S. J. Wright, *Behavior of accelerated gradient methods near critical points of nonconvex functions*, Math. Program., 176 (2019), pp. 403–427.
[40] M. S. Raghunathan, *A proof of Oseledec's multiplicative ergodic theorem*, Israel J. Math., 32 (1979), pp. 356–362.
[41] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Math. Program., 144 (2014), pp. 1–38.
[42] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program., 156 (2016), pp. 433–484.
[43] D. Ruelle, *Ergodic theory of differentiable dynamical systems*, Publications mathématiques, Institut des hautes études scientifiques, 50 (1979), pp. 27–58.
[44] D. Ruelle, *Characteristic exponents and invariant manifolds in Hilbert space*, Ann. Math., (1982), pp. 243–290.
[45] A. Saha and A. Tewari, *On the nonasymptotic convergence of cyclic coordinate descent methods*, SIAM J. Optimiz., 23 (2013), pp. 576–601.
[46] H.-J. M. Shi, S. Tu, Y. Xu, and W. Yin, *A Primer on Coordinate Descent Algorithms*, preprint, arXiv:1610.00040, 2016.
[47] M. Shub, *Global Stability of Dynamical Systems*, Springer Science & Business Media, New York, 1987.
[48] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, Found. Comput. Math., 18 (2018), pp. 1131–1198.

[49] R. Sun and M. Hong, *Improved iteration complexity bounds of cyclic block coordinate descent for convex problems*, Adv. Neural Inform. Process. Syst., 28 (2015), pp. 1306–1314.
[50] R. Sun and Y. Ye, *Worst-case complexity of cyclic coordinate descent: $o(n^2)$ gap with randomized version*, Math. Program., (2019), pp. 1–34.
[51] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.
[52] P. Walters, *A dynamical proof of the multiplicative ergodic theorem*, Trans. Amer. Math. Soc., 335 (1993), pp. 245–257.
[53] T. Wanner, *Linearization of random dynamical systems*, in Dynamics Reported, Springer, New York, 1995, pp. 203–268.
[54] S. J. Wright, *Accelerated block-coordinate relaxation for regularized optimization*, SIAM J. Optim., 22 (2012), pp. 159–186.
[55] S. J. Wright, *Coordinate descent algorithms*, Math. Program., 151 (2015), pp. 3–34.
[56] S. J. Wright and C.-P. Lee, *Analyzing random permutations for cyclic coordinate descent*, Math. Comp., 89 (2020), pp. 2217–2248.