

Stability vs Implicit Bias of Gradient Methods on Separable Data and Beyond

Matan Schliserman

Blavatnik School of Computer Science, Tel Aviv University

SCHLISERMAN@MAIL.TAU.AC.IL

Tomer Koren

Blavatnik School of Computer Science, Tel Aviv University, and Google Research, Tel Aviv

TKOREN@TAUEX.TAU.AC.IL

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

An influential line of recent work has focused on the generalization properties of unregularized gradient-based learning procedures applied to separable linear classification with exponentially-tailed loss functions. The ability of such methods to generalize well has been attributed to their implicit bias towards large margin predictors, both asymptotically as well as in finite time. We give an additional unified explanation for this generalization and relate it to two simple properties of the optimization objective, that we refer to as realizability and self-boundedness. We introduce a general setting of unconstrained stochastic convex optimization with these properties, and analyze generalization of gradient methods through the lens of algorithmic stability. In this broader setting, we obtain sharp stability bounds for gradient descent and stochastic gradient descent which apply even for a very large number of gradient steps, and use them to derive general generalization bounds for these algorithms. Finally, as direct applications of the general bounds, we return to the setting of linear classification with separable data and establish several novel test loss and test accuracy bounds for gradient descent and stochastic gradient descent for a variety of loss functions with different tail decay rates. In some of these cases, our bounds significantly improve upon the existing generalization error bounds in the literature.

1. Introduction

There is a significant interest nowadays in understanding the generalization properties of unregularized gradient-based learning procedures. This has been motivated by empirical observations in the context of modern deep learning, where minimizing the training error without any explicit attempt to constrain model complexity or to avoid overfitting using explicit regularization, often results with models that generalizes well. It has been argued that this phenomenon is explained by the “implicit bias” of the learning algorithm, whereby the dynamics of the optimization method itself serve as a form of regularization that biases the algorithm towards favorable “simple” models that will not overfit (e.g., [Neyshabur et al., 2014, 2017](#)).

In this context, the setting of linear classification with separable data has attracted particular attention. A compelling result of [Soudry et al. \(2018\)](#) showed that by simply minimizing the (unregularized) empirical risk over a linearly-separable training set using plain gradient descent, the trained predictor asymptotically converges (in direction) to the max-margin solution over the dataset, provided that we use an exponentially-tailed classification loss such as the logistic loss. Consequently, by virtue of standard margin-based generalization bounds for linear predictors, we obtain that the gradient descent solution does not overfit when the number of steps T is sufficiently large, even though its magnitude (i.e., Euclidean norm) grows indefinitely with T due to the lack of

explicit regularization. This result has been later extended in several ways to accommodate for other optimization algorithms and loss functions (Ji and Telgarsky, 2018, 2019; Nacson et al., 2019a,b; Ji et al., 2020).

In terms of non-asymptotic rates, the convergence in direction to a large margin solution established in these works is rather slow and decays only logarithmically with T ; the implied generalization error bounds for the trained predictors thus come into effect only when T is very large, and at least exponentially large in the size of the training set and in the maximal margin. In a more recent work, via a refined analysis of the margins attained by the normalized iterates of gradient methods, Shamir (2021) established finite-time generalization error bounds that apply to smaller and more realistic values of T . For example, for gradient descent used to minimize the average empirical logistic loss over a separable training set of size n with margin γ , these bounds assume the form $\tilde{O}(1/\gamma^2 T + 1/\gamma^2 n)$, suppressing logarithmic factors in T . As discussed by Shamir (2021), such a bound is essentially optimal to within these logarithmic factors.

Notably, all of the aforementioned generalization error bounds apply to the zero-one accuracy of the normalized predictor, rather than to the loss of the unnormalized model actually being optimized. In fact, a main popularized takeaway from this line of work is that generalization in terms of zero-one prediction accuracy keeps improving as the number of gradient steps T grows, regardless of the test loss which could actually increase asymptotically (see a detailed discussion in Soudry et al., 2018). In other words, by inspecting the loss on a validation set one might be led to think that the learning algorithm is overfitting as T grows, whereas the underlying predictor actually keeps improving. However, the increase in loss Soudry et al. (2018) demonstrate (theoretically) is again only logarithmic in T , and thus affects only the very late stages of optimization. A natural question is then: are the zero-one generalization guarantees strictly better than the test loss guarantees also earlier in the optimization, after a realistic number of steps? And how do the achievable test loss bounds depend on properties of the particular loss function? Even further, could we in fact benefit, in terms of generalization, by an early-stopping of the optimization process?

In this paper, we give positive answers to these questions and provide sharp test loss bounds that match, up to logarithmic factors, analogous existing bounds for the zero-one accuracy. First, for exponentially-decaying losses such as the logistic loss, we give a bound of the form $\tilde{O}(1/\gamma^2 T + 1/\gamma^2 n)$ over the test loss of the model trained with T steps of either GD or SGD over a separable training set of size n and margin γ . Thus, up to logarithmic factors, the real-valued loss of the trained (unnormalized) model exhibits the same ideal generalization rates known for the zero-one prediction accuracy. In other words, unless the number of steps T is exponentially large (in $1/\gamma$ and n), the test loss does in fact diminish to zero at a nearly optimal rate, closely following the zero-one accuracy.

We also establish analogous results for a variety of other loss functions, including polynomially-tailed, sub-exponentially tailed, and super-exponentially tailed losses; these are summarized in Table 1. Since the loss functions we consider are all surrogates of the zero-one loss, these bounds immediately imply identical bounds for the test zero-one prediction accuracy of the trained models. To the best of our knowledge, these are the first non-asymptotic results for those loss functions in this context.

In fact, our analysis applies much more broadly than just to separable linear classification. We consider a general unbounded stochastic convex optimization problem of the form $\min_w F(w) := \mathbb{E}_{z \sim \mathcal{D}}[f(w, z)]$ where the objective f is a nonnegative and smooth convex function over \mathbb{R}^d , and identify two simple and general conditions on f that enable sharp generalization bounds, as in the more specific separable linear classification setup, for both gradient descent (GD) and stochastic

gradient descent (SGD). The first is a “realizability” condition that captures the rate at which the tail of f diminishes to zero as the Euclidean norm of w grows; this is merely a quantitative analogue of the separability assumption that an optimal zero population loss F is attained “at infinity.” The second condition is a “self-boundedness” property of the gradient operator of f , of the form $\|\nabla f(w, z)\| \leq c(f(w, z))^{1-\delta}$ for all w and z , for some constants $c, \delta \geq 0$. As has been observed in previous work (Ji and Telgarsky, 2018), while for general nonnegative 1-smooth functions this condition holds with $\delta = \frac{1}{2}$ (and $c = \sqrt{2}$; see, e.g., Nesterov, 2003), exponentially-tailed losses satisfy a stronger version of this relation with $\delta = 0$. More generally, we observe that a similar self-boundedness condition holds in fact for a large variety of loss functions, albeit with a possibly different setting of δ and c .

Given that these realizability and self-boundedness conditions hold, and assuming $\delta \rightarrow 0$ and $c = \Theta(1)$ to simplify this informal presentation, our generalization loss bounds for both GD and SGD take the form $O(\varepsilon + \rho(\varepsilon)^2/T + \rho(\varepsilon)^2/n)$ for an arbitrary $\varepsilon > 0$, where $\rho(\varepsilon)$ stands for the minimal possible norm $\|w^*\|$ of a reference point $w^* \in \mathbb{R}^d$ where the test loss drops below ε . (The existence of such a point is guaranteed by our realizability condition.) For example, in the specific case of separable linear classification with margin γ and an exponentially-tailed loss, it is not hard to show that there exists a reference point with test loss $< \varepsilon$ and norm $\rho(\varepsilon) = O((1/\gamma) \log(1/\varepsilon))$, which implies the bound we stated earlier for this case if we set $\varepsilon = \Theta(1/T)$. Similarly, the other results in Table 1 follow directly from this general bound, up to small variations in the self-boundedness parameters δ and c , and whether or not the loss function is Lipschitz on the entire space.

To prove these results, we appeal to algorithmic stability arguments (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009). Typical stability bounds for gradient methods degrade (at least) linearly with the number of gradient steps (Hardt et al., 2016; Chen et al., 2018; Bassily et al., 2020; Attia and Koren, 2021). We show that under the self-boundedness condition with a sufficiently small δ , a significantly more moderate degradation in stability is possible. Notably, our stability bounds scale roughly with the optimization error of the gradient method iterates, so that a smaller training loss directly translates to a better stability bound, and in turn, to a tighter bound on the generalization gap. Thus, in contrast to traditional generalization error analyses that aim to strike a balance between the training error and the generalization gap, our analysis shows that for self-bounded objectives (with sufficiently small δ) these in fact do not stand at odds with each other, and improving the former directly leads to stronger generalization guarantees. We remark that recently Lei and Ying (2020) established refined stability bounds under a realizability assumption similar to ours; however, since they do not make use of the stronger self-boundedness condition, their bounds degrades quickly (roughly linearly) with the number of gradient steps.

1.1. Summary of contributions

To summarize, the main contribution of this paper are as follows:

- We introduce a general setting of unconstrained stochastic convex optimization, that captures the well-studied setting of unregularized (and unconstrained) separable linear classification as a special case (see Section 2). In this framework, we identify two simple conditions—realizability and self-boundedness—under which strong generalization loss bounds are attainable.

Loss function	Tail behavior	$\rho(\varepsilon)$	(c, δ)	Generalization loss
Logistic	$\exp(-x)$	$\frac{1}{\gamma} \log(\frac{1}{\varepsilon})$	$(1, 0)$	$O\left(\frac{\log^2(T)}{\gamma^2 T} + \frac{\log^2(T)}{\gamma^2 n}\right)$
Multi-class logistic	$\exp(-x)$	$\frac{1}{\gamma} \log(\frac{K}{\varepsilon})$	$(2, 0)$	$O\left(\frac{\log^2(KT)}{\gamma^2 T} + \frac{\log^2(KT)}{\gamma^2 n}\right)$
Polynomial	$x^{-\alpha}, \alpha > 0$	$\frac{1}{\gamma} \varepsilon^{-1/\alpha}$	$(\alpha, 0)$	$O\left(\left(\frac{\alpha}{\gamma}\right)^{\frac{2\alpha}{2+\alpha}} \left(\frac{1}{T^{\frac{\alpha}{2+\alpha}}} + \frac{T^{\frac{2}{2+\alpha}}}{n}\right)\right)$
Sub-exponential	$\exp(-x^\alpha), \alpha < 1$	$\frac{1}{\gamma} \log^{1/\alpha}(\frac{1}{\varepsilon})$	$(\alpha, 0)$	$O\left(\frac{\alpha \log^{\frac{2}{\alpha}}(T)}{\gamma^2 T} + \frac{\alpha^2 \log^{\frac{2}{\alpha}}(T)}{\gamma^2 n}\right)$
Super-exponential	$\exp(-x^\alpha), \alpha > 1$	$\frac{1}{\gamma} \log^{1/\alpha}(\frac{1}{\varepsilon})$	$(\frac{\alpha}{2\delta}, \delta), \forall \delta > 0$	$O\left(\frac{\alpha^2 \log^{\frac{2}{\alpha}}(T)}{\gamma^2 T} + \frac{\alpha^3 \log^{\frac{2+\alpha}{\alpha}}(T)}{\gamma^2 n}\right)$
Probit	$\exp(-x^2)$	$\frac{1}{\gamma} \log^{1/2}(\frac{1}{\varepsilon})$	$(\frac{3}{\delta}, \delta), \forall \delta > 0$	$O\left(\frac{\log(T)}{T\gamma^2} + \frac{\log(T)^4}{n\gamma^4}\right)$

Table 1: Generalization loss bounds established in this paper for constant step-size GD and SGD on separable linear classification, for loss functions which satisfy Assumptions 1 and 2 along with the relevant parameters. Here, T is the number of gradient steps, n is the size of the training set, and γ is the margin over the training set. (For the multi-class logistic loss, K is the number of classes.)

- We obtain sharp algorithmic stability bounds for gradient descent and stochastic gradient descent in the general unconstrained stochastic convex optimization setup, which directly translate to bounds on the generalization gap (in terms of actual loss values) of the trained model (see Theorems 8, 10 and 11 in Section 3). Our stability bounds scale with the optimization error of the gradient method iterates and allows for a favorable trade-off between training error and generalization gap, which applies even after a very large number of gradient steps.
- Combined with standard convergence bounds for smooth optimization, we derive generalization loss bounds for gradient descent and stochastic gradient descent in the general stochastic convex optimization setting (see Theorems 12 and 14 in Section 4).
- Finally, as direct applications of the general bounds, we also achieve several new generalization results in the setting of unregularized linear classification with separable data (see Section 5). We consider a variety of loss functions with different tail decay rates, and establish test loss bounds for gradient descent and stochastic gradient descent for the corresponding classification problem. These bounds are summarized in Table 1, and to the best of our knowledge, are new to the literature.

1.2. Discussion and some implications

On the role of early stopping: For polynomially-tailed losses, with a tail decay rate of $x^{-\alpha}$ ($\alpha \geq 1$), our generalization error bounds are optimized when $T = n$, and degrade in quality polynomially-fast with T . Interestingly, the test performance implied by our upper bounds after $T = n$ steps of GD is strictly better than the performance of the asymptotic direction to which GD converges as $T \rightarrow \infty$: while Ji et al. (2020) show that the asymptotic margin is in general $\Theta(n^{-1/(\alpha+1)})$ (see their Proposition 12), which by standard margin bounds translates to a generalization error bound of $\Theta(n^{-(\alpha-1)/(\alpha+1)})$, we obtain a test error bound of $O(n^{-\alpha/(\alpha+2)})$, which is always better than the former bound—the

best that one can achieve without early stopping.¹ (E.g., for $\alpha = 1$ the asymptotic bound becomes vacuous, while our bound for $T = n$ is $O(n^{-1/3})$.) This addresses open questions from earlier works (e.g., Ji et al. (2020) mention this explicitly) regarding the role of early stopping: we see that with a polynomially-decaying loss, not only that early stopping does not harm test performance, but it can actually strictly improve the latter.

On the optimality of exponential tails: Our results shed some light on what enables exponentially-tailed losses in attaining optimal test loss bounds, and thus provides a different justification than that of Ji et al. (2020) for this widely popular choice of a loss function. Crucially, the rapid decay of an exponential tail allows for a faster optimization rate, as the training loss drops below ε at distance from origin of at most $O(\log(1/\varepsilon))$ as a function of ε . As already discussed, this faster optimization rate directly translates to a tighter generalization loss bound. It would be tempting to conclude that decay rate even faster than exponential would lead to better bounds. However, our analysis of super-exponentially decaying tails, e.g., tails of the form $\exp(-x^\alpha)$ for $\alpha > 1$, suggests that faster decaying tails actually allow for weaker stability bounds that become worse by a factor of α . Thus, somewhat remarkably, an exponentially-decaying tail strikes a fine balance between optimization rate and model stability.

On implicit bias vs. algorithmic stability: It is interesting to relate two very different aspects that govern generalization in the context of separable linear classification: implicit bias and algorithmic stability. Our results suggest that while the effect of implicit biases of gradient methods (towards large margin solutions) kicks in rather late in the optimization, stability seems to control generalization more tightly earlier on. Intriguingly, for exponentially-tailed losses that have been the extensive focus of previous work, the stability and implicit bias (i.e., margin based) arguments both imply the same tight generalization rates for the downstream prediction accuracy of the trained model. However, this does not seem to be the case more generally, e.g., in the case of polynomially-tailed losses, where generalization bounds based on stability are strictly better than those obtained from margin-based arguments: indeed, the analysis of Shamir (2021) in this case yielded suboptimal bounds compared to the asymptotic bounds of Ji et al. (2020) which, as we have just discussed, are weaker than our finite-time bounds based on stability.

2. Unconstrained Stochastic Convex Optimization

We consider the following setting of unconstrained and unregularized stochastic convex optimization over \mathbb{R}^d , that subsumes separable linear classification as a special case. Let \mathcal{D} be a distribution over a probability space \mathcal{Z} . We measure the loss of a model w on an example z by a loss function $f(w, z)$ where $f : \mathbb{R}^d \times \mathcal{Z} \rightarrow \mathbb{R}$ be a positive, convex and L -smooth function (with respect to w , for all $z \in \mathcal{Z}$). We are interested to find a prediction w which has a small *population loss*, defined as

$$F(w) = \mathbb{E}_{z \sim \mathcal{D}}[f(w, z)].$$

For finding such a model w , we use a set of training examples $S = \{z_1, \dots, z_n\}$ which drawn i.i.d. from \mathcal{D} and an empirical proxy, the *empirical loss*, defined as

$$\widehat{F}(w) = \frac{1}{n} \sum_{i=1}^n f(w, z_i). \tag{1}$$

1. For this discussion, we treat both the max-margin γ and the decay exponent α as constants; see the bounds in Table 1 for the precise dependence on these parameters.

2.1. The Realizability and Self-boundedness conditions

We additionally make the following two key assumptions on the objective f :

Assumption 1 (Realizability). There exists a monotonic decreasing function $\rho : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that for every $\varepsilon > 0$, there exists $w^* \in \mathbb{R}^d$ such that $f(w^*, z) \leq \varepsilon$ for all z , and $\|w^*\| \leq \rho(\varepsilon)$.

Assumption 2 (Self boundedness). There exist $c > 0$ and $0 \leq \delta \leq \frac{1}{2}$, such that $\|\nabla f(w, z)\| \leq c f(w, z)^{1-\delta}$ for all w and z .

As discussed in the introduction, Assumption 1 is a quantitative abstraction of the separability assumption in linear classification that an optimal zero population loss is attained “at infinity.” Without loss of generality, we will assume that $\rho(\varepsilon) \geq 1$ for all ε ; otherwise, we may replace ρ with the function $\varepsilon \rightarrow \max\{\rho(\varepsilon), 1\}$.

Assumption 2 on the other hand is the main assumption that enables our algorithmic stability and generalization results. We recall that the self-boundedness condition always holds with $c = \sqrt{2\beta}$ and $\delta = \frac{1}{2}$ for nonnegative β -smooth objectives (e.g., Nesterov, 2003). However, we observe that for many loss functions a significantly stronger version of this condition holds with $\delta \ll \frac{1}{2}$, which allows for stronger generalization bounds; see the examples in Table 1, and more details in Section 5. The generality of the self-boundedness condition can be seen from the following simple observation:

Lemma 3. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$ be a positive and differentiable function. Then,*

- (i) *f satisfies Assumption 2 with $\delta = 0, c > 0$ if and only if $\log(f)$ is c -Lipschitz;*
- (ii) *f satisfies Assumption 2 with $\delta > 0, c > 0$ if and only if f^δ is $c\delta$ -Lipschitz.*

For example, it is not hard to see that the exponential loss $f(x) = e^{-x}$ and the logistic loss admit the first property above for $c = 1$; and that a polynomial loss $f(x) = x^\alpha$ ($\alpha \geq 1$) admits the second property for $\delta = 1/\alpha$ and $c = \alpha$.

2.2. Example: linear classification with margin

The setting of linear classification with separable data is obtained as a special case of our main setting. We mainly consider binary classification with margin $\gamma > 0$, where the examples are labeled examples (x_i, y_i) and there exists a unit vector w^* such that $yw^* \cdot x \geq \gamma$ almost surely (with respect to a distribution \mathcal{D} over instances (x, y)). For brevity, we denote $z = y \cdot x$ and assume that data is scaled so that $\|z\| \leq 1$ with probability 1. In this case, we have $f(w, z_i) = \ell(w \cdot z_i)$ for all i , where ℓ is a convex, positive and monotonically decreasing loss function, such as the logistic loss $\ell(x) = \log(1 + \exp(-x))$. For such functions, Assumption 1 holds:

Lemma 4. *If $\ell : \mathbb{R} \rightarrow \mathbb{R}^+$ is positive, convex and strictly monotonically decreasing such that $\lim_{x \rightarrow \infty} \ell(x) = 0$. Then, the inverse ℓ^{-1} is a well defined function and $f(w, z) = \ell(w \cdot z)$ satisfies Assumption 1 with $\rho(\varepsilon) = (1/\gamma)\ell^{-1}(\varepsilon)$.*

Assumption 2 also holds for many asymptotically-vanishing loss functions. As discussed in more detail in Section 5, for exponentially, sub-exponentially, and polynomially tailed losses it is satisfied for $\delta = 0$; for super-exponentially tailed losses (including the Probit loss) Assumption 2 is satisfied for any $\delta > 0$, albeit with $c = \Theta(1/\delta)$. Table 1 summarizes the relevant parameters ρ, c, δ for the various loss functions.

3. Stability bounds for self-bounded objectives

For bounding the generalization gap we use algorithmic stability. In this section we define the notion of stability that we use, on-average-leave-one-out model stability and prove that popular gradient methods, like gradient descent and stochastic gradient descent, are stable with respect to this notion.

3.1. Preliminary: On-Average Leave-One-Out model stability

We first introduce the stability notions we will rely on, and note their connection to generalization. Specifically, that notion of stability that we consider is on-average leave-one-out (on-average-loo) model stability. For this definition, we assume without loss of generality that there exists an example $z_0 \in \mathcal{Z}$ for which $f(w, z_0) = 0$ for all w . (Otherwise, we can artificially augment the sample space with such an instance.) Now, given an i.i.d. sample $S = (z_1, \dots, z_n)$, with the corresponding empirical risk (Eq. (1)), we define the leave-one-out samples $S_i = (z_1, \dots, z_{i-1}, z_0, z_{i+1}, \dots, z_n)$ for all $i \in [n]$, with the corresponding empirical risks:

$$\forall i \in [n], \quad \widehat{F}_i(w) = \frac{1}{n} \sum_{z \in S_i} f(w, z) = \frac{1}{n} \sum_{j \neq i} f(w, z_j).$$

We can now define the on-average-loo model stability for learning algorithms. This notion of stability amounts to the distance (in Euclidean norm) between the output models of a learning algorithm on the the full sample S and the leave-one-out samples S_i , averaged over the choice of i . We refer to two notions of loo-model-stability ℓ_1 -on-average-loo model stability, which will be used for achieving a generalization bound for Lipschitz functions, and ℓ_2 -on-average-loo model stability, we be used for non-Lipschitz functions.

Throughout the paper, we bound the expected stability and generalization of learning algorithms. We denote by \mathbb{E}_S the expectation with respect to the sample $S \sim \mathcal{Z}^n$, and by \mathbb{E}_A the expectation with respect to the randomization in the learning algorithm (if it is randomized).

Definition 5 (on-average leave-out-out model stability). Let $A : \mathcal{Z}^n \rightarrow \mathbb{R}^d$ be a learning algorithm. We say that A is ℓ_1 -on-average-loo model ε -stable if for *any* sample S ,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_A [\|A(S) - A(S_i)\|] \leq \varepsilon, \quad (2)$$

and that A is ℓ_2 -on-average-loo model ε -stable if for any sample S ,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_A [\|A(S) - A(S_i)\|^2] \leq \varepsilon. \quad (3)$$

We will denote by $\varepsilon_{\text{stab}}^{(1)}$ the infimum over all ε for which Eq. (2) holds, and by $\varepsilon_{\text{stab}}^{(2)}$ the infimum over all ε for which Eq. (3) holds.

Note that for any algorithm A , the output $A(S_i)$ given the sample S_i viewed as a random variable that is independent of the sample point z_i . We can use this fact to get a connection between on-average-loo model stability and generalization.

Lemma 6. *Let A be an ℓ_1 - ε -on-average-loo model stable learning algorithm. Then, if for every z , $f(w, z)$ is convex and G -Lipschitz with respect to w ,*

$$\mathbb{E}_{S,A} [F(A(S))] \leq \mathbb{E}_{S,A} [\widehat{F}(A(S))] + 2G\varepsilon.$$

Lemma 7. *Let A be an ℓ_2 - ε -on-average-loo model stable learning algorithm. Then, if for every z , $f(w, z)$ is convex and L -smooth with respect to w ,*

$$\mathbb{E}_{S,A}[F(A(S))] \leq 4\mathbb{E}_{S,A}[\widehat{F}(A(S))] + 3L\varepsilon.$$

3.2. Gradient Descent

First, we establish a stability bound for GD under Assumption 2. We consider standard gradient descent (GD) with a fixed step size $\eta > 0$ applied to the empirical risk \widehat{F} ; this method is initialized at a point w_1 and at each step $t = 1, \dots, T$ performs an update

$$w_{t+1} = w_t - \eta \nabla \widehat{F}(w_t). \quad (4)$$

The algorithm returns the final model, w_T .

We begin with bounding the ℓ_1 -on-average-loo model stability of GD with self-boundedness.

Theorem 8. *Suppose Assumption 2 holds and $f(w, z)$ is positive, convex and L -smooth with respect to w , for all z . Then, GD with step size $\eta \leq 2/L$ is ℓ_1 -on-average-loo model stable with*

$$\varepsilon_{\text{stab}}^{(1)} \leq \frac{c\eta T^\delta}{n} \left(\sum_{t=1}^T \widehat{F}(w_t) \right)^{1-\delta}.$$

We will be mostly interested in the regime where Assumption 2 holds with a sufficiently small δ so that c and T^δ are constants (namely, where $\delta = O(1/\log T)$). In this case, we obtain a stability bound of the form $\varepsilon_{\text{stab}}^{(1)} = O(\frac{1}{n} \sum_{t=1}^T \widehat{F}(w_t))$, which increases very moderately with the number of steps T provided that GD properly minimizes the training error.

In our stability analysis below, we will use the following standard lemma in smooth convex optimization (e.g., [Hardt et al., 2016](#)).

Lemma 9. *If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and L -smooth and $0 < \eta \leq 2/L$, then for every $u, v \in \mathbb{R}^d$,*

$$\|(u - \eta \nabla f(u)) - (v - \eta \nabla f(v))\| \leq \|u - v\|.$$

Proof of Theorem 8. Let $\{w_t\}_t$ be the iterates of GD on \widehat{F} and $\{w_t^i\}_t$ be the iterates of GD on \widehat{F}_i . Then for every t ,

$$w_{t+1}^i = w_t^i - \frac{\eta}{n} \sum_{j \neq i} \nabla f(w_t^i, z_j).$$

As a result, by Lemma 9,

$$\begin{aligned} \|w_{t+1} - w_{t+1}^i\| &= \left\| w_t - \frac{\eta}{n} \sum_{j=1}^n \nabla f(w_t, z_j) - w_t^i + \frac{\eta}{n} \sum_{j \neq i} \nabla f(w_t^i, z_j) \right\| \\ &= \left\| \frac{1}{n} \sum_{j \neq i} (w_t - \eta \nabla f(w_t, z_j) - w_t^i + \eta \nabla f(w_t^i, z_j)) + \frac{1}{n} w_t - \frac{1}{n} w_t^i - \eta \nabla f(w_t, z_i) \right\| \\ &\leq \frac{1}{n} \sum_{j \neq i} \|w_t - \eta \nabla f(w_t, z_j) - w_t^i + \eta \nabla f(w_t^i, z_j)\| + \frac{1}{n} \|w_t - \eta \nabla f(w_t, z_i) - w_t^i\| \end{aligned}$$

$$\leq \frac{1}{n} \sum_{j \neq i}^n \|w_t - w_t^j\| + \frac{1}{n} \|w_t - w_t^i\| + \frac{\eta}{n} \|\nabla f(w_t, z_i)\|.$$

By Assumption 2 we get,

$$\|w_{t+1} - w_{t+1}^i\| \leq \|w_t - w_t^i\| + \frac{\eta c}{n} f(w_t, z_i)^{1-\delta}. \quad (5)$$

Now, the function $g(x) = x^{1-\delta}$ is concave, thus, by Jensen's inequality,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|w_{t+1} - w_{t+1}^i\| &\leq \frac{1}{n} \sum_{i=1}^n \left(\|w_t - w_t^i\| + \frac{\eta c}{n} f_i(w_t)^{1-\delta} \right) \\ &\leq \frac{1}{n} \sum_{i=1}^n \|w_t - w_t^i\| + \frac{\eta c}{n} \frac{1}{n} \sum_{i=1}^n f_i(w_t)^{1-\delta} \\ &\leq \frac{1}{n} \sum_{i=1}^n \|w_t - w_t^i\| + \frac{\eta c}{n} \widehat{F}(w_t)^{1-\delta}. \end{aligned}$$

By summing over $t = 1, \dots, T-1$ and another use of Jensen inequality, we get,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|w_T - w_T^i\| &\leq \frac{\eta c}{n} \sum_{t=1}^{T-1} \widehat{F}(w_t)^{1-\delta} \\ &\leq \frac{\eta c}{n} \sum_{t=1}^T \widehat{F}(w_t)^{1-\delta} \\ &\leq \frac{\eta T c}{n} \left(\frac{1}{T} \sum_{t=1}^T \widehat{F}(w_t) \right)^{1-\delta} \\ &= \frac{\eta T^\delta c}{n} \left(\sum_{t=1}^T \widehat{F}(w_t) \right)^{1-\delta}. \end{aligned}$$

■

Now we bound the ℓ_2 -on-average-loo model stability of Gradient Descent on self-bounded loss functions. The resulting bound is stated in the following theorem.

Theorem 10. *Suppose Assumption 2 holds. If, for every z , $f(w, z)$ is positive, L -smooth, convex with respect to w , then, GD with step size $\eta \leq 2/L$ is ℓ_2 -on-average-loo model stable with*

$$\varepsilon_{\text{stab}}^{(2)} \leq \frac{c^2 \eta^2 T^{2\delta}}{n^{1+2\delta}} \left(\sum_{t=1}^T \widehat{F}(w_t) \right)^{2(1-\delta)}.$$

Again, we will be mostly interested in the regime where c and T^δ are constants. In this case, we get a stability bound of the form $O(\frac{1}{n} (\sum_{t=1}^T \widehat{F}(w_t))^2)$, which again scales favorably with the number of steps T given that GD is properly tuned for convergence on the empirical risk.

Proof. Let $\{w_t\}_t$ be the iterates of GD on \widehat{F} and $\{w_t^i\}_t$ be the iterates of GD on \widehat{F}_i . By a similar derivation as in the proof of Theorem 8,

$$\|w_{t+1} - w_{t+1}^i\| \leq \|w_t - w_t^i\| + \frac{\eta c}{n} f(w_t, z_i)^{1-\delta}. \quad (6)$$

Then, by summing Eq. (6) over $t = 1, \dots, T-1$, we get

$$\|w_T^i - w_T\| \leq \frac{\eta c}{n} \sum_{t=1}^{T-1} f(w_t, z_i)^{1-\delta} \leq \frac{\eta c}{n} \sum_{t=1}^T f(w_t, z_i)^{1-\delta}.$$

As a result,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|w_T^i - w_T\|^2 &\leq \frac{c^2 \eta^2}{n^3} \sum_{i=1}^n \left(\sum_{t=1}^T f(w_t, z_i)^{1-\delta} \right)^2 \\ &= \frac{c^2 \eta^2 T^2}{n^3} \sum_{i=1}^n \left(\frac{1}{T} \sum_{t=1}^T f(w_t, z_i)^{1-\delta} \right)^2 \\ &\leq \frac{c^2 \eta^2 T^{2\delta}}{n^3} \sum_{i=1}^n \left(\sum_{t=1}^T f(w_t, z_i) \right)^{2(1-\delta)} && \text{(Concavity of } x \rightarrow x^{1-\delta}\text{)} \\ &\leq \frac{c^2 \eta^2 T^{2\delta}}{n^3} \left(\sum_{t=1}^T \sum_{i=1}^n f(w_t, z_i) \right)^{2(1-\delta)} && (\delta \leq \frac{1}{2}, f(w_t, z_i) \geq 0) \\ &= \frac{c^2 \eta^2 T^{2\delta}}{n^{1+2\delta}} \left(\sum_{t=1}^T \widehat{F}(w_t) \right)^{2(1-\delta)}. \end{aligned}$$

■

3.3. Stochastic Gradient Descent

We now turn to establish a stability bound for Stochastic Gradient Descent (SGD) under Assumption 2. Given a dataset S of size n , SGD is initialized at a point $w_1 \in \mathbb{R}^d$ and at each step $t = 1, \dots, T$, samples randomly an index $i_t \in [n]$ and performs an update

$$w_{t+1} = w_t - \eta \nabla f(w_t, z_{i_t}), \quad (7)$$

where $\eta > 0$ is the step size of the algorithm. We consider a standard variant of SGD that returns the average iterate, namely $\bar{w}_T = \frac{1}{T} \sum_{t=1}^T w_t$.

We discuss the performance of SGD that runs for T iterations on data set with size n , and show ℓ_1 -on-average-loo model stability bound for the algorithm. The stability bound that we get is identical to the ℓ_1 -on-average-loo model stability bound of GD and stated in the following theorem (proof deferred to the full version of the paper (Schliserman and Koren, 2022)).

Theorem 11. *Suppose Assumption 2 holds. If, for every z , $f(w, z)$ is positive, L -smooth, convex with respect to w , then, SGD with step size $\eta \leq 2/L$ is ℓ_1 -on-average-loo model stable with*

$$\varepsilon_{\text{stab}}^{(1)} \leq \frac{\eta T^\delta c}{n} \left(\sum_{t=1}^T \widehat{F}(w_t) \right)^{1-\delta}.$$

4. Generalization loss bounds

We now establish new generalization bounds for gradient methods on self-bounded objectives, using the stability bounds developed in the previous section.

4.1. Gradient Descent

We begin with GD and show the following generalization bound:

Theorem 12. *Suppose that for any z , the loss function $f(w, z)$ is nonnegative, convex, L -smooth with respect to w , and further satisfies Assumptions 1 and 2. Then the output of GD with step size $\eta \leq 1/2L$ initialized at $w_1 = 0$ has, for any ε such that $0 < \varepsilon/\rho(\varepsilon)^2 \leq 1/\eta T$,*

$$\mathbb{E}[F(w_T)] = O\left(\frac{\rho(\varepsilon)^2}{\eta T} + \frac{L\rho(\varepsilon)^{4(1-\delta)}c^2\eta^{2\delta}T^{2\delta}}{n^{1+2\delta}}\right).$$

If in addition f is G -Lipschitz, the output w_T also satisfies

$$\mathbb{E}[F(w_T)] = O\left(\frac{\rho(\varepsilon)^2}{\eta T} + \frac{\rho(\varepsilon)^{2(1-\delta)}\eta^\delta T^\delta Gc}{n}\right).$$

We remark that the condition on ε is not very restrictive; for example, $\varepsilon = 1/\eta T$ is always a valid choice and gives a nontrivial convergence bound. (In applications of the bound, we will sometimes use a better choice of ε that satisfies this condition.)

To obtain this result, we first bound the training error of GD in the smooth and convex regime, under Assumption 1. The proof is standard and based on basic techniques in convex optimization.

Lemma 13. *Under Assumptions 1 and 2, if for every z $f(w, z)$ is L -smooth, convex and positive, let $\{w_t\}_t$ be produced by the GD update rule (Eq. (4)) with $\eta \leq 1/2L$ on S and $w_1 = 0$. It holds for any $\varepsilon > 0$ that*

$$\widehat{F}(w_T) \leq \frac{1}{T} \sum_{t=1}^T \widehat{F}(w_t) \leq \frac{2\rho(\varepsilon)^2}{\eta T} + 2\varepsilon.$$

Now we turn to proving the generalization bound for gradient methods on self bounded objectives, as stated in Theorem 12. Here we only show the bound in the Lipschitz regime. The full proof is implied by using Lemma 7 and Theorem 10 and appears in the full version of the paper (Schliserman and Koren, 2022).

Proof of Theorem 12. First, for any z , the loss function $f(w, z)$ is nonnegative, convex, L -smooth and G -Lipschitz with respect to w . Moreover, that f satisfies Assumptions 1 and 2. Then, we know by Lemma 6 that

$$\mathbb{E}[F(w_t)] \leq \mathbb{E}[\widehat{F}(w_t)] + \mathbb{E}\left[\frac{2L}{n} \sum_{i=1}^n \|w_t - w_t^i\|\right].$$

By Lemma 13 and Theorem 8 we get,

$$\mathbb{E}[F(w_t)] \leq \left(\frac{2\rho(\varepsilon)^2}{\eta T} + 2\varepsilon\right) + \frac{2\eta T^\delta Gc}{n} \left(\frac{2\rho(\varepsilon)^2}{\eta} + 2\varepsilon T\right)^{1-\delta}.$$

Finally, if $\varepsilon \leq \rho(\varepsilon)^2/\eta T$ we get,

$$\begin{aligned}\mathbb{E}[F(w_t)] &\leq \left(\frac{2\rho(\varepsilon)^2}{\eta T}\right) + \frac{2\eta T^\delta Gc}{n} \left(\frac{2\rho(\varepsilon)^2}{\eta}\right)^{1-\delta} \\ &= O\left(\frac{\rho(\varepsilon)^2}{\eta T} + \frac{\rho(\varepsilon)^{2(1-\delta)}\eta^\delta T^\delta Gc}{n}\right).\end{aligned}$$

■

4.2. Stochastic Gradient Descent

In this section we state and show a generalization bound for SGD on self-bounded losses.

Theorem 14. *Suppose that for any z , the loss function $f(w, z)$ is nonnegative, convex, G -Lipschitz and L -smooth with respect to w . Further assume that f satisfies Assumptions 1 and 2. Then the output of SGD with step size $\eta \leq 1/2L$ initialized at $w_1 = 0$ has, for any ε such that $0 < \varepsilon/\rho(\varepsilon)^2 \leq 1/\eta T$,*

$$\mathbb{E}[F(w_T)] = O\left(\frac{\rho(\varepsilon)^2}{\eta T} + \frac{\rho(\varepsilon)^{2(1-\delta)}\eta^\delta T^\delta Gc}{n}\right),$$

when the expectation is on the randomness of the algorithm and on the data examples.

As in the case of GD, we use the stability bound to get a generalization error bound. In the same manner, we begin in bounding the optimization error of SGD (see Lemma 23 in the full version of the paper (Schliserman and Koren, 2022)). then, we prove generalization in a similar way as in GD, except using Lemma 23 and Theorem 11 instead of Lemma 13 and Theorem 8. The full proof appears in the full version of the paper (Schliserman and Koren, 2022).

5. Applications to separable linear classification

As detailed in Section 2, the setting of linear classification with separable data (with margin $\gamma > 0$) is captured by our general framework. In this section, we demonstrate how to apply our general generalization bounds for gradient methods in this setting with several popular choices of loss functions. Most of our application are of binary classification, where $f_i(w) = \ell(w \cdot z_i)$ for every i , where $\ell : \mathbb{R} \rightarrow \mathbb{R}^+$ is a nonnegative loss function and $z_i = y_i \cdot x_i$ for the labeled examples (x_i, y_i) . Here we provide the derivation of the bounds for the logistic loss and for polynomially-tailed losses. Details for the remaining bounds can be found in the full version of the paper (Schliserman and Koren, 2022)

5.1. Logistic loss

We start with functions with exponential tails. The first loss function that we consider is the logistic loss, $\ell(y) = \log(1 + e^{-y})$. This function is convex, 1-Lipschitz, 1-smooth and $(1, 0)$ -self-bounded (see Lemma 25 in the full version of the paper (Schliserman and Koren, 2022)). Moreover, by Lemma 4, Assumption 1 holds with $\rho(\varepsilon) = (1/\gamma) \log(1/\varepsilon)$. For this function, we can choose $\varepsilon = 1/T$ and obtain the following generalization bound:

Corollary 15. *If ℓ is the logistic loss, then for gradient descent on \widehat{F} with step size $\eta = \frac{1}{2}$ and $w_1 = 0$:*

$$\mathbb{E}[F(w_T)] = O\left(\frac{\log^2 T}{\gamma^2 T} + \frac{\log^2 T}{\gamma^2 n}\right).$$

The implied generalization bound over the zero-one accuracy of the model matches, up to $\log T$ factors, the bounds obtained by Shamir (2021) for the normalized predictor.

5.2. Polynomially-tailed losses

Now we turn to discuss loss functions with polynomially-decaying tails. On such functions, Assumption 2 holds for $\delta = 0$. For concreteness, we will focus on the following loss function, but our arguments hold more generally for any Lipschitz loss with a similar tail decay rate:

$$\ell(y) = \begin{cases} (1+y)^{-\alpha} & y \geq 0; \\ 1 - \alpha y & y < 0, \end{cases} \quad (8)$$

for $\alpha > 0$, which is convex, α -Lipschitz, $\alpha(\alpha+1)$ -smooth, and $(\alpha, 0)$ -self-bounded (see Lemma 27 in the full version of the paper (Schliserman and Koren, 2022)). In addition, by Lemma 4, Assumption 1 holds for $\rho(\varepsilon) = (1/\gamma)\varepsilon^{-1/\alpha}$. For a suitable choice of ε we can obtain the following generalization bound:

Corollary 16. *If ℓ is the the function defined in Eq. (8), the output of gradient descent on \widehat{F} with step size $\eta = \frac{1}{2L} = \frac{1}{2\alpha(\alpha+1)}$ and $w_1 = 0$ holds,*

$$\mathbb{E}[F(w_T)] = O\left(\left(\frac{\alpha}{\gamma}\right)^{\frac{2\alpha}{2+\alpha}} \left(\frac{1}{T^{\frac{\alpha}{2+\alpha}}} + \frac{T^{\frac{2}{\alpha+2}}}{n}\right)\right).$$

We see that as long as $\alpha \ll T$, as the degree of the polynomial gets higher, gradient methods will optimize the test loss faster. Also note that this bound is optimized for $T = n$, for any degree $\alpha > 0$. Finally, we remark that the rate we established in Corollary 16 is essentially the best one could hope for, as it matches the optimal (training) optimization rate of GD on polynomially-tailed functions.

Lemma 17. *Let $\alpha \geq 1$. There exists a function $f : \mathbb{R} \rightarrow \mathbb{R}^+$ that is convex, α -Lipschitz, $\alpha(\alpha + 1)$ -smooth and $(c, 0)$ -self-bounding over \mathbb{R}^d with $\rho(\varepsilon) = \varepsilon^{-1/\alpha}$ and for every $t \geq 1$, the iterate w_t of gradient descent with $\eta = 1/\alpha(\alpha + 1)$ and $w_1 = 0$ has $f(w_t) = \Omega(t^{-\frac{\alpha}{\alpha+2}})$.*

Acknowledgments

This work has received support from the Israeli Science Foundation (ISF) grant no. 2549/19, the Len Blavatnik and the Blavatnik Family Foundation, the Deutsch Foundation and the Yandex Initiative in Machine Learning.

References

Amit Attia and Tomer Koren. Algorithmic instabilities of accelerated gradient descent. *Advances in Neural Information Processing Systems*, 34, 2021.

- Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
- Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.
- Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.
- Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.
- Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *Journal of Environmental Sciences (China) English Ed*, 2019.
- Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2109–2136. PMLR, 09–12 Jul 2020.
- Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.
- Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019a.
- Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019b.
- Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.
- Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.
- Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.
- Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. *arXiv preprint arXiv:2202.13441*, 2022.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, page 5, 2009.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.