CONVERGENCE ANALYSIS OF STOCHASTIC GRADIENT DESCENT

ABSTRACT. In this note, we consider the asymptotic behavior of stochastic gradient descent under Robbins-Monro type decreasing learning rates. We review the results of Benaïm and Hursch: under standard growth conditions on the optimization objective and moment conditions on the gradient oracle, almost every trajectory of stochastic gradient descent converges to a critical point. Moreover, using dynamical systems theory, we sketch an argument by Mertikopoulos, Hallak, Kavis, and Cevher that shows that strict saddle points are avoided.

CONTENTS

Let $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be a collection of labelled data. Our goal is to find, from some parameterized class of functions $\{f_\theta : \mathbb{R}^d \to \mathbb{R} : \theta \in \Theta\}$, that minimizes the *empirical risk*, i.e., for some loss function $\ell : \mathbb{R}^2 \to \mathbb{R}_+$,

$$\min_{\theta\in\Theta} \Lambda(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell(f_\theta(x_i), y_i).$$

Throughout, we will use the canonical choice of

$$\ell(y_1, y_2) = |y_1 - y_2|^2.$$

The empirical risk minimizer does not usually admit a closed form, and we rely on algorithms that use queries of the function value or its derivatives (or more) to find the minimizer. For example, one can perform gradient descent, an iteration of the form

$$\theta_{k+1} = \theta_k - \eta_k \nabla\Lambda(\theta_k). \tag{GD}$$ {eq:gd}

Under mild smoothness conditions of Λ, we know that the iteration converges asymptotically to a critical point, i.e., $\lim_{k\to\infty} \|\nabla\Lambda(\theta_k)\|_2 = 0$.

Taking a closer look, however, reveals a practical problem: the number of gradient computations scales linearly in n! Though harmless in complexity land, it is causing problems on real computers. The resolution is simple: at each iteration k, we draw a random size-m subset D_k of $\{1, \dots, n\}$ independently of the past, and compute

$$G_k = \frac{1}{m}\sum_{i\in D_k} \nabla_\theta \left(|f_\theta(x_i) - y_i|^2\right).$$

That is, we perform the iteration

$$\theta_{k+1} = \theta_k - \eta_k G_k.$$

This works tremendously well in terms of both computational complexity and performance. We would like to know why...

0.1. **Standing assumptions.** To unify discussion, we provide an abstract framework for SGD and place some standing assumptions throughout. Suppose that parameter space $\Theta = \mathbb{R}^d$ is Euclidean, we will let

$$\Lambda : \mathbb{R}^d \to \mathbb{R}$$

be the objective function to minimize. At each step, the algorithm evolves according to some (random) gradient oracle V and the dynamics read

{eq:sgd}
$$\theta_{k+1} = \theta_k + \eta_k \mathsf{G}(\theta_k, \xi_k) \tag{SGD}$$

where $(\xi_k)_{k\in\mathbb{N}}$ are independent, identically-distributed noises. We place no assumptions on the initial condition θ_0. We are interested in the pathwise, long-time behavior of θ_k, that is,

Does there exist a set $\mathcal{C} \subset \mathbb{R}^d$ such that $\lim_{k\to\infty} d(\theta_k, \mathcal{C}) = 0$ with probability one?

For this, we propose the following set of assumptions.

std-ass-obj}
Assumption 0.1. We assume the following about the objective function Λ:

(1) $\nabla\Lambda$ is bounded by c_0 and c_1-Lipschitz;
(2) Λ and its gradient have bounded (sub)-level sets, i.e., for any $c \in \mathbb{R}$, the sets

$$\{\theta \in \mathbb{R}^d : \Lambda(\theta) \le c\} \quad \text{and} \quad \{\theta \in \mathbb{R}^d : \|\nabla\Lambda(\theta)\| \le c\}$$

are bounded;

td-ass-grad}
Assumption 0.2. We further assume the following about the gradient oracle:

(1) the gradient oracle is unbiased, i.e.,

$$\mathbb{E}\left[\mathsf{G}(\theta_k, \xi_k)\right] = \nabla\Lambda(\theta_k);$$

(2) the gradient oracle has finite second moment, that is, there exists $\sigma > 0$ such that

$$\sup_{\theta\in\mathbb{R}^d, k\in\mathbb{N}} \mathbb{E}\left[\|\mathsf{G}(\theta, \xi_k)\|^2\right] < \sigma.$$

Lastly, we assume the following on the learning rate:

(1) the learning rate satisfies

$$\sum_{k\in\mathbb{N}} \eta_k = +\infty \quad \text{and} \quad \sum_{k\in\mathbb{N}} \eta_k^2 < \infty.$$

Remark 0.3. The formulation of (SGD), particularly with the introduction of V, is written in a way that invites generalization to the forthcoming theory in cases where, e.g., the noise is non-additive. We will remark these generalizations when possible.

1. Is (SGD) a gradient flow?

Taking θ_k to the left-hand side of (SGD) and taking η_k small, it is tempting to consider the gradient flow instead, i.e., the solution to

{eq:gf}
$$\dot{\theta}_t = -\nabla\Lambda(\theta_t). \tag{GF}$$

We will show that this comparison is indeed valid for the step-size regime considered. In fact, this theory has been known for some time as part of *stochastic approximation*, in particular, as the *method of ordinary differential equations*; see [5] for a thorough overview.

Remark 1.1. The method of ordinary differential equations, initiated largely by Kushner and Clark [6], has led to a slightly different set of techniques than the ones presented below. The approach is more probabilistic, namely, it uses the theory of weak convergence, which allows for more detailed analysis of typical fluctuations and rare events [5, Section 8]. It also allows for constant step sizes, as demanded sometimes from the application, whereas the pathwise approach, soon to be introduced, requires diminishing step sizes.

1.1. **Asymptotic pseudotrajectories and their limit sets.** We begin with a formal notion of trajectories that are asymptotically alike. This will serve as the basis of the comparison between (SGD) and (GF).

{eq:def-apt}

Definition 1.2 ([3, Page 141])**.** Let $\Phi : \mathbb{R}_+ \times \mathbb{R}^d \ni (t, x) \mapsto \Phi_t(x) \in \mathbb{R}^d$ be a semiflow, i.e.,

$$\Phi_0 = \mathrm{id} \quad \text{and} \quad \Phi_s \circ \Phi_t = \Phi_{s+t} \text{ for all } 0 \leq s, t.$$

Then, we say that a continuous function $\theta : \mathbb{R}_+ \to \mathbb{R}^d$ is an *asymptotic pseudotrajectory* of the semiflow Φ if

$$\lim_{t\to\infty} \sup_{h\in[0,t]} \|\Phi_s(\theta_h) - \theta_{s+h}\| = 0 \quad \text{for all } s > 0.$$

There are many asymptotic pseudotrajectories, random and deterministic, that naturally occur in applications; see [3, Section 2-6] for some. In particular, it can be shown generally that the asymptotic behavior of asymptotic pseudotrajectories is similar to that of the corresponding semiflow [1, 2]. We focus on the particular case when Φ admits a Lyapunov function. Recall that,

Definition 1.3. Let $\mathcal{C} \subset \mathbb{R}^d$ be a compact invariant set of the semiflow Φ. A *Lyapunov function* $\mathsf{V} : \mathbb{R}^d \to \mathbb{R}$ for $\mathcal{C}$ is such that the map $s \mapsto \mathsf{V}(\Phi_s(x))$ is constant for all $x \in \mathcal{C}$ and strictly decreasing for $x \notin \mathbb{R}^d$. The Lyapunov function is *strict* is $\mathcal{C}$ coincides with the equilibria set, that is,

$$\mathcal{C} = \{x \in \mathbb{R}^d : \Phi_s(x) = x \text{ for all } s \geq 0\}.$$

{prop:apt-to

Proposition 1.4 ([2, Corollary 6.6])**.** *Let Φ be the semiflow generated by (GF) with Λ satisfying Assumption 0.1. Then, if θ is an precompact asymptotic pseudotrajectory of Φ, its limit set contains the equilibria set of Φ:*

$$\mathsf{L}(\theta) := \bigcap_{s\geq 0} \overline{\theta([s,\infty))} \subset \{x \in \mathbb{R}^d : \nabla\Lambda(x) = 0\}.$$

First, we notice that (GF) admits a unique global solution. Moreover, it is clear that $\mathsf{V} = \Lambda$ is a (strict) Lyapunov function for the set of critical points. So, the question boils down to whether the set of limit points of θ coincides with the set of critical points of Λ. For this, we invoke a more general result, which involves a few definitions.

Definition 1.5. A (δ, t)-pseudo orbit from $a \in \mathcal{C}$ to $b \in \mathcal{C}$ is a finite collection of trajectories

$$\{\Phi_s(x_i) : 0 \leq s \leq t_i\} \text{ for } i = 0, \dots, k-1, \text{ and } t_i \geq t,$$

such that

$$\|x_0 - a\| < \delta, \; x_k = b, \text{ and } \|\Phi_{t_i}(x_i) - x_{i+1}\| < \delta \text{ for } i = 0, \dots, k-1.$$

Then, we say a set $\mathcal{L}$ is *internally chain transitive* if, for every $\delta > 0$ and $t > 0$, there is a (δ, t)-pseudo orbit connecting any two points in $\mathcal{L}$.

In the case of gradient flows (more generically, if Φ admits a Lyapunov function), we know that internally chain transitive sets are connected sets of critical points $\mathcal{C}$. This is because trajectories generated by Φ can only flow down the energy V, which means that pseudo orbits can never be constructed from a point of lower energy to higher energy given δ sufficiently small and t sufficiently large. On the other hand, in the case of gradient flow, invariant sets are equilibria and pseudo orbits

are simply chains of overlapping δ-balls, hence, internally chain transitive sets are connected critical points. We provide a more precise statement below.

Proposition 1.6 ([2, Proposition 6.4]). *Let $\mathcal{C} \subset \mathbb{R}^d$ be a compact invariant set and $\mathsf{V} : \mathbb{R}^d \to \mathbb{R}$ be a Lyapunov function for $\mathcal{C}$. Assume that $\mathsf{V}(\mathcal{C})$ has empty interior. Then, every internal chain transitive set $\mathcal{L}$ is contained in $\mathcal{C}$ and $\mathsf{V}|_{\mathcal{L}}$ is constant.*

Under Assumption 0.1, we can apply Sard's theorem [13] to conclude that $\mathsf{V}(\mathcal{C})$ has zero Lebesgue measure, hence, an empty interior. Next, we state the main theorem that relates limit sets of asymptotic pseudotrajectories with those of the corresponding flow.

Theorem 1.7 ([3, Theorem 8.2]). *Let θ be a precompact asymptotic pseudotrajectory of Φ. Then $\mathsf{L}(\theta)$ is internally chain transitive.*

Since all internally chain transitive sets are connected sets of critical points and the limit set of θ is internally chain transitive, we know that the limit set of θ must be contained in *one of* the set of connected critical points.

1.2. (SGD) **almost always goes to a critical point.** All that is needed from here is to show that (SGD) tracks (GF) in the sense of it being a (precompact) asymptotic psuedotrajectory. First, we can embed realizations of (SGD) via a linear interpolation: let $\tau_n = \sum_{k=1}^{n} \eta_k$, then

$$\theta_t = \theta_n + \frac{t - \tau_n}{\tau_{n+1} - \tau_n}(\theta_{n+1} - \theta_n) \quad \text{for } t \in [\tau_n, \tau_{n+1}).$$

From here, the proof relies on some elaborate estimates, which we some will be omitted for entertainment purposes. We will begin with a sufficient condition on a linearly-interpolated trajectories being an asymptotic pseudotrajectory. The proof of this sufficient condition uses fairly standard analytic arguments, and will be omitted.

Lemma 1.8 ([2, Proposition 4.1 and 4.2]). *Suppose that Λ satisfies Assumption 0.1 and the gradient oracle satisfies Assumption 0.2. Moreover, let $m(t) = \sup\{k \geq 0 : t \geq \tau_k\}$ and*

$$\Delta(n,t) = \sup\left\{ \left\| \sum_{i=n+1}^{k} \eta_i \left(\mathsf{G}(\theta_i, \xi_i) - \mathbb{E}\left[\mathsf{G}(\theta_i, \xi_i) | \theta_{i-1} \right] \right) \right\| : k = n+1, \dots, m(\tau_n + t) \right\}.$$

If for all $t > 0$,

$$\lim_{n \to \infty} \Delta(n,t) = 0 \quad \textit{almost surely}, \tag{1}$$

then θ is an asymptotic pseudotrajectory of (GF).

To show that (1) holds, we need to use several technical yet fairly standard techniques in probability. The first is the observation that "if there is a martingale, there is a way." Owing to the unbiased assumption, we know that the noise has zero mean and

$$M_n = \sum_{k=1}^{n} \left(\mathsf{G}(\theta_k, \xi_k) - \mathbb{E}\left[\mathsf{G}(\theta_k, \xi_k) | \theta_{k-1} \right] \right)$$

is a martingale. Since it is of this differene form, it is also called the *martingale difference sequence* and has much received much love from the stochastic approximation literature.

Looking at (1), we would like to know the something about the excursion of a martingale. The remarkable (and now, standard) result of Burkholder, Davis, and Gundy relates the supremum deviation with the fluctuation of the martingale.

Lemma 1.9 (Burkholder-Davis-Gundy). *There exists a universal constant $c > 0$ such that*

$$\mathbb{E}\left[\sup_{n < k \leq m} \|M_k\|^2 \right] \leq c \sum_{n < k \leq m} \gamma_k^2 \, \mathbb{E}\left[\|\mathsf{G}(\theta_k) - \mathbb{E}\left[\mathsf{G}(s, \theta_k) | \theta_{k-1} \right]\|^2 \right].$$

Remark 1.10. In fact, the Burkholder-Davis-Gundy inequality holds more generally: there exists $c_1, c_2 > 0$ such that

$$c_1 \mathbb{E}\left[[M]_t^{p/2}\right] \leq \mathbb{E}\left[\sup_{0\leq s\leq t} |M_s|^p\right] \leq c_2 \mathbb{E}\left[[M]_t^{p/2}\right]$$

for $p \in [1, \infty)$, where $[M]$ is the *quadratic variation* of the martingale.

For any $\epsilon > 0$, we have

$$\sum_{k=1}^{\infty} \mathbb{P}(\Delta(kt, t) > \epsilon) \lesssim \frac{1}{\epsilon} \sum_{k=1}^{\infty} \gamma_k^2 < \infty.$$

By Borel-Cantelli, we know that

$$\lim_{k\to\infty} \Delta(kt, t) = 0 \quad \text{almost surely.}$$

Lastly, using the fact that for $kt \leq s \leq (k+1)t$,

$$\Delta(s, t) \leq \Delta(kt, t) + \Delta((k+1)t, t).$$

This concludes the proof for why (1) holds.

Showing precompactness, i.e., $\sup_{t\geq 0} \|\theta_t\| < \infty$, is much more complicated; see [9, Theorem 1]. Combining these two results, we know that (SGD) asymptotically follows (GF) in the sense that it goes to internally chain transient points, which are critical points.

2. Which critical points do (SGD) avoid?

Recall that first-order optimality is only sufficient when the objective function Λ is convex, which we do not (and should not) assume. To facilitate discussion, let's give ourselves one more derivative: {ass:obj-hes

Assumption 2.1. The objective function Λ is twice-continuously differentiable everywhere with

$$\sup_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\| = L < \infty$$

while satisfying Assumption 0.1.

Now, the set of local minimizers can be easily defined via the spectrum of the Hessian. Let $\mathcal{C}^*$ be the set of local minimizers, i.e.,

$$\mathcal{C}^* = \{x \in \mathbb{R}^d : \nabla\Lambda(x) = 0 \text{ and } \lambda_{\min}(\nabla^2\Lambda(x)) \geq 0\}.$$

For convenience, also define the set of strict saddle points

$$\mathcal{S} = \{x \in \mathbb{R}^d : \lambda_{\min}(\nabla^2\Lambda(x)) < 0\}.$$

We will show that, in different ways, the (stochastic) gradient descent avoids saddle points $\mathcal{S}$. The protagonist here is the *center-stable manifold theorem*; we will review the appropriate versions as they come up.

2.1. (GD) avoids strict saddle points.

For demonstration purposes, we consider (GD) with constant step size $\eta_k = \eta$. Consider the following example.

Example 2.2 ([10, Section 1.2.3])**.** Consider the "degenerate double-well"

$$f(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$$

be the objective function, and there are three critical points:

$$x_1^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; x_2^* = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \; x_3^* = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The point x_1^* is a saddle point, and others are local minima. Gradient descent starting from any point with second coordinate being zero converges to x_1^* since the update reads:

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} (1-\eta)x_k \\ (1-\eta)y_k - \eta y_k^3 \end{pmatrix}.$$

All other points converge to local minimas. Therefore, the set of initialization that converge to saddles has Lebesgue measure zero. Coincidentally, computing the Hessian

$$\nabla^2 f(x,y) = \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix}$$

and observe that, at x_1^*, the positive eigenvector spans the x-axis. This agrees with the characterization above.

The proof is surprisingly short, given that the center manifold theorem is not unfamiliar.

Theorem 2.3 (The center manifold theorem)**.** *Let x^* be a fixed point of a $\mathcal{C}^r$-local diffeomorphism $g : \mathbb{R}^d \to \mathbb{R}^d$. Let E_{cs} be the center-stable subspace, i.e., spanned by eigenvectors of $\nabla g(x^*)$ with eigenvalue no more than one. Then, there is a $\mathcal{C}^r$-embedded manifold W_{cs} tangent to E_{cs} at x^* and a neighborhood $B \ni x^*$ such that*

$$g(W_{\mathsf{cs}}) \cap B \subset W_{\mathsf{cs}} \quad \textit{and} \quad \bigcap_{k=0}^{\infty} g^{-k}(B) \subset W_{\mathsf{cs}}.$$

Theorem 2.4 ([7, Theorem 2])**.** *Let Λ satisfy Assumption 2.1 and $\eta \leq 1/L$. Then*

$$\left| \left\{ \theta_0 \in \mathbb{R}^d : \lim_{k\to\infty} \theta_k \in \mathcal{S} \right\} \right| = 0.$$

Proof. By [7, Proposition 2], with $\eta \leq 1/L$, the iteration $g = \mathbb{I}_d - \eta\nabla\Lambda$ is a $\mathcal{C}^1$-diffeomorphism with $\mathcal{S}$ being the unstable fixed points.

For each $x^* \in \mathcal{S}$, let B_{x^*} be the neighborhood offered by the center manifold theorem. Then we form an open cover $\mathcal{S}$ by a countable number of B_{x^*}, i.e.,

$$\mathcal{S} \subset \bigcup_{i=1} B_{x_i^*}.$$

Now, take a point θ_0 that converges to $\mathcal{S}$. Then, there is a $m \in \mathbb{N}$ such that $\theta_k \in B_{x_i^*}$ for some $i \in \mathbb{N}$ for all $k \geq m$. So

$$g^{m+k}(\theta_0) \in B_{x_i^*} \quad \text{for all } k \geq 0,$$

in other words,

$$g^m(\theta_0) \in \bigcap_{k=0}^{\infty} g^k(B_{x_i^*}) =: S_i \subset W_{\mathsf{cs}}(x_i^*).$$

But the center-stable manifold has codimension at least one, which has Lebesgue measure zero. Also, since m depends on θ_0, we can agnostically (and generously) write that

$$\theta_0 \in \cup_{j=1}^{\infty} g^{-j}(S_i).$$

Finally, since θ_0 was chosen arbitrarily, we can union over choices of i to get that

$$\left\{ \theta_0 \in \mathbb{R}^d : \lim_{k\to\infty} \theta_k \in \mathcal{S} \right\} \subset \bigcup_{i=1}^{\infty} \bigcup_{j=1}^{\infty} g^j(S_i),$$

which has Lebesgue measure zero. □

Remark 2.5. The assumption on the potential and step size can be relaxed following similar technology [11]. Also, while asymptotically, gradient descent does not stay in saddle points, it can take an exponential amount of time to escape [4].

2.2. (SGD) **avoids hyperbolic saddle points.** The plot thickens dramatically in the stochastic case. Even though we know that the trajectories of (SGD) is close to that of the gradient flow, the theorem given by the theory of asymptotic pseudotrajectories is actually much weaker. The comparison with (GF) is then, unfortunately, quite useless. We leave the following as an exercise to readers who are familiar with the center manifold theorem.

Theorem 2.6. *If the Hessian of the potential is uniformly hyperbolic, that is,*

$$\inf_{x\in\mathbb{R}^d} \lambda_{\min}(\nabla^2\Lambda(x)) > 0,$$

then

$$\left|\left\{\theta_0 \in \mathbb{R}^d : \lim_{t\to\infty} \theta_t \in \mathcal{S}\right\}\right| = 0.$$

The analog of the above theorem—particularly, the hyperbolic assumption—was used to first show that (SGD) avoids saddle points [12]. In particular, it required an additional non-degeneracy condition:

Assumption 2.7. For every $\theta \in \mathbb{R}^d$ such that $\|\theta\| = 1$, we have

$$\inf_{x\in\mathbb{R}^d} \mathbb{E}\left[\langle \mathsf{G}(x,\xi_k),\theta\rangle_+\right] > 0$$

Without this condition, one can choose adversarially so that (SGD) moves deterministically into a saddle point; though in light of the (GD) result, this might not occur often.[1]

We present the idea of the proof, in line with [2, Section 9] and [9, Theorem 3]. The goal is to show that, with probability one, (SGD) exists the stable-center manifold around a saddle point. To do so, one construct an energy functional

$$E(x) = \int_0^\tau \|\Phi_{-s}(x) - \Pi(\Phi_{-s}(x))\|\,\mathrm{d}s,$$

where Π is a (carefully defined) projection onto the center-stable manifold. Then, one tracks the increment of the energy by defining

$$Y_{k+1} = \begin{cases} E(\theta_{k+1}) - E(\theta_k), & \text{if } \theta \text{ haven't escaped the locally-invariant neighborhood;} \\ \eta_k, & \text{otherwise.} \end{cases}$$

Defining $E_n = \sum_{k=1}^n Y_k$, one can use the geometric construction to show that

$$\mathbb{E}\left[E_{n+1}^2 - E_n^2 | \theta_n\right] \gtrsim \frac{1}{n^{2p}}$$

for any $p \in (0,1)$, and consequently,

$$\mathbb{P}\left(\lim_{n\to\infty} E_n = 0\right) = 0.$$

See also [8] for more general conditions.

[1] I believe this is open: showing (SGD) avoids saddle more directly by comparison with (GF).

References

[1] M. Benaim. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, 1996.

[2] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1998.

[3] M. Benaïm and M. W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

[4] 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. *Advances in neural information processing systems*, 30, 2017.

[5] H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, 2006.

[6] H. J. Kushner and D. S. Clark. Convergence wp 1 for unconstrained systems. In *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, pages 19–99. Springer, 1978.

[7] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

[8] J. Liu and Y. Yuan. Almost sure convergence rates analysis and saddle avoidance of stochastic gradient methods. *Journal of Machine Learning Research*, 25(271):1–40, 2024.

[9] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

[10] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[11] I. Panageas and G. Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

[12] R. Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, pages 698–712, 1990.

[13] A. Sard. The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society*, 48(12):883–890, 1942.