Black-Box Generalization: Stability of Zeroth-Order Learning

Konstantinos E. Nikolakakis
Yale University
konstantinos.nikolakakis@yale.edu

Farzin Haddadpour
Yale University
farzin.haddadpour@yale.edu

Dionysios S. Kalogerias
Yale University
dionysis.kalogerias@yale.edu

Amin Karbasi
Yale University
amin.karbasi@yale.edu

Abstract

We provide the first generalization error analysis for black-box learning through derivative-free optimization. Under the assumption of a Lipschitz and smooth unknown loss, we consider the *Zeroth-order Stochastic Search* (ZoSS) algorithm, that updates a d-dimensional model by replacing stochastic gradient directions with stochastic differences of $K+1$ perturbed loss evaluations per dataset (example) query. For both unbounded and bounded possibly nonconvex losses, we present the first generalization bounds for the ZoSS algorithm. These bounds coincide with those for SGD, and they are independent of d, K and the batch size m, under appropriate choices of a slightly decreased learning rate. For bounded nonconvex losses and a batch size $m = 1$, we additionally show that both generalization error and learning rate are independent of d and K, and remain essentially the same as for the SGD, even for two function evaluations. Our results extensively extend and consistently recover established results for SGD in prior work, on both generalization bounds and corresponding learning rates. If additionally $m = n$, where n is the dataset size, we recover generalization guarantees for full-batch GD as well.

1 Introduction

Learning methods often rely on empirical risk minimization objectives that highly depend on a limited training data-set. Known gradient-based approaches such as SGD train and generalize effectively in reasonable time [1]. In contrast, emerging applications such as convex bandits [2–4], black-box learning [5], federated learning [6], reinforcement learning [7, 8], learning linear quadratic regulators [9, 10], and hyper-parameter tuning [11] stand in need of *gradient-free learning algorithms* [11–14] due to an unknown loss/model or impossible gradient evaluation.

Given two or more function evaluations, zeroth-order algorithms (see, e.g., [14, 15]) aim to estimate the true gradient for evaluating and updating model parameters (say, of dimension d). In particular, Zeroth-order Stochastic Search (ZoSS) [13, Corollary 2], [16, Algorithm 1] uses $K+1$ function evaluations ($K \geq 1$), while deterministic zeroth-order approaches [5, Section 3.3] require at least $K \geq d+1$ queries. The optimization error of the ZoSS algorithm is optimal as shown in prior work for convex problems [13], and suffers at most a factor of $\sqrt{d/K}$ in the convergence rate as compared with SGD. In addition to the optimization error, the importance of generalization error raises the question of how well zeroth-order algorithms generalize to unseen examples. In this paper, we show that the generalization error of ZoSS essentially coincides with that of SGD, under the choice of a slightly decreased learning rate. Assuming a Lipschitz and smooth loss function, we

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

establish generalization guarantees for ZoSS, by extending stability-based analysis for SGD [1], to the gradient-free setting. In particular, we rely on the celebrated result that uniform algorithmic stability implies generalization [1, 17, 18].

Early works [17, 19–22] first introduced the notion of stability, and the connection between (uniform) stability and generalization. Recently, alternative notions of stability and generalization gain attention such as locally elastic stability [23], VC-dimension/flatness measures [24], distributional stability [25–27], information theoretic bounds [16, 28–33] mainly based on assuming a sub-Gaussian loss, as well as connections between differential privacy and generalization [34–37].

In close relation to our paper, Hardt et al. [1] first showed uniform stability final-iterate bounds for vanilla SGD. More recent works develop alternative generalization error bounds based on high probability guarantees [38–41] and data-dependent variants [42], or under different assumptions than those of prior works such as as strongly quasi-convex [43], non-smooth convex [44–47], and pairwise losses [48, 49]. In the nonconvex case, [50] provide bounds that involve on-average variance of the stochastic gradients. Generalization performance of other algorithmic variants lately gain further attention, including SGD with early momentum [51], randomized coordinate descent [52], look-ahead approaches [53], noise injection methods [54], and stochastic gradient Langevin dynamics [55–62].

Recently, stability and generalization of full-bath GD has also been studied; see, e.g., [63–67]. In particular, Charles and Papailiopoulos. [64] showed instability of GD for nonconvex losses. Still, such instability does not imply a lower bound on the generalization error of GD (in expectation). In fact, Hoffer et al. [63] showed empirically that the generalization of GD is not affected by the batch-size, and for large enough number of iterations GD generalizes comparably to SGD. Our analysis agree with the empirical results of Hoffer et al. [63], as we show that (for smooth losses) the generalization of ZoSS (and thus of SGD) is independent of the batch size.

Notation. We denote the training data-set S of size n as $\{z_i\}_{i=1}^n$, where z_i are i.i.d. observations of a random variable Z with unknown distribution $\mathcal{D}$. The parameters of the model are vectors of dimension d, denoted by $W \in \mathbb{R}^d$, and W_t is the output at time t of a (randomized) algorithm A_S. The (combined) loss function $f(\cdot, z) : \mathbb{R}^d \to \mathbb{R}^+$ is uniformly Lipschitz and smooth for all $z \in \mathcal{Z}$. We denote the Lipschitz constant as L and the smoothness parameter by β. The number of function (i.e., loss) evaluations (required at each iteration of the ZoSS algorithm) is represented by $K + 1 \in \mathbb{N}$. We denote by Δf the smoothed approximation of the loss gradient, associated with parameter μ. The parameter $\Gamma_K^d \triangleq \sqrt{(3d-1)/K} + 1$ prominently appears in our results. We denote the gradient of the loss function with respect to model parameters W, by $\nabla f(W, z) \equiv \nabla_w f(w, z)|_{w=W}$. We denote the mini batch at t by J_t, and $m \triangleq |J_t|$.

1.1 Contributions

Under the assumption of Lipschitz and smooth loss functions, we provide generalization guarantees for black-box learning, extending the analysis of prior work by Hardt et al. [1] to the gradient free setting. In particular, we establish uniform stability and generalization error bounds for the final iterate of the ZoSS algorithm; see Table 1 for a summary of the results. In more detail, the contributions of this work are as follows:

- For unbounded *and* bounded losses, we show generalization error bounds identical to SGD, with a slightly decreased learning rate. Specifically, the generalization error bounds are independent of the dimension d, the number of evaluations K and the batch-size m. Further, a large enough number of evaluations (K) provide fast generalization even in the high dimensional regime.
- For bounded nonconvex losses and single (example) query updates ($m = 1$), we show that both the ZoSS generalization error and learning rate are independent of d and K, similar to that of SGD [1, Theorem 3.8]. This property guarantees efficient generalization even with two function evaluations.
- In the full information regime (i.e., when the number of function evaluations K grow to ∞), the ZoSS generalization bounds also provide guarantees for SGD by recovering the results in prior work [1]. Further, we derive novel SGD bounds for unbounded nonconvex losses, as well as mini-batch SGD for any batch size. Our results subsume generalization guarantees for full-batch ZoSS and GD algorithms.

Generalization Error Bounds: ZoSS vs SGD				
Algorithm	Bound	NC	UB	MB
ZoSS (**this work**) $\alpha_t \leq C/(t\Gamma_K^d)$	$\frac{1+(C\beta)^{-1}}{n}\big((2+c)CL^2\big)^{\frac{1}{C\beta+1}}(eT)^{\frac{C\beta}{C\beta+1}}$	✓	✗	✗
SGD, $\alpha_t \leq C/t$ Hardt et al. [1]	$\frac{1+(C\beta)^{-1}}{n}\big(2CL^2\big)^{\frac{1}{C\beta+1}}(eT)^{\frac{C\beta}{C\beta+1}}$	✓	✗	✗
ZoSS (**this work**) $\alpha_t \leq C/t$	$\frac{3e\big(1+(C\beta)^{-1}\big)^2}{2n}\big(1+(2+c)CL^2\big)T$ (*independent of both* d *and* K)	✓	✗	✗
ZoSS (**this work**) $\alpha_t \leq \frac{\log\left(1+\frac{C\beta}{\Gamma_K^d}(\Gamma_K^d-1)\right)}{T\beta\sqrt{(3d-1)/K}}$	$\frac{(2+c)CL^2}{n}$	✗	✓	✓
SGD, $\alpha_t \leq C/T$ Hardt et al. [1]	$\frac{2CL^2}{n}$	✗	✓	✓
ZoSS (**this work**) $\alpha_t \leq C/(T\Gamma_K^d)$	$\frac{(2+c)L^2(e^{C\beta}-1)}{n\beta}$	✓	✓	✓
ZoSS (**this work**) $\alpha_t \leq \frac{\log(1+C\beta)}{T\beta\Gamma_K^d}$	$\frac{(2+c)CL^2}{n}$ (proper choice of C in previous bound)	✓	✓	✓
ZoSS (**this work**) $\alpha_t \leq C/(t\Gamma_K^d)$	$\frac{(2+c)L^2(eT)^{C\beta}}{n}\min\{C+\beta^{-1}, C\log(T)\}$	✓	✓	✓

Table 1: A list of the generalization error bounds developed herein for ZoSS (Eq. 6) in comparison with SGD, with $\mu \leq cL\Gamma_K^d/n\beta(3+d)^{3/2}$, for $c > 0$. In the table, "NC" and "UB" stand for "nonconvex" and "unbounded", respectively. "MB" corresponds to the mini-batch algorithm and for any batch size. Also, α_t denotes the stepsize of ZoSS/SGD, and T the total number of iterations.

2 Problem Statement

Given a data $S \triangleq \{z_i\}_{i=1}^n$ of i.i.d samples z_i from an unknown distribution $\mathcal{D}$, our goal is to find the parameters w^* of a learning model such that $w^* \in \arg\min_w R(w)$, where $R(w) \triangleq \mathbb{E}_{Z\sim\mathcal{D}}[f(w,Z)]$. Since the distribution $\mathcal{D}$ is not known, we consider the empirical risk

$$R_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n f(w, z_i), \tag{1}$$

and the corresponding empirical risk minimization (ERM) problem to find $w_s^* \in \arg\min_w R_S(w)$. For a (randomized) algorithm A_S with input S and output $W = A(S)$, the excess risk ϵ_{excess} is bounded by the sum of the generalization error ϵ_{gen} and the optimization error ϵ_{opt},

$$\epsilon_{\text{excess}} \triangleq \mathbb{E}_{S,A}[R(W)] - R(w^*) = \underbrace{\mathbb{E}_{S,A}[R(W) - R_S(W)]}_{\epsilon_{\text{gen}}} + \underbrace{(\mathbb{E}_{S,A}[R_S(W)] - R(w^*))}_{\epsilon_{\text{opt}}}. \tag{2}$$

To analyze and control ϵ_{gen}, we prove uniform stability bounds which imply generalization [1, Theorem 2.2]. Specifically, if for all i.i.d. sequences $S, S' \in \mathcal{Z}^n$ that differ in one entry, we have $\sup_z \mathbb{E}_A[f(A(S),z) - f(A(S'),z)] \leq \epsilon_{\text{stab}}$, for some $\epsilon_{\text{stab}} > 0$, then $\epsilon_{\text{gen}} \leq \epsilon_{\text{stab}}$. Because the loss is L-Lipschitz, ϵ_{stab} may then be chosen as $L\sup_{S,S'}\mathbb{E}_A\|A(S) - A(S')\|$.

Our primary goal in this work is to develop uniform stability bounds for a *gradient-free* algorithm A_S of the form $w_{t+1} = w_t - \alpha_t \Delta f_{w_t,z}$, where $\Delta f_{w_t,z}$ only depends on loss function evaluations. To achieve this without introducing unnecessary assumptions, we consider a novel algorithmic stability error decomposition approach. In fact, the stability error introduced at time t by A_S breaks down into the stability error of SGD and an approximation error due to missing gradient information. Let $G_t(\cdot)$

and $G'_t(\cdot)$ be the following SGD update rules

$$G_t(w) \triangleq w - \alpha_t \nabla f(w, z_{i_t}), \quad G'_t(w) \triangleq w - \alpha_t \nabla f(w, z'_{i_t}), \tag{3}$$

under inputs S, S' respectively, and let $i_t \in \{1, 2, \ldots, n\}$ be a random index chosen uniformly and independently by the random selection rule of the algorithm, for all $t \leq T$. Similarly we use the notation $\tilde{G}(\cdot)$ and $\tilde{G}'(\cdot)$ to denote the iteration mappings of A_S, i.e.,

$$\tilde{G}_t(w) \triangleq w - \alpha_t \Delta f_{w, z_{i_t}}, \quad \tilde{G}'_t(w) \triangleq w - \alpha_t \Delta f_{w, z'_{i_t}}. \tag{4}$$

Then, as we also discuss later on (Lemma 1), the iterate stability error $\tilde{G}_t(w) - \tilde{G}'_t(w')$ of A_S, for any $w, w' \in \mathbb{R}^d$ and for all at $t \leq T$, may be decomposed as

$$\tilde{G}_t(w) - \tilde{G}'_t(w') \propto \underbrace{G_t(w) - G'_t(w')}_{\epsilon_{\text{GBstab}}} + \underbrace{\left[\nabla f(w, z_{i_t}) - \Delta f_{w, z_{i_t}}\right] + \left[\nabla f(w', z'_{i_t}) - \Delta f_{w', z'_{i_t}}\right]}_{\epsilon_{\text{est}}}, \tag{5}$$

where ϵ_{GBstab} denotes the gradient-based stability error (associated with SGD), and ϵ_{est} denotes the gradient approximation error. We now proceed by formally introducing ZoSS.

3 Zeroth-Order Stochastic Search (ZoSS)

As a *gradient-free* alternative of the classical SGD algorithm, we consider the ZoSS scheme, with iterates generated according to the following (single-example update) rule

$$W_{t+1} = W_t - \alpha_t \frac{1}{K} \sum_{k=1}^{K} \frac{f(W_t + \mu U_k^t, z_{i_t}) - f(W_t, z_{i_t})}{\mu} U_k^t, \quad U_k^t \sim \mathcal{N}(0, I_d), \quad \mu \in \mathbb{R}^+, \tag{6}$$

where $\alpha_t \geq 0$ is the corresponding learning rate (for the mini-batch update rule we refer the reader to Section 5). At every iteration t, ZoSS generates K i.i.d. standard normal random vectors $U_k^t, k = 1, \ldots, K$, and obtains $K + 1$ loss evaluations on perturbed model inputs. Then ZoSS evaluates a smoothed approximation of the gradient for some $\mu > 0$. In light of the discussion in Section 2, we define the ZoSS smoothed gradient step at time t as

$$\Delta f_{w, z_{i_t}}^{K, \mu} \equiv \Delta f_{w, z_{i_t}}^{K, \mu, \mathbf{U}^t} \triangleq \frac{1}{K} \sum_{k=1}^{K} \frac{f(w + \mu U_k^t, z_{i_t}) - f(w, z_{i_t})}{\mu} U_k^t. \tag{7}$$

3.1 ZoSS Stability Error Decomposition

To show stability bounds for ZoSS, we decompose its error into two parts through the stability error decomposition discussed in Section 2. Under the ZoSS update rule, Eq. (5) holds by considering the directions $\Delta f_{w, z_{i_t}}$ and $\Delta f_{w', z'_{i_t}}$ according to ZoSS smoothed approximations (7). Then for any $w, w' \in \mathbb{R}^d$, the iterate stability error $\tilde{G}_t(w) - \tilde{G}'_t(w')$ of ZoSS at t, breaks down into the gradient based error ϵ_{GBstab} and approximation error ϵ_{est}.

The error term ϵ_{GBstab} expresses the stability error of the gradient based mappings [1, Lemma 2.4] and inherits properties related to the SGD update rule. The error ϵ_{est} captures the approximation error of the ZoSS smoothed approximation and depends on K and μ. The consistency of the smoothed approximation with respect to SGD follows from $\lim_{K \uparrow \infty, \mu \downarrow 0} \Delta f_{w,z}^{K,\mu} = \nabla f(w, z)$ for all $w \in \mathbb{R}$ and $z \in \mathcal{Z}$. Further, the stability error is also consistent since $\lim_{K \uparrow \infty, \mu \downarrow 0} |\epsilon_{\text{est}}| = 0$. Later on, we use the ZoSS error decomposition in Eq. (5) together with a variance reduction lemma (Lemma 10), to derive exact expressions on the iterate stability error $\tilde{G}_t(w) - \tilde{G}'_t(w')$ for fixed K and $\mu > 0$ (see Lemma 1). Although in this paper we derive stability bounds and bounds on the ϵ_{gen}, the excess risk ϵ_{excess} depends on both errors ϵ_{gen} and ϵ_{opt}. In the following section, we briefly discuss known results on the ϵ_{opt} of zeroth-order methods, including convex and nonconvex losses.

3.2 Optimization Error in Zeroth-Order Stochastic Approximation

Convergence rates of the ZoSS optimization error and related zeroth-order variants have been extensively studied in prior works; see e.g., [14, 15, 68]. For the convex loss setting, when $K + 1$

function evaluations are available and no other information regarding the loss is given, the ZoSS algorithm achieves optimal rates with respect to the optimization error ϵ_{opt}. Specifically, under the assumption of a closed and convex loss, Duchi et al. [13] provided a lower bound for the minimax convergence rate and showed that $\epsilon_{\text{opt}} = \Omega(\sqrt{d/K})$, for any algorithm that approximates the gradient given $K+1$ evaluations. In the nonconvex setting Ghadimi et al. [69, 70] established sample complexity guarantees for the zeroth-order approach to reach an approximate stationary point.

4 Main Results

For our analysis, we introduce the same assumptions on the loss function (Lipschitz and smooth) as appears in prior work [1]. Additionally, we exploit the η-expansive and σ-bounded properties of the SGD mappings $G_t(\cdot)$ and $G'_t(\cdot)$ in Eq. (3).[1] The mappings $G_t(\cdot)$ and $G'_t(\cdot)$ are introduced for analysis purposes due to the stability error decomposition given in Eq. (5) and no further assumptions or properties are required for the zeroth-order update rules $\tilde{G}_t(\cdot)$ and $\tilde{G}'_t(\cdot)$ given in Eq. (4). The η-expansivity of $G_t(\cdot)$ holds for $\eta = 1 + \beta\alpha_t$ if the loss is nonconvex, and $\eta = 1$ if the loss is convex and $\alpha_t \leq 2/\beta$ [1, Lemma 3.6]. Note that $G_t(\cdot)$ is always σ-bounded ($\sigma = L\alpha_t$) [1, Lemma 3.3.].

4.1 Stability Analysis

We derive generalization error bounds through uniform stability. To study the stability of ZoSS, we apply a variance reduction lemma that we provide in Appendix A. Exploiting the variance reduction lemma, we show a growth recursion lemma for the iterates of the ZoSS.

Lemma 1 (ZoSS Growth Recursion) *Consider the sequences of updates $\{\tilde{G}_t\}_{t=1}^T$ and $\{\tilde{G}'_t\}_{t=1}^T$. Let $w_0 = w'_0$ be the starting point, $w_{t+1} = \tilde{G}_t(w_t)$ and $w'_{t+1} = \tilde{G}'_t(w'_t)$ for any $t \in \{1, \ldots, T\}$. Then for any $w_t, w'_t \in \mathbb{R}^d$ and $t \geq 0$ the following recursion holds*

$$\mathbb{E}[\|\tilde{G}_t(w_t) - \tilde{G}'_t(w'_t)\|] \leq \begin{cases} \left(\eta + \alpha_t\sqrt{\frac{3d-1}{K}}\beta\right)\|w_t - w'_t\| + \mu\beta\alpha_t(3+d)^{3/2}, & \text{if } \tilde{G}_t(\cdot) = \tilde{G}'_t(\cdot), \\ \|w_t - w'_t\| + 2\alpha_t L\Gamma_K^d + \mu\beta\alpha_t(3+d)^{3/2}, & \text{if } \tilde{G}_t(\cdot) \neq \tilde{G}'_t(\cdot). \end{cases}$$

The growth recursion of ZoSS characterizes the stability error that it is introduced by the ZoSS update and according to the outcome of the randomized selection rule at each iteration. Lemma 1 extends growth recursion results for SGD in prior work [1, Lemma 2.5] to the setting of the ZoSS algorithm. If $K \to \infty$ and $\mu \to 0$ (while the rest of the parameters are fixed), then $\Gamma_K^d \to 1$, and the statement recovers that of the SGD [1, Lemma 2.5].

Proof of Lemma 1. Let S and S' be two samples of size n differing in only a single example, and let $\tilde{G}_t(\cdot), \tilde{G}'_t(\cdot)$ be the update rules of the ZoSS for each of the sequences S, S' respectively. First under the event $\mathcal{E}_t \triangleq \{\tilde{G}_t(\cdot) \equiv \tilde{G}'_t(\cdot)\}$ (see Eq. (4)), by applying the Taylor expansion there exist vectors $W^*_{k,t}$ and $W^\dagger_{k,t}$ with j^{th} coordinates in the intervals $\left(w_t^{(j)}, w_t^{(j)} + \mu U_{k,t}^{(j)}\right) \cup \left(w_t^{(j)} + \mu U_{k,t}^{(j)}, w_t^{(j)}\right)$ and $\left(w_t'^{(j)}, w_t'^{(j)} + \mu U_{k,t}^{(j)}\right) \cup \left(w_t'^{(j)} + \mu U_{k,t}^{(j)}, w_t'^{(j)}\right)$, respectively, such we find that for any $w_t, w'_t \in \mathbb{R}^d$ it is true that

$$\begin{aligned} &\tilde{G}_t(w_t) - \tilde{G}'_t(w'_t) = \tilde{G}_t(w_t) - \tilde{G}_t(w'_t) \\ &= w_t - w'_t - \frac{\alpha_t}{K}\sum_{k=1}^K \langle \nabla f(w_t, z_{i_t}) - \nabla f(w'_t, z_{i_t}), U_k^t\rangle U_k^t \\ &\quad - \frac{\alpha_t}{K}\sum_{k=1}^K \left(\frac{\mu}{2} U_k^{\mathrm{T}} \nabla_w^2 f(w, z_{i_t})|_{w=W^*_{k,t}} U_k^t\right) U_k^t + \frac{\alpha_t}{K}\sum_{k=1}^K \left(\frac{\mu}{2} U_k^{\mathrm{T}} \nabla_w^2 f(w, z_{i_t})|_{w=W^\dagger_{k,t}} U_k^t\right) U_k^t \\ &= \underbrace{w_t - \alpha_t \nabla f(w_t, z_{i_t})}_{G(w_t)} - \underbrace{(w'_t - \alpha_t \nabla f(w'_t, z_{i_t}))}_{G'(w'_t) \equiv G(w'_t)} \end{aligned} \tag{8}$$

[1] [1, Definition 2.3]: An update rule $G(\cdot)$ is η-expansive if $\|G(w) - G(w')\| \leq \eta\|w - w'\|$ for all $w, w' \in \mathbb{R}^d$. If $\|w - G(w)\| \leq \sigma$ then it is σ-bounded.

$$- \frac{\alpha_t}{K}\sum_{k=1}^{K}\left(\frac{\mu}{2}U_k^{\mathrm{T}}\nabla_w^2 f(w,z_{i_t})|_{w=W_{k,t}^*}U_k^t\right)U_k^t + \frac{\alpha_t}{K}\sum_{k=1}^{K}\left(\frac{\mu}{2}U_k^{\mathrm{T}}\nabla_w^2 f(w,z_{i_t})|_{w=W_{k,t}^\dagger}U_k^t\right)U_k^t$$

$$- \alpha_t\left(\frac{1}{K}\sum_{k=1}^{K}\langle\nabla f(w_t,z_{i_t})-\nabla f(w_t',z_{i_t}),U_k^t\rangle U_k^t - (\nabla f(w_t,z_{i_t})-\nabla f(w_t',z_{i_t}))\right). \tag{9}$$

We find (9) by adding and subtracting $\alpha_t\nabla f(w_t,z_{i_t})$ and $\alpha_t\nabla f(w_t',z_{i_t})$ in Eq. (8). Recall that U_k^t are independent for all $k\le K$, $t\le T$ and that the mappings $G(\cdot)$ and $G'(\cdot)$ defined in Eq. (9), are η-expansive. The last display and the triangle inequality give

$$\mathbb{E}[\|\tilde{G}_t(w_t)-\tilde{G}_t(w_t')\|]$$

$$\le \|G(w_t)-G(w_t')\| + \frac{2\alpha_t}{K}\sum_{k=1}^{K}\frac{\mu\beta}{2}\mathbb{E}\left[\|U_k^t\|^3\right] + \alpha_t\sqrt{\frac{3d-1}{K}}\mathbb{E}[\|\nabla f(w_t,z_{i_t})-\nabla f(w_t',z_{i_t})\|] \tag{10}$$

$$\le \eta\|w_t-w_t'\| + \frac{2\alpha_t}{K}\sum_{k=1}^{K}\frac{\mu\beta}{2}\mathbb{E}\left[\|U_k^t\|^3\right] + \alpha_t\sqrt{\frac{3d-1}{K}}\beta\|w_t-w_t'\| \tag{11}$$

$$\le \left(\eta+\alpha_t\sqrt{\frac{3d-1}{K}}\beta\right)\|w_t-w_t'\| + \mu\beta\alpha_t(3+d)^{3/2}, \tag{12}$$

where (10) follows from (9) and Lemma 10, and for (11) we applied the η-expansive property of $G(\cdot)$ (see [1, Lemma 2.4 and Lemma 3.6]) and the β-smoothness of the loss function.[2] Finally (12) holds since the random vectors $U_k^t\sim\mathcal{N}(0,I_d)$ are identically distributed for all $k\in\{1,2,\dots,K\}$ and $\mathbb{E}\|U_k^t\|^3\le(3+d)^{3/2}$. Eq. (12) gives the first part of the recursion.

Similar to (9), under the event $\mathcal{E}_t^c\triangleq\{\tilde{G}_t(\cdot)\neq\tilde{G}_t'(\cdot)\}$, we find

$$\tilde{G}_t(w_t)-\tilde{G}_t'(w_t')$$

$$= \underbrace{w_t-\alpha_t\nabla f(w_t,z_{i_t})}_{G(w_t)} - \underbrace{\left(w_t'-\alpha_t\nabla f(w_t',z_{i_t}')\right)}_{G'(w_t')}$$

$$- \frac{\alpha_t}{K}\sum_{k=1}^{K}\left(\frac{\mu}{2}U_k^{\mathrm{T}}\nabla_w^2 f(w,z_{i_t})|_{w=\tilde{W}_{k,t}^*}U_k^t\right)U_k^t + \frac{\alpha_t}{K}\sum_{k=1}^{K}\left(\frac{\mu}{2}U_k^{\mathrm{T}}\nabla_w^2 f(w,z_{i_t}')|_{w=\tilde{W}_{k,t}^\dagger}U_k^t\right)U_k^t$$

$$- \alpha_t\left(\frac{1}{K}\sum_{k=1}^{K}\langle\nabla f(w_t,z_{i_t})-\nabla f(w_t',z_{i_t}'),U_k^t\rangle U_k^t - (\nabla f(w_t,z_{i_t})-\nabla f(w_t',z_{i_t}'))\right). \tag{13}$$

By using the last display, triangle inequality, Lemma 10 and β-smoothness, we find

$$\mathbb{E}[\|\tilde{G}_t(w_t)-\tilde{G}_t(w_t')\|]$$

$$\le \|G(w_t)-G'(w_t')\| + \frac{2\alpha_t}{K}\sum_{k=1}^{K}\frac{\mu\beta}{2}\mathbb{E}[\|U_k^t\|^3] + \alpha_t\sqrt{\frac{3d-1}{K}}\mathbb{E}[\|\nabla f(w_t,z_{i_t})-\nabla f(w_t',z_{i_t}')\|]$$

$$\le \min\{\eta,1\}\delta_t + 2\sigma_t + \frac{2\alpha_t}{K}\sum_{k=1}^{K}\frac{\mu\beta}{2}\mathbb{E}[\|U_k^t\|^3] + 2L\alpha_t\sqrt{\frac{3d-1}{K}} \tag{14}$$

$$\le \delta_t + 2\alpha_t L\Gamma_K^d + \mu\beta\alpha_t(3+d)^{3/2}, \tag{15}$$

where (14) follows from the triangle inequality and $L-$Lipschitz condition, while the upper bound on $\|G(w_t)-G'(w_t')\|$ comes from [1, Lemma 2.4]. Finally, (15) holds since $\eta\ge1$ for both convex and nonconvex losses, $\sigma_t=L\alpha_t$ and $\mathbb{E}\|U_k^t\|^3\le(3+d)^{3/2}$ for all $k\in\{1,\dots,K\}$. This shows the second part of recursion. □

For sake of brevity, let $\mathcal{I}$ be an adapted stopping time that corresponds to the first iteration index that the single distinct instance of the two data-sets S,S' is sampled by ZoSS. For any $t_0\in\{0,1,\dots,n\}$ we define the event $\mathcal{E}_{\delta_{t_0}}\triangleq\{\mathcal{I}>t_0\}\equiv\{\delta_{t_0}=0\}$. The next result provides the stability bound.

[2] For all $z\in\mathcal{Z}$ and $W\in\mathbb{R}^d$ it is true that $\|\nabla_w^2 f(w,z)|_{w=W}\|\le\beta$.

Lemma 2 (ZoSS Stability | Nonconvex Loss) *Assume that the loss function $f(\cdot, z)$ is L-Lipschitz and β-smooth for all $z \in \mathcal{Z}$. Consider the ZoSS algorithm* (6) *with final-iterate estimates W_T and W'_T, corresponding to the data-sets S, S', respectively (that differ in exactly one entry). Then the discrepancy $\delta_T \triangleq \|W_T - W'_T\|$, under the event $\mathcal{E}_{\delta_{t_0}}$, satisfies the inequality*

$$\mathbb{E}[\delta_T|\mathcal{E}_{\delta_{t_0}}] \leq \left(\frac{2L}{n}\Gamma_K^d + \mu\beta(3+d)^{3/2}\right) \sum_{t=t_0+1}^{T} \alpha_t \prod_{j=t+1}^{T} \left(1 + \beta\alpha_j\Gamma_K^d\left(1 - \frac{1}{n}\right)\right). \tag{16}$$

The corresponding bound of Lemma 2 for convex losses is slightly tighter than the bound in (16). Since the two bounds differ only by a constant, the consequent results of Lemma 2 are essentially identical for convex losses as well. We provide the equivalent version of Lemma 2 for convex losses in Appendix B.

Proof of Lemma 2. Consider the events $\mathcal{E}_t \triangleq \{\tilde{G}_t(\cdot) \equiv \tilde{G}'_t(\cdot)\}$ and $\mathcal{E}_t^c \triangleq \{\tilde{G}_t(\cdot) \neq \tilde{G}'_t(\cdot)\}$ (see Eq. (4)). Recall that $\mathbb{P}(\mathcal{E}_t) = 1 - 1/n$ and $\mathbb{P}(\mathcal{E}_t^c) = 1/n$ for all $t \leq T$. For any $t_0 \geq 0$, a direct application of Lemma 1 gives

$$\begin{aligned}
\mathbb{E}[\delta_{t+1}|\mathcal{E}_{\delta_{t_0}}] &= \mathbb{P}(\mathcal{E}_t)\mathbb{E}[\delta_{t+1}|\mathcal{E}_t, \mathcal{E}_{\delta_{t_0}}] + \mathbb{P}(\mathcal{E}_t^c)\mathbb{E}[\delta_{t+1}|\mathcal{E}_t^c, \mathcal{E}_{\delta_{t_0}}] \\
&= \left(1 - \frac{1}{n}\right)\mathbb{E}[\delta_{t+1}|\mathcal{E}_t, \mathcal{E}_{\delta_{t_0}}] + \frac{1}{n}\mathbb{E}[\delta_{t+1}|\mathcal{E}_t^c, \mathcal{E}_{\delta_{t_0}}] \\
&\leq \left(\eta + \alpha_t\beta\sqrt{\frac{3d-1}{K}} + \frac{1}{n}\left(1 - \eta - \alpha_t\beta\sqrt{\frac{3d-1}{K}}\right)\right)\mathbb{E}[\delta_t|\mathcal{E}_{\delta_{t_0}}] \\
&\quad + \frac{2\alpha_t L}{n}\Gamma_K^d + \mu\beta\alpha_t(3+d)^{3/2}.
\end{aligned} \tag{17}$$

With $R_t \triangleq (\eta + \alpha_t\beta(\Gamma_K^d - 1) + (1 - \eta - \alpha_t\beta(\Gamma_K^d - 1))/n)$ solving the recursion in (17) gives

$$\mathbb{E}[\delta_T|\mathcal{E}_{\delta_{t_0}}] \leq \left(\frac{2L}{n}\Gamma_K^d + \mu\beta(3+d)^{3/2}\right) \sum_{t=t_0+1}^{T} \alpha_t \prod_{j=t+1}^{T} R_j. \tag{18}$$

We consider the last inequality for nonconvex loss functions with $\eta = 1 + \beta\alpha_t$ and convex loss functions with $\eta = 1$ to derive Lemma 2 and Lemma 11 respectively (Appendix B). □

4.2 Generalization Error Bounds

For the first generalization error bound, we evaluate the right part of the inequality (16) for decreasing step size and bounded nonconvex loss. Then the Lipschitz condition provides a uniform stability condition for the loss and yields the next theorem.

Theorem 3 (Nonconvex Bounded Loss | Decreasing Stepsize) *Assume that the loss $f(\cdot, z) \in [0, 1]$ is L-Lipschitz and β-smooth for all $z \in \mathcal{Z}$. Consider the ZoSS update rule* (6) *with T the total number of iterates, $\alpha_t \leq C/t\Gamma_K^d$ for some (fixed) $C > 0$ and for all $t \leq T$, and fixed $\mu \leq cL\Gamma_K^d/n\beta(3+d)^{3/2}$ for some $c > 0$. Then the generalization error of ZoSS is bounded by*

$$|\epsilon_{\text{gen}}| \leq \frac{\left((2+c)CL^2\right)^{\frac{1}{C\beta+1}}(eT)^{\frac{C\beta}{C\beta+1}}}{n} \max\left\{1, 1 + (C\beta)^{-1} - \frac{e^{\beta C}}{\beta C^{\frac{1}{C\beta+1}}}\left(\frac{(2+c)L^2}{eT}\right)^{\frac{C\beta}{C\beta+1}}\right\} \tag{19}$$

$$\leq \frac{\left(1 + (C\beta)^{-1}\right)\left((2+c)CL^2\right)^{\frac{1}{C\beta+1}}}{n}(eT)^{\frac{C\beta}{C\beta+1}}. \tag{20}$$

Inequality (19), as a tighter version of (20), provides a meaningful bound in marginal cases, i.e.,

$$\lim_{\beta \downarrow 0} \mathbb{E}\left[|f(W_T, z) - f(W'_T, z)|\right] \leq \frac{(2+c)CL^2}{n} \max\left\{\log\left(\frac{eT}{(2+c)CL^2}\right), 1\right\}. \tag{21}$$

By neglecting the negative term in (19) we find (20), that is the ZoSS equivalent of SGD [1, Theorem 3.8]. When $K \to \infty$ and $c \to 0$, then $\Gamma_K^d \to 1$, and the inequalities (19), (20) reduce to a

generalization bound for SGD. Inequality (20) matches that of [1, Theorem 3.8], and (19) provides a tighter generalization bound for SGD as well. We show Theorem 3 in Appendix A.

Next, we provide a bound on the generalization error for nonconvex losses that comes directly from Theorem 3. In contrast to Theorem 3, the next result provides learning rate and a generalization error bounds, both of which are independent of the dimension and the number of function evaluations.

Corollary 4 *Assume that the loss function $f(\cdot, z) \in [0, 1]$ is L-Lipschitz and β-smooth for all $z \in \mathcal{Z}$. Consider the ZoSS update rule* (6) *with $\mu \leq cL\Gamma_K^d/(n\beta(3+d)^{3/2})$, T the total number of iterates, and $\alpha_t \leq C/t$ for some (fixed) $C > 0$ and for all $t \leq T$. Then the generalization error of ZoSS is bounded by*

$$|\epsilon_{\text{gen}}| \leq \left(1 + (\beta C)^{-1}\right)^2 \left(1 + (2+c)CL^2\right) \frac{3Te}{2n}. \tag{22}$$

As a consequence, even in the high dimensional regime $d \to \infty$, two function evaluations (i.e., $K = 1$) are sufficient for the ZoSS to achieve $\epsilon_{\text{gen}} = \mathcal{O}(T/n)$, with the learning rate being no smaller than that of SGD. We continue by providing the proof of Theorem 3. For the proof of Corollary 4, see Appendix A.3.

In light of Theorem 3 and Corollary 4, we observe that the over-fitting phenomenon occurs in the gradient-free approach similarly to gradient-based algorithms. For general nonconvex (and convex) losses under standard step-size choices, the generalization error increases with respect to T. Further, the effect of β affects both the stability (similarly to SGD in prior work) of the algorithm and the error approximation of the ZoSS. If β is large, then the expected approximation error (due to limited function evaluations) is also large [15] and the dependence on smoothness is unavoidable in black-box learning. In our results, this is expressed through the Growth Recursion of ZoSS (Lemma 2), that involves both the stability and approximation error per iteration. However, a smaller step-size ($\alpha_t = 1/2t\beta\Gamma_K^d$) mitigates the effect of β on the bound. We refer the reader to Appendix E for a unified analysis of the excess risk, that captures the over-fitting and under-fitting trade-off.

Additionally, the number of iterations T is considered to be fixed and known (as in prior works including on average and high probability results on generalization). This is reasonable and quite standard because given the theoretical results, we know beforehand the appropriate choices of T that provide a good trade-off between generalization and optimization. A classical setting is that of a fixed step-size $\alpha_t = 1/T$ with $T = \sqrt{n}$, which provides the well known generalization error bound for SGD with order $\mathcal{O}(1/\sqrt{n})$, as appears in very recent and timely prior works [32, Section 3.1], [45].

In the unbounded loss case, we apply Lemma 2 by setting $t_0 = 0$ (recall that t_0 is a free parameter, while the algorithm depends on the random variable $\mathcal{I}$). The next result provides a generalization error bound for the ZoSS algorithm with constant step size. In the first case of the theorem, we also consider the convex loss as a representative result, as we show the same bound holds for an appropriate choice of greater learning rate than the learning rate of the nonconvex case. The convex case for the rest of the results of this work can be similarly derived.

Theorem 5 (Unbounded Loss | Constant Step Size) *Assume that the loss $f(\cdot, z)$ is L-Lipschitz, β-smooth for all $z \in \mathcal{Z}$. Consider the ZoSS update rule* (6) *with $\mu \leq cL\Gamma_K^d/(n\beta(3+d)^{3/2})$ for some $c > 0$. Let T be the total number of iterates and for any $t \leq T$,*

- *if $f(\cdot, z)$ is convex for all $z \in \mathcal{Z}$ and $\alpha_t \leq \min\{\log\left(1 + C\beta(1 - 1/\Gamma_K^d)\right)/T\beta(\Gamma_K^d - 1), 2/\beta\}$, or if $f(\cdot, z)$ is nonconvex and $\alpha_t \leq \log\left(1 + C\beta\right)/T\beta\Gamma_K^d$, for $C > 0$ then*

$$|\epsilon_{\text{gen}}| \leq \frac{(2+c)CL^2}{n}, \tag{23}$$

- *if $f(\cdot, z)$ is nonconvex and $\alpha_t \leq C/T\Gamma_K^d$, for some $C > 0$, then*

$$|\epsilon_{\text{gen}}| \leq \frac{L^2\,(2+c)\,(e^{C\beta} - 1)}{n\beta}. \tag{24}$$

For the proof of Theorem 5 see Appendix A.4. In the following, we present the generalization error of ZoSS for an unbounded loss with a decreasing step size. Recall that the results for unbounded nonconvex loss also hold for the case of a convex loss with similar bounds on the generalization error and learning rate (see the first case of Theorem 5).

Theorem 6 (Unbounded Loss | Decreasing Step Size) *Assume that the loss $f(\cdot, z)$ is L-Lipschitz, β-smooth for all $z \in \mathcal{Z}$. Consider ZoSS with update rule* (6)*, T the total number of iterates, $\alpha_t \leq C/t\Gamma_K^d$ for all $t \leq T$ and for some $C > 0$, and $\mu \leq cL\Gamma_K^d/(n\beta(3+d)^{3/2})$ for some $c > 0$. Then the generalization error of ZoSS is bounded by*

$$|\epsilon_{\text{gen}}| \leq \frac{(2+c)L^2(eT)^{C\beta}}{n} \min\left\{C + \beta^{-1}, C\log(eT)\right\}. \tag{25}$$

For the proof of Theorem 6 see Appendix A.5. Note that the constant C is free and controls the learning rate. Furthermore, it quantifies the trade-off between the speed of training and the generalization of the algorithm. In the next section, we consider the ZoSS algorithm with a mini-batch of size m for which we provide generalization error bounds. These results hold under the assumption of unbounded loss and for any batch size m including the case $m = 1$.

5 Generalization of Mini-Batch ZoSS

For the *mini-batch* version of ZoSS, at each iteration t, the randomized selection rule (uniformly) samples a batch J_t of size m and evaluates the new direction of the update by averaging the smoothed approximation $\Delta f_{w,z}^{K,\mu}$ over the samples $z \in J_t$ as

$$\Delta f_{w,J_t}^{K,\mu} \equiv \Delta f_{w,J_t}^{K,\mu,\mathbf{U}^t} \triangleq \frac{1}{mK}\sum_{i=1}^{m}\sum_{k=1}^{K}\frac{f(w+\mu U_{k,i}^t, z_{J_{t,i}}) - f(w, z_{J_{t,i}})}{\mu}U_{k,i}^t, \tag{26}$$

where $U_{k,i}^t \sim \mathcal{N}(0, I_d)$ are i.i.d. (standard normal), and $\mu \in \mathbb{R}^+$. The update rule of the mini-batch ZoSS is $W_{t+1} = W_t - \alpha_t \Delta f_{W_t,J_t}^{K,\mu}$ for all $t \leq T$, and we define $\tilde{G}_{J_t}(w) \triangleq w - \alpha_t \Delta f_{w,J_t}^{K,\mu}$, $\tilde{G}'_{J'_t}(w) \triangleq w - \alpha_t \Delta f_{w,J'_t}^{K,\mu}$ for $J_t \subset S$ and $J'_t \subset S'$ respectively. Due to space limitation, we refer the reader to Appendix C for the detailed stability analysis of ZoSS with mini-batch. Specifically, we prove a growth recursion lemma for the mini-batch ZoSS updates (see Appendix C.1 for proof).

Lemma 7 (Mini-Batch ZoSS Growth Recursion) *Consider the sequences of updates $\{\tilde{G}_{J_t}\}_{t=1}^T$ and $\{\tilde{G}'_{J_t}\}_{t=1}^T$ and $\mu \leq cL\Gamma_K^d/(n\beta(3+d)^{3/2})$. Let $w_0 = w'_0$ be the starting point, $w_{t+1} = \tilde{G}_{J_t}(w_t)$ and $w'_{t+1} = \tilde{G}'_{J_t}(w'_t)$ for any $t \in \{1, \ldots, T\}$. Then for any $w_t, w'_t \in \mathbb{R}^d$ and $t \geq 0$ the following recursion holds*

$$\mathbb{E}[\|\tilde{G}_{J_t}(w_t) - \tilde{G}'_{J_t}(w'_t)\|] \leq \begin{cases} \left(1 + \beta\alpha_t\Gamma_K^d\right)\delta_t + \frac{cL\alpha_t}{n}\Gamma_K^d & \text{if } \tilde{G}_{J_t}(\cdot) = \tilde{G}'_{J_t}(\cdot) \\ \left(1 + \frac{m-1}{m}\beta\alpha_t\Gamma_K^d\right)\delta_t + \frac{2L\alpha_t}{m}\Gamma_K^d + \frac{cL\alpha_t}{n}\Gamma_K^d & \text{if } \tilde{G}_{J_t}(\cdot) \neq \tilde{G}'_{J_t}(\cdot). \end{cases}$$

Although the iterate stability error (at time t) in the growth recursion depends on the batch size m under the event $\{\tilde{G}_{J_t}(\cdot) \neq \tilde{G}'_{J_t}(\cdot)\}$, the stability bound on the final iterates is independent of m, and coincides with the single example updates ($m = 1$, Lemma 2). Herein, we provide an informal statement of the result.

Lemma 8 (Mini-Batch ZoSS Stability | Nonconvex Loss) *Consider the mini-batch ZoSS with any batch size $m \leq n$, and iterates $W_{t+1} = W_t - \alpha_t \Delta f_{W_t,J_t}^{K,\mu}$, $W'_{t+1} = W'_t - \alpha_t \Delta f_{W'_t,J'_t}^{K,\mu}$, for all $t \leq T$, with respect to the sequences S, S'. Then the stability error δ_T satisfies the inequality of Lemma 2.*

We refer the reader to Appendix Section C.1, Theorem 14 for the formal statement of the result.[3] Through the Lipschitz condition of the loss and Lemma 8, we show that the mini-batch ZoSS enjoys the same generalization error bounds as in the case of single-query ZoSS ($m = 1$). As a consequence, the batch size does not affect the generalization error.

Theorem 9 (Mini-batch ZoSS | Generalization Error) *Let the loss function $f(\cdot, z)$ be L-Lipschitz and β-smooth (possibly nonconvex, possibly unbounded) for all $z \in \mathcal{Z}$. Then the bounds of Theorem 5 and Theorem 6 hold for the mini-batch ZoSS with iterate $W_{t+1} = W_t - \alpha_t \Delta f_{W_t,J_t}^{K,\mu}$, for all $t \leq T$ and any batch size $m \leq n$.*

[3] As in the single-query ($m = 1$) ZoSS, under the assumption of convex loss, the stability error of mini-batch ZoSS satisfies the inequality (46), Appendix B, Lemma 11.

By letting $K \to \infty$ and $c \to 0$, the generalization error bounds of mini-batch ZoSS reduce to those of mini-batch SGD, extending results of the single-query ($m = 1$) SGD that appeared in prior work [1]. Additionally, once $K \to \infty$, $c \to 0$ and $m = n$ we obtain generalization guarantees for full-batch GD. For the sake of clarity and completeness we provide dedicated stability and generalization analysis of full-batch GD in Appendix D, Corollary 15.

6 Discussion: Black-box Adversarial Attack Design and Future Work

A standard, well-cited example of ZoSS application is adversarial learning as considered in [5], when the gradient is not known for the adversary (for additional applications for instance federated/reinforcement learning, linear quadratic regulators; see also Section 1 for additional references). Notice that the algorithm in [5] is restrictive in the high dimensional regime since it requires $2d$ function evaluations per iteration. In contrast, ZoSS can be considered with any $K \geq 2$ functions evaluations (the trade-off is between accuracy and resource allocation, which is also controlled through K). If $K = d + 1$ evaluations are available we recover guarantees for the deterministic zeroth-order approaches (similar to [5]).

Retrieving a large number of function evaluations often is not possible in practice. When a limited amount of function evaluations is available, the adversary obtains the solution (optimal attack) with an optimization error that scales by a factor of $\sqrt{d/K}$, and the generalization error of the attack is of the order $\sqrt{T}/n$ under appropriate choices of the step-size, the smoothing parameter μ and K. Fine tuning of the these parameters might be useful in practice, but in general K should be chosen as large as possible. In contrast, μ should be small and satisfy the inequality $\mu \leq cL\Gamma_K^d/n\beta(3+d)^{3/2}$ (Theorem 6). For instance, in practice μ is often chosen between 10^{-10} and 10^{-8} (or even lower) and the ZoSS algorithm remains (numerically) stable.

For neural networks with smooth activation functions [71–73], the ZoSS algorithm does not require the smoothness parameter β to be necessarily known, however if β is large then the guarantees of the estimated model would be pessimistic. To ensure that the learning procedure is successful, the adversary can approximate β (since the loss is not known) by estimating the (largest eigenvalue of the) Hessian through the available function evaluations [74, Section 4.1].

Although the non-smooth (convex) loss setting lies out of the scope of this work, it is expected to inherit properties and rates of the SGD for non-smooth losses (at least for sufficiently small smoothing parameter μ). In fact, [45, page 3, Table 1] developed upper and lower bounds for the SGD in the non-smooth case, and they showed that standard step-size choices provide vacuous stability bound. Due to these inherent issues of non-smooth (and often convex only cases), the generalization error analysis of ZoSS for non-smooth losses remains open. Finally, information-theoretic generalization error bounds of ZoSS can potentially provide further insight into the problem, due to the noisy updates of the algorithm, and consist part of future work.

7 Conclusion

In this paper, we characterized the generalization ability of black-box learning models. Specifically, we considered the Zeroth-order Stochastic Search (ZoSS) algorithm, which evaluates smoothed approximations of the unknown gradient of the loss by only relying on $K + 1$ loss evaluations. Under the assumptions of a Lipschitz and smooth (unknown) loss, we showed that the ZoSS algorithm achieves the same generalization error bounds as that of SGD, while the learning rate is slightly decreased compared to that of SGD. The efficient generalization ability of ZoSS, together with strong optimality results related to the optimization error by Duchi et al. [13], makes it a robust and powerful algorithm for a variety of black-box learning applications and problems.

References

[1] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL: https://proceedings.mlr.press/v48/hardt16.html.

[2] Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper/2011/file/67e103b0761e60683e83c559be18d40c-Paper.pdf.

[3] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017. URL: https://www.jmlr.org/papers/volume18/16-632/16-632.pdf.

[4] Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9017–9027. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/6646b06b90bd13dabc11ddba01270d23-Paper.pdf.

[5] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017. URL: https://dl.acm.org/doi/10.1145/3128572.3140448.

[6] Zan Li and Li Chen. Communication-efficient decentralized zeroth-order method on heterogeneous data. In *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6, 2021. doi:10.1109/WCSP52459.2021.9613620.

[7] Anirudh Vemula, Wen Sun, and J. Bagnell. Contrasting exploration in parameter and action space: A zeroth-order optimization perspective. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2926–2935. PMLR, 16–18 Apr 2019. URL: https://proceedings.mlr.press/v89/vemula19a.html.

[8] Harshat Kumar, Dionysios S. Kalogerias, George J. Pappas, and Alejandro Ribeiro. Actor-only deterministic policy gradient via zeroth-order gradient oracles in action space. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1676–1681, 2021. doi:10.1109/ISIT45174.2021.9518023.

[9] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2916–2925. PMLR, 16–18 Apr 2019. URL: https://proceedings.mlr.press/v89/malik19a.html.

[10] Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R. Jovanović. On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, 5(3):989–994, 2021. doi:10.1109/LCSYS.2020.3006256.

[11] Jeremy Rapin and Olivier Teytaud. Nevergrad - A gradient-free optimization platform. https://GitHub.com/FacebookResearch/Nevergrad, 2018.

[12] Luis Miguel Rios and Nikolaos V Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013. URL: https://link.springer.com/article/10.1007/s10898-012-9951-y.

[13] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi:10.1109/TIT.2015.2409256.

[14] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. URL: https://link.springer.com/article/10.1007/s10208-015-9296-2.

[15] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42, 2021. URL: https://link.springer.com/article/10.1007/s10208-021-09499-8.

[16] Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26370–26381. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/ddbc86dc4b2fbfd8a62e12096227e068-Paper.pdf.

[17] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002. URL: https://www.jmlr.org/papers/v2/bousquet02a.html.

[18] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010. URL: https://www.jmlr.org/papers/volume11/shalev-shwartz10a/shalev-shwartz10a.pdf.

[19] Luc P. Devroye and et al. Distribution-free performance bounds for potential function rules, 1979. URL: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.8772.

[20] Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999. doi:10.1162/089976699300016304.

[21] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006. URL: https://link.springer.com/article/10.1007/s10444-004-7634-z.

[22] Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001. URL: https://proceedings.neurips.cc/paper/2000/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[23] Zhun Deng, Hangfeng He, and Weijie Su. Toward better generalization bounds with locally elastic stability. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2590–2600. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/deng21b.html.

[24] Yiding Jiang*, Behnam Neyshabur*, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=SJgIPJBFvH.

[25] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 1046–1059, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2897518.2897566.

[26] Jonathan Ullman, Adam Smith, Kobbi Nissim, Uri Stemmer, and Thomas Steinke. The limits of post-selection generalization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/77ee3bc58ce560b86c2b59363281e914-Paper.pdf.

[27] Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. *A New Analysis of Differential Privacy's Generalization Guarantees (Invited Paper)*, page 9. Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3406325.3465358.

[28] Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020. URL: https://proceedings.mlr.press/v125/steinke20a.html.

[29] Ibrahim Alabdulmohsin. An Information-Theoretic Route from Generalization in Expectation to Generalization in Probability. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings*

of Machine Learning Research, pages 92–100. PMLR, 20–22 Apr 2017. URL: https://proceedings.mlr.press/v54/alabdulmohsin17a.html.

[30] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/ad71c82b22f4f65b9398f76d8be4c615-Paper.pdf.

[31] Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591, 2019. doi:10.1109/ISIT.2019.8849590.

[32] Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3526–3545. PMLR, 15–19 Aug 2021. URL: https://proceedings.mlr.press/v134/neu21a.html.

[33] Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24670–24682. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/cf0d02ec99e61a64137b8a2c3b03e030-Paper.pdf.

[34] Fengxiang He, Shaopeng Fu, Bohan Wang, and Dacheng Tao. Robustness, privacy, and generalization of adversarial training. *arXiv preprint arXiv:2012.13573*, 2020. URL: https://arxiv.org/abs/2012.13573.

[35] Fengxiang He, Bohan Wang, and Dacheng Tao. Tighter generalization bounds for iterative differentially private learning algorithms. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 802–812. PMLR, 27–30 Jul 2021. URL: https://proceedings.mlr.press/v161/he21a.html.

[36] Zhenhuan Yang, Yunwen Lei, Siwei Lyu, and Yiming Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2026–2034. PMLR, 13–15 Apr 2021. URL: https://proceedings.mlr.press/v130/yang21c.html.

[37] Puyu Wang, Zhenhuan Yang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private empirical risk minimization for auc maximization. *Neurocomputing*, 461:419–437, 2021.

[38] Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/05a624166c8eb8273b8464e8d9cb5bd9-Paper.pdf.

[39] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1270–1279. PMLR, 25–28 Jun 2019. URL: https://proceedings.mlr.press/v99/feldman19a.html.

[40] Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv e-prints*, pages arXiv–2006, 2020. URL: https://arxiv.org/abs/2006.05610.

[41] Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate o(1/n). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5065–5076. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/286674e3082feb7e5afb92777e48821f-Paper.pdf.

[42] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2815–2824. PMLR, 10–15 Jul 2018. URL: https://proceedings.mlr.press/v80/kuzborskij18a.html.

[43] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/qian19b.html.

[44] Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/8c01a75941549a705cf7275e41b21f0d-Paper.pdf.

[45] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4381–4391. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/2e2c4bf7ceaa4712a72dd5ee136dc9a8-Paper.pdf.

[46] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5809–5819. PMLR, 13–18 Jul 2020. URL: https://proceedings.mlr.press/v119/lei20c.html.

[47] Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *The Journal of Machine Learning Research*, 22(25):1–41, 2021. URL: http://jmlr.org/papers/v22/19-716.html.

[48] Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of SGD for pairwise learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21216–21228. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/b1301141feffabac455e1f90a7de2054-Paper.pdf.

[49] Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21236–21246. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/f3173935ed8ac4bf073c1bcd63171f8a-Paper.pdf.

[50] Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of SGD in nonconvex optimization. *Machine Learning*, pages 1–31, 2021. URL: https://link.springer.com/article/10.1007/s10994-021-06056-w.

[51] Ali Ramezani-Kebrya, Ashish Khisti, and Ben Liang. On the generalization of stochastic gradient descent with momentum. *arXiv preprint arXiv:2102.13653*, 2021.

[52] Puyu Wang, Liang Wu, and Yunwen Lei. Stability and generalization for randomized coordinate descent. *arXiv preprint arXiv:2108.07414*, 2021. URL: https://arxiv.org/abs/2108.07414.

[53] Pan Zhou, Hanshu Yan, Xiaotong Yuan, Jiashi Feng, and Shuicheng Yan. Towards understanding why lookahead generalizes better than sgd and beyond. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27290–27304. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf.

[54] Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors,

Advances in Neural Information Processing Systems, volume 34, pages 26523–26535. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/df1f1d20ee86704251795841e6a9405a-Paper.pdf.

[55] Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550, 2018. doi:10.1109/ISIT.2018.8437571.

[56] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 605–638. PMLR, 06–09 Jul 2018. URL: https://proceedings.mlr.press/v75/mou18a.html.

[57] Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=SkxxtgHKPS.

[58] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/05ae14d7ae387b93370d142d82220f1b-Paper.pdf.

[59] Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pritham Pingali, Chao Chen, and Mayank Goswami. Stability of SGD: Tightness analysis and improved bounds. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL: https://openreview.net/forum?id=Sl-zm08j5lq.

[60] Tyler Farghly and Patrick Rebeschini. Time-independent generalization bounds for SGLD in non-convex settings. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19836–19846. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/a4ee59dd868ba016ed2de90d330acb6a-Paper.pdf.

[61] Bohan Wang, Huishuai Zhang, Jieyu Zhang, Qi Meng, Wei Chen, and Tie-Yan Liu. Optimizing information-theoretical generalization bound via anisotropic noise of SGLD. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26080–26090. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/db2b4182156b2f1f817860ac9f409ad7-Paper.pdf.

[62] Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of SGLD using properties of Gaussian channels. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24222–24234. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/cb77649f5d53798edfa0ff40dae46322-Paper.pdf.

[63] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/a5e0ff62be0b08456fc7f1e88812af3d-Paper.pdf.

[64] Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 745–754. PMLR, 10–15 Jul 2018. URL: https://proceedings.mlr.press/v80/charles18a.html.

[65] Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman

Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25033–25043. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/d27b95cac4c27feb850aaa4070cc4675-Paper.pdf.

[66] Idan Amir, Tomer Koren, and Roi Livni. SGD generalizes better than GD (and regularization doesn't help). In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 63–92. PMLR, 15–19 Aug 2021. URL: https://proceedings.mlr.press/v134/amir21a.html.

[67] Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8609–8621. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/483101a6bc4e6c46a86222eb65fbcb6a-Paper.pdf.

[68] Yurii Nesterov. Random gradient-free minimization of convex functions. core discussion papers 2011001, université catholique de louvain. *Center for Operations Research and Econometrics (CORE)*, 2011. URL: https://econpapers.repec.org/paper/corlouvco/2011001.htm.

[69] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *Found Comput Math*, 23:35–76, 2022. URL: https://doi.org/10.1007/s10208-021-09499-8.

[70] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016. URL: https://link.springer.com/content/pdf/10.1007/s10107-014-0846-1.pdf.

[71] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. URL: https://arxiv.org/abs/1606.08415.

[72] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. URL: https://arxiv.org/abs/1710.05941.

[73] Hao Zheng, Zhanlei Yang, Wenju Liu, Jizhong Liang, and Yanpeng Li. Improving deep neural networks using softplus units. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–4, 2015. doi:10.1109/IJCNN.2015.7280459.

[74] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022. URL: https://doi.org/10.1137/120880811.

[75] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015. URL: https://arxiv.org/abs/1509.01240.

[76] Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the Lambert W function. *Advances in Computational mathematics*, 5(1):329–359, 1996. URL: https://link.springer.com/content/pdf/10.1007/BF02124750.pdf.

Acknowledgments and Disclosure of Funding

We would like to thank the four anonymous reviewers for providing valuable comments and suggestions, which have improved the presentation of the results and the overall quality of our paper.

Amin Karbasi acknowledges funding in direct support of this work from NSF (IIS-1845032), ONR (N00014- 19-1-2406), and the AI Institute for Learning-Enabled Optimization at Scale (TILOS).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See abstract and introduction.
   (b) Did you describe the limitations of your work? [Yes] See Section 4 for discussions on the assumptions adopted, which are standard and common in the related literature.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This paper presents theoretical contributions of generic nature and broad applicability.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Table 1, as well as all statement of our results.
   (b) Did you include complete proofs of all theoretical results? [Yes] All results include detailed proofs, either in the paper or in the supplement.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]