Stability and Generalization of Learning Algorithms that Converge to Global Optima

Zachary Charles [1] **Dimitris Papailiopoulos** [1]

Abstract

We establish novel generalization bounds for learning algorithms that converge to global minima. We derive black-box stability results that only depend on the convergence of a learning algorithm and the geometry around the minimizers of the empirical risk function. The results are shown for non-convex loss functions satisfying the Polyak-Łojasiewicz (PL) and the quadratic growth (QG) conditions, which we show arise for 1-layer neural networks with leaky ReLU activations and deep neural networks with linear activations. We use our results to establish the stability of first-order methods such as stochastic gradient descent (SGD), gradient descent (GD), randomized coordinate descent (RCD), and the stochastic variance reduced gradient method (SVRG), in both the PL and the strongly convex setting. Our results match or improve state-of-the-art generalization bounds and can easily extend to similar optimization algorithms. Finally, although our results imply comparable stability for SGD and GD in the PL setting, we show that there exist simple quadratic models with multiple local minima where SGD is stable but GD is not.

1. Introduction

The recent success of training complex models at state-of-the-art accuracy in many common machine learning tasks has sparked significant interest and research in algorithmic machine learning. In practice, not only can these complex deep neural models yield zero training loss, they can also generalize surprisingly well (Zhang et al., 2016; Lin & Tegmark, 2016). Although there has been significant recent work in analyzing the generalization capacity of various learning algorithms, our theoretical understanding of their generalization properties falls far below what has been observed empirically.

A useful proxy for analyzing the generalization performance of learning algorithms is that of *stability*. A training algorithm is stable if small changes in the training set result in small differences in the output predictions of the trained model. In their foundational work, Bousquet & Elisseeff (2002) establish that *stability begets generalization*.

While there has been stability analysis for empirical risk minimizers (Bousquet & Elisseeff, 2002; Mukherjee et al., 2006), there are far fewer results for commonly used iterative learning algorithms. In a recent novel work, Hardt et al. (2016) establish stability bounds for SGD, and discuss algorithmic heuristics that provably increase the stability of SGD models. Unfortunately, showing non-trivial stability for more involved algorithms like SVRG (Johnson & Zhang, 2013) (even in the convex case), or SGD in more nuanced non-convex setups is not straightforward. While Hardt et al. (2016) provide an elegant analysis that shows stability of SGD for non-convex loss functions, the result requires very small step-sizes. The step-size is small enough that under standard smoothness assumptions (Ghadimi & Lan, 2013), approximate convergence may require an exponential number of steps (see Appendix section B.3). Generally, there seems to be a trade-off between convergence and stability of algorithms. In this work we show that under certain geometric assumptions on the loss function around global minima, we can actually leverage the convergence properties of an algorithm to prove that it is stable.

The goal of this work is to provide black-box and easy-to-use stability results for a variety of learning algorithms in non-convex settings. We show that this is in some cases possible by decoupling the stability of global minima and their proximity to models trained by learning algorithms.

Our Contributions: We establish that models trained by algorithms that converge to global minima are stable under the Polyak-Łojasiewicz (PL) and the quadratic growth (QG) conditions (Karimi et al., 2016). Informally, these conditions assert that the suboptimality of a model is upper

[1] Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Wisconsin, USA. Correspondence to: Zachary Charles <zcharles@wisc.edu>, Dimitris Papailiopoulos <dimitris@ece.wisc.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

bounded by the norm of its gradient and lower bounded by its distance to the closest global minimizer. These are weaker conditions than considered in previous work, but still match several known stability bounds. For example, in (Hardt et al., 2016) the authors require convexity or strong convexity. Gonen & Shalev-Shwartz (2017) prove the stability of ERMs for non-convex but locally strongly convex loss functions obeying strict saddle inequalities. By contrast, we develop comparable stability results for a large class of non-convex functions. Although (Hardt et al., 2016) establishes the stability of SGD for smooth non-convex objectives, the step size selection can be prohibitively small for convergence. Our bounds make no assumptions on the hyper-parameters of the algorithm.

We use our black-box results to directly compare the generalization performance of popular first-order methods in general learning setups. While direct proofs of stability may require novel algorithm-specific analysis, our results are derived from known convergence rates of popular algorithms. For strong convexity (a special case of the PL condition), we recover order-wise the stability bounds of Hardt et al. (2016), but for a larger family of optimization algorithms (*e.g.,* SGD, GD, SVRG, etc). We show that many of these algorithms offer order-wise similar stability as saddle-point avoiding algorithms in non-convex problems where all local minima are global (Gonen & Shalev-Shwartz, 2017).

We give examples of some machine learning scenarios where the PL condition mentioned above holds true. Adapting techniques from (Hardt & Ma, 2016), we show that deep networks with linear activation functions are PL almost everywhere in the parameter space. Our theory allows us to derive results similar to those in (Kawaguchi, 2016) about local/global minimizers in linear neural networks and reformulate them in terms fo the PL condition. We also show that 1-layer neural networks with leaky ReLU activations satisfy the PL condition.

Finally, we show that while SGD and GD have analogous stability in the convex setting, this breaks down in the non-convex setting. We give an explicit example of a simple 1-layer neural network on which SGD is stable but GD is not. Such an example was theorized in (Hardt et al., 2016) (see Figure 10 in that paper); here we formalize the authors' intuition. Our results offer yet another indication that models trained via SGD generalize better than those trained by full-batch GD.

Prior Work: The idea of stability analysis has been around for more than 30 years since Devroye & Wagner (1979). Bousquet & Elisseeff (2002) defined several notions of algorithmic stability and used them to derive bounds on generalization error. Further work has focused on stability of randomized algorithms (Elisseeff et al., 2005) and the interplay between uniform convergence and generalization (Shalev-Shwartz et al., 2010). Mukherjee et al. (2006) show that stability implies consistency of empirical risk minimization. Shalev-Shwartz et al. (2010) show that stability can also imply learnability in some problems.

Hardt et al. (2016) establish stability bounds for stochastic gradient descent (SGD) in the convex, strongly convex, and non-convex case. Work by Lin et al. (2016) shows that stability of SGD can be controlled by forms of regularization. In (Kuzborskij & Lampert, 2017), the authors give stability bounds for SGD that are data-dependent. These bounds are smaller than those in (Hardt et al., 2016), but require assumptions on the underlying data. Liu et al. give a related notion of *uniform hypothesis stability* and show that it implies guarantees on the generalization error (Liu et al., 2017).

Stability is closely related to the notion of *differential privacy* (Dwork, 2006). Roughly speaking, differential privacy ensures that the probability of observing any outcome from a statistical query changes if you modify any single dataset element. It was later shown that differentially private algorithms generalize well (Dwork et al., 2015; Nissim & Stemmer, 2015).

2. Preliminaries

We first introduce some notation we will use in this paper. For $w \in \mathbb{R}^m$, we will let $\|w\|$ denote the 2-norm of w. For a matrix $A \in \mathbb{R}^{n\times m}$, we will let $\sigma_{\min}(A)$ denote its minimum singular value, and $\|A\|_F$ denote its Frobenius norm.

Let $S = \{z_1, \ldots, z_n\}$ be a set of training data, where $z_i \overset{iid}{\sim} \mathcal{D}$. For a model w and a loss function ℓ, let $\ell(w; z)$ be the error of w on the training example z. We define the *expected risk* of a model w by $R[w] := \mathbb{E}_{z\sim\mathcal{D}}\, \ell(w; z)$. Since, we do not have access to the underlying distribution $\mathcal{D}$ optimizing $R[w]$ directly is not possible. Instead, we will measure the *empirical risk* of a model w on a set S, given by:

$$R_S[w] := \frac{1}{n}\sum_{i=1}^{n} \ell(w; z_i).$$

The generalization of our model is measured by the *generalization gap*, $\epsilon_{\text{gen}}(w) := |R_S[w] - R[w]|$.

For our purposes, w will be the output of some (potentially randomized) learning algorithm $\mathcal{A}$, trained on some data set S. We will denote this output by $\mathcal{A}(S)$. Let $S' = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$, where $z_i' \sim \mathcal{D}$. We then have the following notion of uniform stability that was first introduced in (Bousquet & Elisseeff, 2002).

Definition 1 (Uniform Stability)**.** *An algorithm $\mathcal{A}$ is uniformly ϵ-stable if for all data sets S, S' differing in at most one example,* $\sup_z \mathbb{E}_{\mathcal{A}}\left[\ell(\mathcal{A}(S); z) - \ell(\mathcal{A}(S'); z)\right] \leq \epsilon$.

The expectation is taken with respect to the randomness of the algorithm $\mathcal{A}$. Bousquet and Elisseeff establish that uniform stability implies small generalization gap (Bousquet & Elisseeff, 2002).

Theorem 1. *Suppose $\mathcal{A}$ is uniformly ϵ-stable. Then* $|\mathbb{E}_{S,\mathcal{A}}\left[R_S[\mathcal{A}(S)] - R[\mathcal{A}(S)]\right]| \leq \epsilon$.

In practice, uniform stability may be too restrictive, since the bound above must hold for all z, irrespective of its marginal distribution. The following notion of stability, while weaker, is still enough to control the generalization gap. Given a data set $S = \{z_1, \ldots, z_n\}$ and $i \in \{1, \ldots, n\}$, we define S^i as $S \backslash z_i$.

Definition 2 (Pointwise Hypothesis Stability, Bousquet & Elisseeff (2002))**.** *$\mathcal{A}$ has pointwise hypothesis stability β with respect to a loss function ℓ if $\forall i \in \{1, \ldots, n\}$,* $\mathbb{E}_S[|\ell(\mathcal{A}(S); z_i) - \ell(\mathcal{A}(S^i); z_i)|] \leq \beta$.

Pointwise hypothesis stability is a weaker notion than uniform stability, but can still be used to establish non-trivial generalization bounds.

Theorem 2 (Elisseeff et al. (2005))**.** *Suppose $\mathcal{A}$ has pointwise hypothesis stability β with respect to ℓ where $0 \leq \ell(w; z) \leq M$. For any δ, with probability at least $1 - \delta$,*

$$R[\mathcal{A}(S)] \leq R_S[\mathcal{A}(S)] + \sqrt{\frac{M^2 + 12Mn\beta}{2n\delta}}.$$

While this result was initially proved only for non-random algorithms, (Elisseeff et al., 2005) later extended this type of argument to random algorithms using similar notions of stability.

In the following, we derive stability bounds for models trained on empirical risk functions satisfying the PL and QG conditions. To do so, we will assume that the functions in question are L-Lipschitz.

Definition 3. *A function $f : \Omega \to \mathbb{R}$ is L-Lipschitz if for all $x_1, x_2 \in \Omega$, $|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|$.*

If f is assumed to be differentiable, this is equivalent to saying that for all x, $\|\nabla f(x)\| \leq L$.

In recent work, Karimi et al. (2016) used the *Polyak-Łojasiewicz condition* to prove simplified nearly-optimal convergence rates for several first-order methods. Notably, there are some non-convex functions that satisfy the PL condition. The condition is defined below.

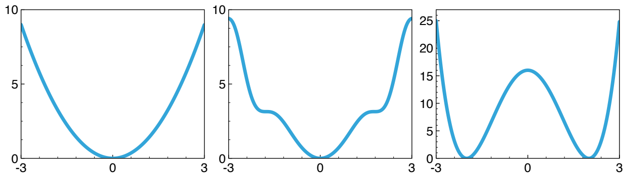
Definition 4 (Polyak-Łojasiewicz)**.** *Fix a set $\mathcal{X}$ and let $f^* = \min_{x \in \mathcal{X}} f(x)$. Then f satisfies the Polyak-Łojasiewicz (PL) condition with parameter $\mu > 0$ on $\mathcal{X}$ if for all $x \in \mathcal{X}$, $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$.*

We often refer to such functions as μ-PL. Note that for PL functions, every critical point is a global minimizer. While strong convexity implies PL, the reverse is not true. Moreover, PL functions are in general non-convex (*e.g.,* invex functions). We also consider a strictly larger family of functions that satisfy the quadratic growth condition.

Definition 5 (Quadratic Growth)**.** *A function f satisfies the quadratic growth (QG) condition on $\mathcal{X}$ with parameter $\mu > 0$ if for all $x \in \mathcal{X}$, $f(x) - f^* \geq \frac{\mu}{2}\|x - x_p\|^2$, where x_p denotes the euclidean projection of x onto the set of global minimizers of f in $\mathcal{X}$ (i.e., x_p is the closest point to x in $\mathcal{X}$ satisfying $f(x_p) = f^*$).*

We often refer to such functions as μ-QG. Both of these conditions have been considered in previous studies. The PL condition was first introduced by Polyak in (Lojasiewicz, 1963), who showed that under this assumption, gradient descent converges linearly. The QG condition has been considered under various guises (Bonnans & Ioffe, 1995; Ioffe, 1994) and can imply important properties about the geometry of critical points. For example, Anitescu (2000) showed that local minima of non-linear programs satisfying the QG condition are actually isolated stationary points. These kinds of geometric implications will allow us to derive stability results for large classes of algorithms.

In general, the PL condition retains many properties of strong convexity (such as the fact that under the PL assumption, all critical points are local minima) without requiring convexity. The QG condition further relaxes this, allowing for critical points that are not global minima, while still enforcing that locally, the function grows quadratically away from global minima. Figure 1 gives examples of μ-strongly convex, μ-PL, and μ-QG functions for $\mu = 2$.



(a) x^2 (b) $x^2 + \sin^2(x)$ (c) $(x-2)^2(x+2)^2$

Figure 1. Examples of (a) strongly convex, (b) PL, and (c) QG functions.

3. Stability of Approximate Global Minima

In this section, we establish the stability of large classes of learning algorithms under the PL and QG conditions presented above. Our stability results are "black-box" in the sense that our bounds are decomposed as a sum of two terms: a term concerning the convergence of the algorithm to a global minimizer, and a term relevant to the geometry of the loss function around the global minima. Both terms are used to establish good generalization and provide some insights into the way that learning algorithms perform.

For an algorithm $\mathcal{A}$, let w_S denote its output on S. The empirical training error on a data set S is denoted $f_S(w)$ and defined as $f_S(w) = \frac{1}{|S|}\sum_{z\in S}\ell(w;z)$. We assume that $\ell(\cdot, w)$ is L-Lipschitz w.r.t. w for all z.

3.1. Pointwise Hypothesis Stability for PL/QG Loss Functions

We will show that if f_S satisfies the PL or QG condition, we will be able to quantify the stability of $\mathcal{A}$. Although these conditions may at first seem unnatural, we show in Section 4 that they arise in a large number of machine learning settings, including in certain deep linear neural networks. Let $\mathcal{X}_{\min}$ denote the set of global minima of f_S.

Theorem 3. *Assume that for all S and $w \in \mathcal{X}$, f_S is PL with parameter μ. We assume that applying $\mathcal{A}$ to f_S produces output w_S that is converging to some global minimizer w_S^*. Then $\mathcal{A}$ has pointwise hypothesis stability with parameter ϵ_{stab} satisfying the following conditions.*

Case 1: *If for all S, $\|w_S - w_S^*\| \le \epsilon_{\mathcal{A}}$ then*

$$\epsilon_{stab} \le 2L\epsilon_{\mathcal{A}} + \frac{2L^2}{\mu(n-1)}.$$

Case 2: *If for all S, $|f_S(w_S) - f_S(w_S^*)| \le \epsilon'_{\mathcal{A}}$ then*

$$\epsilon_{stab} \le 2L\sqrt{\frac{2\epsilon'_{\mathcal{A}}}{\mu}} + \frac{2L^2}{\mu(n-1)}.$$

Case 3: *If for all S, $\|\nabla f_S(w_S)\| \le \epsilon''_{\mathcal{A}}$, then*

$$\epsilon_{stab} \le \frac{2L\epsilon''_{\mathcal{A}}}{\mu} + \frac{2L^2}{\mu(n-1)}.$$

Suppose our loss functions are PL and our algorithm $\mathcal{A}$ is an oracle that returns a global optimizer w_S^*. Then the terms $\epsilon_{\mathcal{A}}, \epsilon'_{\mathcal{A}}, \epsilon''_{\mathcal{A}}$ above are all equal to 0, leading to the following corollary.

Corollary 4. *Suppose f_S is PL with parameter μ and let $\mathcal{A}(S) = \arg\min_{w\in\mathcal{X}} f_S(w)$. Then, $\mathcal{A}$ has pointwise hypothesis stability with $\epsilon_{stab} = \frac{2L^2}{\mu(n-1)}$.*

Bousquet & Elisseeff (2002) considered the stability of empirical risk minimizers where the loss function satisfied strong convexity. Their work implies that for λ-strongly convex functions, the empirical risk minimizer has stability satisfying $\epsilon_{\text{stab}} \le \frac{L^2}{\lambda n}$. Since λ-strongly convex implies λ-PL, Corollary 6 generalizes their result, with only a constant factor loss.

Remark 1. *Theorem 3 holds even if we only have information about $\mathcal{A}$ in expectation. For example, if we only know that $\mathbb{E}_{\mathcal{A}}\|w_S - w_S^*\| \le \epsilon_{\mathcal{A}}$, we still establish pointwise hypothesis stability (in expectation with respect to $\mathcal{A}$), with the same constant as above. This allows us to apply our result to randomized algorithms such as SGD where we are interested in the convergence in expectation.*

A similar result to Theorem 3 can be derived for empirical risk functions that satisfy the QG condition and that are *realizable*, that is, where the global minimum of f, denoted f^*, is 0.

Theorem 5. *Suppose that for all S, f_S is QG with parameter μ and $f^* = 0$. Suppose that applying $\mathcal{A}$ to f_S produces output w_S that is converging to some global minimizer w_S^*. Assume that for all w and z, $|\ell(w;z)| \le c$. Then $\mathcal{A}$ has pointwise hypothesis stability with parameter ϵ_{stab} satisfying the following conditions.*

Case 1: *If for all S, $\|w_S - w_S^*\| \le \epsilon_{\mathcal{A}}$ then*

$$\epsilon_{stab} \le 2L\epsilon_{\mathcal{A}} + 2L\sqrt{\frac{c}{\mu n}}.$$

Case 2: *If for all S, $|f_S(w_S) - f_S(w_S^*)| \le \epsilon'_{\mathcal{A}}$ then*

$$\epsilon_{stab} \le 2L\sqrt{\frac{2\epsilon'_{\mathcal{A}}}{\mu}} + 2L\sqrt{\frac{c}{\mu n}}.$$

Remark 2. *Observe that unlike the case of PL empirical losses, QG empirical losses only allow for a $O(\frac{1}{\sqrt{n}})$ convergence rate of stability. Moreover, similarly to our result for PL loss functions, the result of Theorem 5 holds even if we only have information about the convergence of $\mathcal{A}$ in expectation.*

Finally, we can obtain the following Corollary for empirical risk minimizers.

Corollary 6. *Let f_S satisfy the QG inequality with parameter μ and let $\mathcal{A}(S) = \arg\min_{w\in\mathcal{X}} f_S(w)$. Then, $\mathcal{A}$ has pointwise hypothesis stability with $\epsilon_{stab} = 2L\sqrt{\frac{c}{\mu n}}$.*

3.2. Uniform Stability for PL/QG Loss Functions

Under a more restrictive setup, we can obtain similar bounds for uniform hypothesis stability, which is a stronger stability notion compared to its pointwise hypothesis variant. The usefulness of uniform stability compared to pointwise stability, is that it can lead to generalization bounds that concentrate exponentially faster (Bousquet & Elisseeff, 2002) with respect to the sample size n.

As before, given a data set S, we let denote w_S be the model that $\mathcal{A}$ outputs. Let $\pi_S(w)$ denote the closest optimal point of f_S to w. We will denote $\pi_S(w_S)$ by w_S^*. Let S, S' be data sets differing in at most one entry. We will make the following technical assumption:

Assumption 1. *The empirical risk minimizers for f_S and $f_{S'}$, i.e., $w_S^*, w_{S'}^*$ satisfy $\pi_S(w_{S'}^*) = w_S^*$, where $\pi_S(w)$ is the projection of w on the set of empirical risk minimizers of f_S. Note that this is satisfied if for every data set S, there is a unique minimizer w_S^*.*

Remark 3. *The above assumption is relatively strict, and in general does not apply to empirical losses with infinitely many global minima. To tackle the existence of infinitely many global minima, one could imagine designing $\mathcal{A}(S)$ to output a structured empirical risk minimizer, e.g., one such that if $\mathcal{A}$ is applied on S', its projection on the optima of*

f_S would always yield back $A(S)$. This could be possible if $A(S)$ corresponds to minimizing a regularized, or structured cost function whose set of optimizers only contained a small subset of the global minima of f_S. Unfortunately, coming up with such a structured empirical risk minimizer for general non-convex losses seems far from straightforward, and serves as an interesting open problem.

Theorem 7. *Assume that for all S, f_S satisfies the PL condition with parameter μ, and suppose that Assumption 1 holds. Then $\mathcal{A}$ has uniform stability with parameter ϵ_{stab} satisfying the following conditions.*

Case 1: *If for all S, $\|w_S - w_S^*\| \leq \epsilon_{\mathcal{A}}$ then*

$$\epsilon_{stab} \leq 2L\epsilon_{\mathcal{A}} + \frac{2L^2}{\mu n}.$$

Case 2: *If for all S, $|f_S(w_S) - f_S(w_S^*)| \leq \epsilon'_{\mathcal{A}}$ then*

$$\epsilon_{stab} \leq 2L\sqrt{\frac{2\epsilon'_{\mathcal{A}}}{\mu}} + \frac{2L^2}{\mu n}.$$

Case 3: *If for all S, $\|\nabla f_S(w_S)\| \leq \epsilon''_{\mathcal{A}}$, then*

$$\epsilon_{stab} \leq \frac{2L\epsilon''_{\mathcal{A}}}{\mu} + \frac{2L^2}{\mu n}.$$

Since strong convexity is a special case of PL, this theorem implies that if we run enough iterations of a convergent algorithm $\mathcal{A}$ on a λ-strongly convex loss function, then we obtain uniform stability on the order of $\epsilon_{\text{stab}} = L^2/\lambda n$. In particular, this theorem recovers the stability estimates for ERMs and SGD applied to strongly convex functions proved in (Bousquet & Elisseeff, 2002) and (Hardt et al., 2016), respectively.

In order to make this result more generally applicable, we would like to extend the theorem to a larger class of functions than just globally PL functions. If we assume boundedness of the loss function, then we can derive a similar result for globally QG functions. This leads us to the following theorem:

Theorem 8. *Assume that for all S, f_S satisfies the QG condition with parameter μ, moreover let Assumption 1 hold. Suppose that for all z and $w \in \mathcal{X}$, $\ell(w; z) \leq c$. Then $\mathcal{A}$ is uniformly stable with parameter ϵ_{stab} satisfying:*

Case 1: *If for all S, $\|w_S - w_S^*\| \leq \epsilon_{\mathcal{A}}$ then*

$$\epsilon_{stab} \leq 2L\epsilon_{\mathcal{A}} + 2L\sqrt{\frac{c}{\mu n}}.$$

Case 2: *If for all S, $|f_S(w_S) - f_S^*| \leq \epsilon'_{\mathcal{A}}$ then*

$$\epsilon_{stab} \leq 2L\sqrt{\frac{2\epsilon'_{\mathcal{A}}}{\mu}} + 2L\sqrt{\frac{c}{\mu n}}.$$

Remark 4. *By analogous reasoning to that in Remark 1, both Theorem 7 and Theorem 8 hold if you only have information about the output of $\mathcal{A}$ in expectation.*

4. PL Loss Functions in Practice

Strongly Convex Composed with Piecewise-Linear Functions: As the bounds above show, the PL and QG conditions are sufficient for algorithmic stability and therefore imply good generalization. In this section, we show that the PL condition actually arises in some interesting machine learning setups, including least squares minimization, strongly convex functions composed with piecewise linear functions, and neural networks with linear activation functions. A first step towards a characterization of PL loss functions was proved by Karimi et al. (2016), which established that the composition of a strongly-convex function and a linear function results in a loss that satisfies the PL condition.

We wish to generalize this result to piecewise linear activation functions. Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is defined by $\sigma(z) = c_1 z$, for $z > 0$ and $\sigma(z) = c_2 z, for z \leq 0$. Here $c_i > 0$. For a vector $z \in \mathbb{R}^n$, we denote by $\sigma(z) \in \mathbb{R}^n$ the vector whose ith component is $\sigma(z_i)$. Note that this encompasses leaky-ReLU functions. Following similar techniques to those in (Karimi et al., 2016), we get the following result showing that the composition of strongly convex functions with piecewise-linear functions are PL. The proof can be found in Appendix A.6.

Theorem 9. *Let g be strongly-convex with parameter λ, σ a leaky ReLU activation function with slopes $c1$ and c_2, and X a matrix with minimum singular value $\sigma_{\min}(X)$. Let $c = \min\{|c_1|, |c_2|\}$. Then $f(w) = g(\sigma(Xw))$ is PL almost everywhere with parameter $\mu = \lambda\sigma_{\min}(X)^2c^2$.*

In particular, 1-layer neural networks with a squared error loss and leaky ReLU activations satisfy the PL condition. More generally, this holds for any piecewise-linear activation function with slopes $\{c_i\}_{i=1}^k$. As long as each slope is non-zero and X is full rank, the result above shows that the PL condition is satisfied. This result is closely related to that of (Brutzkus et al., 2017), which shows that SGD converges to global minimizers on such networks when the data is linearly separable.

Linear Neural Networks: The results above only concern one layer neural networks. Given the prevalence of deep networks, we would like to say something about the associated loss function. As it turns out, we can prove that a PL inequality holds in large regions of the parameter space for deep linear networks. Such networks have recently become popular as objects of analysis due to the non-convexity of their landscape. In many cases, all critical points are global minimizers (Kawaguchi, 2016). We will show that a similar theorem can be derived through the lens of the PL condition.

Say we are given a training set $S = \{z_1, \ldots, z_n\}$ where $z_i = (x_i, y_i)$ for $x_i, y_i \in \mathbb{R}^d$. Our neural network will have ℓ fully-connected non-input layers, each with d neurons and linear activation functions. We will parametrize the neural network model via $W_1, \ldots, W_\ell$, where each $W_i \in \mathbb{R}^{d\times d}$. That is, the output at the first non-input layer is

$u_1 = W_1 x$ and the output at layer $k \geq 2$ is $A_k u_{k-1}$. Letting $X, Y \in \mathbb{R}^{d \times N}$ be the matrices with x_i, y_i as their columns (respectively), we can then write our loss function as

$$f(W) = \frac{1}{2}\|W_\ell W_{\ell-1} \ldots W_1 X - Y\|_F^2.$$

Let $W = W_\ell W_{\ell-1} \ldots W_1$. The optimal value of W is $W^* = YX^+$. Here, $X^+ = X^T(XX^T)^{-1}$ is the pseudoinverse of X. We assume that $X \in \mathbb{R}^{d \times N}$ has rank d so that XX^T is invertible. We will also make use of the following lemma which we prove in Appendix A.7.

Lemma 10. *Let $W \in \mathbb{R}^{d \times d}$ be some weight matrix. Then for $C = \|(XX^T)^{-1}X\|_F^2$, we have*

$$C\|(WX - Y)X^T\|_F^2 \geq \|WX - Y\|_F^2 - \|YX^+X - Y\|_F^2.$$

For a matrix A, let $\sigma_{\min}(A)$ denote the smallest singular value of A. For a given $W_1, \ldots, W_\ell$, let $W = W_\ell W_{\ell-1} \ldots W_1$. In Appendix A.8, we prove the following lemma.

Lemma 11. *Suppose that the W_i satisfy $\sigma_{\min}(W_i) \geq \tau > 0$ for all i. Then,*

$$\|\nabla f(W_1, \ldots, W_\ell)\|_F^2 \geq \ell \tau^{2\ell-2}\|(WX - Y)X^T\|_F^2.$$

Combining Lemmas 10 and 11, we derive the following interesting corollary about when critical points are global minimizers. This result is not directly related to the work above, but gives an easy way to understand the landscape of critical points of deep linear networks.

Theorem 12. *Let $(W_1, \ldots, W_\ell)$ be a critical point such that each W_i has full rank. Then $(W_1, \ldots, W_\ell)$ is a global minimizer of f.*

Thematically similar results have been derived previously for 1 layer networks in (Xie et al., 2017) and for deep neural networks in (Zhou & Feng, 2017). In (Kawaguchi, 2016), Kawaguchi derives a similar result to ours for deep linear neural networks, showing that every critical point is either a global minima or a saddle point. Our result, by contrast, implies that all full-rank critical points are global minima.

Lemmas 10 and 11 can also be combined to show that linear networks satisfy the PL condition in large regions of parameter space, as the following theorem states.

Theorem 13. *Suppose our weight matrices $(W_1, \ldots, W_\ell)$ satisfy $\sigma_{\min}(W_i) \geq \tau$ for $\tau > 0$. Then $f(W_1, \ldots, W_\ell)$ satisfies the following PL inequality:*

$$\frac{1}{2}\|\nabla f\|_F^2 \geq \frac{\ell \tau^{2\ell-2}}{\|(XX^T)^{-1}X\|_F^2}(f(W_1, \ldots, W_\ell) - f^*).$$

5. Stability of Some First-order Methods

We wish to apply our bounds from the previous section to popular convergent gradient-based methods. We consider SGD, GD, RCD, and SVRG. When we have L-Lipschitz, μ-PL loss functions f_S and n training examples, Theorem 7 states that any learning algorithm $\mathcal{A}$ has uniform stability ϵ_{stab} satisfying $\epsilon_{\text{stab}} \leq O(L\sqrt{\epsilon_\mathcal{A}/\mu}) + O(L^2/\mu n)$.

Here, $\epsilon_\mathcal{A}$ refers to how quickly $\mathcal{A}$ converges to the optimal value of the loss function. Specifically, this holds if the algorithm produces a model w_S satisfying $|f_S(w_s) - f_S^*| \leq O(\epsilon_\mathcal{A})$.

The convergence rates of SGD, GD, RCD, and SVRG have been studied extensively both for strongly convex functions (Bubeck et al., 2015; Johnson & Zhang, 2013; Nesterov, 2012; 2013) and PL functions (Karimi et al., 2016). These results are summarized in Table 1. When necessary to state the result, we assume a constant step-size of γ. Note that the same convergence rates apply to both λ-strongly convex and λ-PL functions.

Table 1. Convergence rates for T iterations of various gradient-based algorithms with step size γ applied to both λ-SC and λ-PL functions.

ALGORITHM	CONVERGENCE RATE
SGD	$(1 - 2\gamma\lambda)^T + \frac{\gamma L^2}{2\lambda}$
GD	$\left(1 - \frac{\lambda}{L}\right)^T$
RCD	$\left(1 - \frac{\lambda}{dL}\right)^T$
SVRG	$\left(\frac{1}{\lambda\gamma(1-2L\gamma)m} + \frac{2L\gamma}{1-2L\gamma}\right)^T$

We wish to perform enough iterations of our algorithm guarantee to guarantee the same stability as SGD in the strongly convex case. We need to determine how many iterations T we need to perform such that $\epsilon_\mathcal{A} = O(L^2/\mu n)$, as then Corollary 6 implies that our algorithm is uniformly stable with parameter $\epsilon_{\text{stab}} = O(L^2/\mu n)$. The number of iterations needed are summarized in Table 2.

In these settings, the above algorithms all exhibit the same stability for these values of T, despite the potential non-convexity. This is not the case for general non-convex functions. Several studies have observed that small-batch SGD offers superior generalization performance compared to large-batch SGD, or full-batch GD, when training deep neural networks (Keskar et al., 2016).

Unfortunately our bounds are not nuanced enough to capture the difference in generalization performance between mini-batch and large-batch SGD. In Section 6, we will make this observation formal. Although SGD and GD can be equally

Table 2. The number of iterations T that achieves stability $\epsilon_{\text{stab}} = O(L^2/\mu n)$ for various gradient-based algorithms with step size γ in the λ-SC and λ-PL settings.

ALGORITHM	NUMBER OF ITERATIONS
SGD	$\frac{Ln}{\lambda}$
GD	$\frac{\log(L/\lambda n)}{\log(1-\lambda/L)}$
RCD	$\frac{\log(L/\lambda n)}{\log(1-\lambda/dL)}$
SVRG	$\frac{\log(L/\lambda n)}{\log\left((\lambda\gamma(1-2L\gamma)m)^{-1}+2L\gamma(1-2L\gamma)^{-1}\right)}$

stable for non-convex problems satisfying the PL condition, there exist non-convex problems where full-batch GD is not stable and SGD is stable.

6. The Instability of Gradient Descent

In (Hardt et al., 2016), the authors proved bounds on the uniform stability of SGD. They also noted that GD does not appear to be provably as stable for the non-convex case and sketched a situation in which this difference would appear. Due to the similarity of SGD and GD, one may expect similar uniform stability. While this is true in the convex setting (as we show in the Appendix, subsection B.1), this breaks down in the non-convex setting. Below we construct an explicit example where GD is not uniformly stable, but SGD is. This example formalizes the intuition given in (Hardt et al., 2016).

For $x, w \in \mathbb{R}^m$ and $y \in \mathbb{R}$, we let $\ell(w; (x, y)) = (\langle w, x\rangle^2 + \langle w, x\rangle - y)^2$. Intuitively, this is a generalized quadratic model where the predicted label $\hat{y}$ for a given x is given by $\hat{y} = \langle w, x\rangle^2 + \langle w, x\rangle$.

For almost every x, y, the function $\ell(w; (x, y))$ is non-convex as it is a quartic polynomial in the weight vector w. We will use this function to construct data sets S, S' that differ in only one entry, for which GD produces significantly different models. We consider this loss function for all $z = (x, y)$ with $\|z\| \leq C$ for C sufficiently large. When $m = 1$, the loss function simplifies to $\ell(w; (x, y)) = (w^2x^2 + wx - y)^2$.

Define $\alpha := (-1, 1), \beta := (\frac{-1}{2}, 1)$. The graphs of $\ell(w; z)$ at α and β are given in Figure 2. We also graph $g(w)$, defined by $g(w) = \frac{1}{2}(\ell(w; \alpha) + \ell(w; \beta))$.

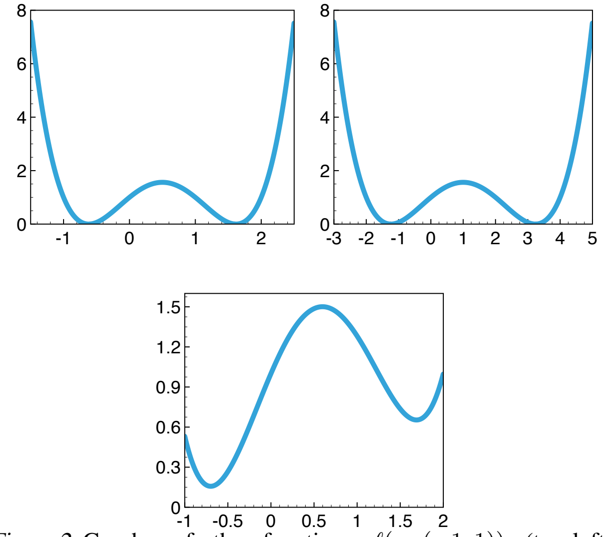


Figure 2. Graphs of the functions $\ell(w; (-1, 1))$ (top-left), $\ell(w; (\frac{-1}{2}, 1))$ (top-right), and $g(w) = \frac{1}{2}[\ell(w; (-1, 1)) + \ell(w; (\frac{-1}{2}, 1))]$ (bottom).

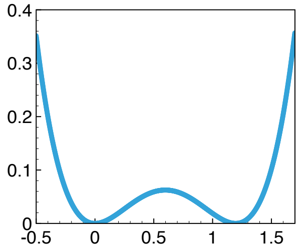


Figure 3. Graph of the function $\ell(w; (-\frac{1}{2\hat{w}}), 0)$. By construction, this function has critical points at $w = 0, \hat{w}, 2\hat{w}$.

Note that the last function has two distinct basins of different heights. Taking the gradient, one can show that the rightmost function in Figure 2 has zero slope at $\hat{w} \approx 0.598004$. Comparing $\ell(w; \alpha)$ and $\ell(w; \beta)$, we see that the sign of their slopes agrees on $(-\frac{1}{2}, \frac{1}{2})$ and on $(1, \frac{3}{2})$. The slopes are of different sign in the interval $[\frac{1}{2}, 1]$. We will use this to our advantage in showing that gradient descent is not stable, while that SGD is.

We will construct points (x_1, y_1) and (x_2, y_2) such that $\ell(w; (x_1, y_1))$ and $\ell(w; (x_2, y_2))$ have positive and negative slope at $\hat{w}$. To do so, we will first construct an example with a slope of zero at $\hat{w}$. A straightforward computation shows that $\ell(w; (\frac{-1}{2\hat{w}}, 0))$ has slope zero at $\hat{w}$. A graph of this loss function is given in Figure 3. Note that $\hat{w}$ corresponds to the concave-down critical point in between the two global minima.

Define $z_\pm = \left(\frac{-1}{2(\hat{w}\pm\epsilon)}, 0\right)$. Straightforward calculations show that $\ell(w; (z_+, 0))$ will have positive slope for $w \in (0, \hat{w} + \epsilon)$, while $\ell(w; (z_-, 0))$ will have negative slope for $w \in (\hat{w} - \epsilon, 2(\hat{w} - \epsilon))$. In particular, their slopes have opposite signs in the interval $(\hat{w} - \epsilon, \hat{w} + \epsilon)$. Define $S = \{z_1, \ldots, z_{n-1}, z_-\}, S' = \{z_1, \ldots, z_{n-1}, z_+\}$, where $z_i = \alpha$ for $1 \leq i \leq \frac{n-1}{2}$, $z_i = \beta$ for $\frac{n-1}{2} < i \leq n - 1$.

By construction, we have

$$f_S(w) = \frac{n-1}{n} g(w) + \ell\left(w; \left(\frac{-1}{2(\hat{w} + \epsilon)}, 0\right)\right).$$

$$f_{S'}(w) = \frac{n-1}{n} g(w) + \ell\left(w; \left(\frac{-1}{2(\hat{w}-\epsilon)}, 0\right)\right).$$

Then $f_S(w)$, $f_{S'}(w)$ will approximately have the shape of the bottom function in Figure 2 above. However, recall that $\frac{d}{dw}g(w) = 0$ at $w = \hat{w}$. Therefore, there is some δ, with $0 < \delta < \epsilon$, such that for all $w \in (\hat{w} - \delta, \hat{w} + \delta)$, $\frac{d}{dw} f_S(w) < 0 < \frac{d}{dw} f_{S'}(w)$.

Now say that we initialize gradient descent with some step-size $\gamma > 0$ in the interval $(w - \delta, w + \delta)$. The above equation implies that the first step of gradient descent on f_S will produce a step moving to the left, but a step moving to the right for $f_{S'}$. This will hold for all $\gamma > 0$. Moreover, the iterations for f_S will continue to move to towards the left basin, while the iterations for $f_{S'}$ will continue to move to the right basin since $\ell(w; (z_+, 0))$ has positive slope for $w \in (0, \hat{w} + \epsilon)$ while $\ell(w; (z_-, 0))$ has negative slope for $w \in (\hat{w} - \epsilon, 2(\hat{w} - \epsilon)$, and that $g(w)$ has positive slope for $w \in (-\frac{1}{2}, \hat{w})$ and negative slope for $w \in (\hat{w}, \frac{3}{2})$.

After enough iterations of gradient descent on $f_S, f_{S'}$, we will obtain models w_S and $w_{S'}$ that are close to the distinct local minima in the right-most graph in Figure 2. This will hold as long as γ is not too large. To ensure this does not happen, we restrict to $\gamma \leq 1$.

Let $w_1 < w_2$ denote the two local minima of $g(w)$. For $z^* = (-\frac{1}{2}, 1)$, plugging these values into $\ell(w; z^*)$ shows that $|\ell(w_1; z^*) - \ell(w_2; z^*)| > 1$. Since w_S is close to w_1 and $w_{S'}$ is close to w_2, we get the following theorem.

Theorem 14. *For all $n, m \geq 1, \gamma \in (0, 1]$, there are $K \geq 1$, $S, S' \subseteq \mathbb{R}^m$ of size n with $|S \cap S'| = n - 1$, and a non-zero measure $A \subseteq \mathbb{R}^m$ such that if we perform at least K iterations of gradient descent with step-size γ on S and S', starting in A, to get outputs $\mathcal{A}(S), \mathcal{A}(S')$, then there is a z^* such that $|\ell(\mathcal{A}(S); z^*) - \ell(\mathcal{A}(S'); z^*)| \geq \frac{1}{2}$.*

This theorem establishes that there exist simple non-convex settings for which the uniform stability of gradient descent does not decrease with n. In light of the work in (Hardt et al., 2016), where the authors show that for very conservative step-sizes, SGD is stable on non-convex loss function, we might wonder whether SGD is stable in this setting. We show in Section B.2 that SGD is stable in this setting. For simplicity of analysis, we focus on the case where $\gamma = 1$. In this section we prove the following theorem.

Theorem 15. *Suppose that we initialize SGD in $[\hat{w} - \eta, \hat{w} + \eta]$ with a step-size of $\gamma = 1$. Let $\mathcal{A}(S), \mathcal{A}(S')$ denote the output of SGD after k iteration for sufficiently large k. For $\|z\| \leq 2, \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}(S); z) - \ell(\mathcal{A}(S'); z)] \leq O(\frac{1}{n})$.*

This is in contrast to gradient descent, which is unstable in this setting. While (Zhang et al., 2016) suggests that SGD is generally more stable in non-convex settings, proving that this holds remains an open problem.

7. Conclusion

The success of machine learning algorithms in practice is often dictated by their ability to generalize. While recent work has developed great insight into the training error of machine learning algorithms, much less is understood about their generalization error. Most prior work has been algorithm-specific or has made strong assumptions on the loss function that may not hold in practice. By decomposing stability into convergence and the geometry of global minimizers, we are able to derive broader results. These easy-to-use stability results encompass a general class of non-convex functions, some of which appear in machine learning setups. Our bounds establish the stability for SGD, GD, SVRG, and RCD and match prior specialized results. Although our bounds are not nuanced enough to compare the generalization of mini-batch and large-batch SGD, we hope that the generality of our bounds serves as a step towards understanding the generalization performance of practical machine learning algorithms.

There are still many exciting open problems concerning the stability and generalization of machine learning and optimization algorithms. We give a few below.

Stability for non-convex loss functions: While our results establish the stability of learning algorithms in some non-convex scenarios, it is unclear how to extend them directly to more general non-convex loss functions. Due to the wide variety of similar but not identical algorithms in machine learning, it would be particularly interesting to derive black-box results on the stability of learning algorithms for general non-convex loss functions, even when it concerns convergence to approximate local minima. In non-convex functions, local minima are not necessarily global. The question remains, among all local minima of a loss function, which one has the smallest generalization gap? The local minima with the smallest generalization error may not be a global minimizer. Even if we restrict to global minimizers, these may have different generalization errors. Is there a simple geometric characterization of their generalization error?

Generalization of SGD vs GD: We showed above that there are settings in which gradient descent is not uniformly stable, but SGD is. Empirically, SGD leads to small generalization error in neural networks (Zhang et al., 2016). Theoretically, it is unclear how widespread this phenomenon is. Does SGD actually lead to more generalizable models than gradient descent?

References

Anitescu, M. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

Bonnans, J. F. and Ioffe, A. Second-order sufficiency and quadratic growth for nonisolated minima. *Mathematics of Operations Research*, 20(4):801–817, 1995.

Bousquet, O. and Elisseeff, A. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, March 2002. ISSN 1532-4435. doi: 10.1162/153244302760200704. URL http://dx.doi.org/10.1162/153244302760200704.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Devroye, L. and Wagner, T. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, Sep 1979. ISSN 0018-9448. doi: 10.1109/TIT.1979.1056087.

Dwork, C. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052, pp. 1–12, Venice, Italy, July 2006. Springer Verlag. ISBN 3-540-35907-9. URL https://www.microsoft.com/en-us/research/publication/differential-privacy/.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 117–126. ACM, 2015.

Elisseeff, A., Evgeniou, T., and Pontil, M. Stability of randomized learning algorithms. *J. Mach. Learn. Res.*, 6:55–79, December 2005. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1046923.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Gonen, A. and Shalev-Shwartz, S. Fast rates for empirical risk minimization of strict saddle problems. *arXiv preprint, arXiv:1701.04271*, 2017.

Hardt, M. and Ma, T. Identity matters in deep learning. *CoRR*, abs/1611.04231, 2016. URL http://arxiv.org/abs/1611.04231.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1225–1234, 2016. URL http://jmlr.org/proceedings/papers/v48/hardt16.html.

Ioffe, A. On sensitivity analysis of nonlinear programs in banach spaces: the approach via composite unconstrained optimization. *SIAM Journal on Optimization*, 4(1):1–43, 1994.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Karimi, H., Nutini, J., and Schmidt, M. *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*, pp. 795–811. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46128-1. doi: 10.1007/978-3-319-46128-1_50. URL http://dx.doi.org/10.1007/978-3-319-46128-1_50.

Kawaguchi, K. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2016.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint, arXiv:1609.04836*, 2016.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. *arXiv preprint, arXiv:1703.01678*, 2017.

Lin, H. W. and Tegmark, M. Why does deep and cheap learning work so well? *arXiv preprint, arXiv:1608.08225*, 2016.

Lin, J., Camoriano, R., and Rosasco, L. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, 2016.

Liu, T., Lugosi, G., Neu, G., and Tao, D. Algorithmic stability and hypothesis complexity. *arXiv preprint, arXiv:1702.08712*, 2017.

Lojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89, 1963.

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. 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.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

Nissim, K. and Stemmer, U. On the generalization properties of differential privacy. *CoRR, abs/1504.05800*, 2015.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

Xie, B., Liang, Y., and Song, L. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pp. 1216–1224, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint, arXiv:1611.03530*, 2016.

Zhou, P. and Feng, J. The landscape of deep learning algorithms. *arXiv preprint, arXiv:1705.07038*, 2017.

A. Proof of Results

A.1. Properties of PL and QG Functions

We have the following equivalent definition of PL due to Karimi et al. (2016).

Lemma 16 (Error bound, (Karimi et al., 2016))**.** *The PL condition is equivalent to the condition that there is some constant $\mu > 0$ such that for all x, $\|\nabla f(x)\| \geq \mu \|x_p - x\|$.*

Karimi et al. (2016) also show that PL functions also satisfy QG.

Lemma 17 ((Karimi et al., 2016))**.** *The PL condition implies the QG condition.*

A.2. Proof of Theorem 3

Proof. Fix a training set S and $i \in \{1, \dots, n\}$. We will show pointwise hypothesis stability for all S, i instead of for them in expectation. Let w_1 denote the output of $\mathcal{A}$ on S, and let w_2 denote the output of $\mathcal{A}$ on S^i. Let w_1^* denote the critical point of f_S to which w_1 is approaching, and w_2^* denote the critical point of f_{S^i} that w_2 is approaching. We then have,

$$\begin{aligned}&|\ell(w_1; z_i) - \ell(w_2; z_i)| \\ &\leq |\ell(w_1; z_i) - \ell(w_1^*; z_i)| + |\ell(w_1^*; z_i) - \ell(w_2^*; z_i)| \\ &\quad + |\ell(w_2^*; z_i) - \ell(w_2; z_i)|. \qquad (1)\end{aligned}$$

We first wish to bound the first and third terms of (1). The bound depends on the case in Theorem 3.

Case 1: By assumption, $\|w_1 - w_1^*\| \leq \epsilon_{\mathcal{A}}$. Since $\ell(\cdot; z_i)$ is L-Lipschitz, this implies

$$|\ell(w_1; z_i) - \ell(w_1^*; z_i)| \leq L\|w_1 - w_1^*\| \leq L\epsilon_{\mathcal{A}}.$$

Case 2: As stated in Lemma 17, the PL condition implies the QG condition. Therefore,

$$\frac{\mu}{2}\|w_1 - w_1^*\|^2 \leq |f_S(w_1) - f_S(w_1^*)|. \qquad (2)$$

By assumption on case 2, $|f_S(w_1) - f_S(w_1^*)| \leq \epsilon'_{\mathcal{A}}$. This implies

$$\begin{aligned}\|w_1 - w_1^*\| &\leq \frac{\sqrt{2}}{\sqrt{\mu}}\sqrt{|f_S(w_1) - f_S(w_1^*)|} \\ &\leq \sqrt{\frac{2\epsilon'_{\mathcal{A}}}{\mu}}.\end{aligned}$$

Case 3: By Lemma 16, the PL condition on w_1, w_1^* implies that

$$\|\nabla f_S(w_1)\| \geq \mu\|w_1 - w_1^*\|.$$

Using the fact that f_S is L-Lipschitz and the fact that $\|\nabla f_S(w_j)\| \leq \epsilon''_{\mathcal{A}}$ by assumption on Case 3, we find

$$\begin{aligned}&|\ell(w_1; z_i) - \ell(w_1^*; z_i)| \\ &\leq L\|w_1 - w_1^*\| \leq \frac{L}{\mu}\|\nabla f_S(w_1)\| \\ &\leq \frac{2L\epsilon''_{\mathcal{A}}}{\mu}.\end{aligned}$$

In the above three cases, we can bound $|\ell(w_2; z_i) - \ell(w_2^*; z_i)|$ in the same manner. We now wish to bound the second term of (1). Note that we can manipulate this term as

$$\begin{aligned}&|\ell(w_1^*; z_i) - \ell(w_2^*; z_i)| \\ &= |(nf_S(w_1^*) - (n-1)f_{S^i}(w_1^*)) \\ &\quad - (nf_S(w_2^*) + (n-1)f_{S^i}(w_2^*))| \\ &\leq n|f_S(w_1^*) - f_S(w_2^*)| + (n-1)|f_{S^i}(w_1^*) - f_{S^i}(w_2^*)|.\end{aligned}$$

By the PL condition, we can find a local minimum u of f_S such that

$$\|\nabla f_S(w_2^*)\|^2 \geq \mu|f_S(w_2^*) - f_S(u)|.$$

Similarly, we can find a local minimum v of f_{S^i} such that

$$\|\nabla f_{S^i}(w_1^*)\|^2 \geq \mu|f_{S^i}(w_1^*) - f_{S^i}(v)|.$$

Note that since $\nabla f_{S^i}(w_2^*) = 0$, we get:

$$\|\nabla f_S(w_2^*)\|^2 = \frac{1}{n^2}\|\nabla \ell(w_2^*; z_i)\|^2 \leq \frac{L^2}{n^2}.$$

Similarly, since $\nabla f_S(w_1^*) = 0$, we get:

$$\|\nabla f_{S^i}(w_1^*)\|^2 = \frac{1}{(n-1)^2}\|\nabla \ell(w_1^*; z_i)\|^2 \leq \frac{L^2}{(n-1)^2}.$$

Since all local minima of a PL function are global minima, we obtain

$$\begin{aligned}&n|f_S(w_1^*) - f_S(w_2^*)| \\ &\leq n|f_S(w_1^*) - f_S(u)| + n|f_S(u) - f_S(w_2^*)| \\ &\leq n\frac{L^2}{\mu n^2} \\ &= \frac{L^2}{\mu n}.\end{aligned}$$

In a similar manner, we get

$$\begin{aligned}&(n-1)|f_{S^i}(w_1^*) - f_{S^i}(w_2^*)| \\ &\leq (n-1)|f_{S^i}(w_1^*) - f_{S^i}(v)| + (n-1)|f_{S^i}(v) - f_{S^i}(w_2^*)| \\ &\leq (n-1)\frac{L^2}{\mu(n-1)^2} \\ &\leq \frac{L^2}{\mu(n-1)}.\end{aligned}$$

Therefore,

$$|\ell(w_1^*; z_i) - \ell(w_2^*; z_i)| \leq \frac{L^2}{\mu n} + \frac{L^2}{\mu(n-1)}.$$

This proves the desired result. □

A.3. Proof of Theorem 5

Proof. By analogous reasoning to the proof of Theorem 3, we have the following decomposition:

$$\begin{aligned} &|\ell(w_1; z_i) - \ell(w_2; z_i)| \\ &\leq |\ell(w_1; z_i) - \ell(w_1^*; z_i)| + |\ell(w_1^*; z_i) - \ell(w_2^*; z_i)| \\ &\quad + |\ell(w_2^*; z_i) - \ell(w_2; z_i)|. \end{aligned} \tag{3}$$

We first wish to bound the first and third terms of (3). The bound depends on the case in Theorem 5.

Case 1: By assumption, $\|w_1 - w_1^*\| \leq \epsilon_{\mathcal{A}}$. Since $\ell(\cdot; z_i)$ is L-Lipschitz, this implies

$$|\ell(w_1; z_i) - \ell(w_1^*; z_i)| \leq L\|w_1 - w_1^*\| \leq L\epsilon_{\mathcal{A}}.$$

Case 2: By the QG condition, we have

$$\frac{\mu}{2}\|w_1 - w_1^*\|^2 \leq |f_S(w_1) - f_S(w_1^*)|. \tag{4}$$

By assumption on case 2, $|f_S(w_1) - f_S(w_1^*)| \leq \epsilon'_{\mathcal{A}}$. This implies

$$\begin{aligned} \|w_1 - w_1^*\| &\leq \frac{\sqrt{2}}{\sqrt{\mu}}\sqrt{|f_S(w_1) - f_S(w_1^*)|} \\ &\leq \sqrt{\frac{2\epsilon'_{\mathcal{A}}}{\mu}}. \end{aligned}$$

In the above two cases, we can bound $|\ell(w_2; z_i) - \ell(w_2^*; z_i)|$ in the same manner. We now wish to bound the second term of (3). By the QG property, we can pick some local minima v of f_S such that

$$\|w_2^* - v\| \leq \frac{2}{\sqrt{\mu}}\sqrt{|f_S(w_2^*) - f_S(v)|}. \tag{5}$$

We then have

$$\begin{aligned} &|\ell(w_1^*; z_i) - \ell(w_2^*; z_i)| \\ &\leq |\ell(w_1^*; z_i) - \ell(v; z_i)| + |\ell(v; z_i) - \ell(w_2^*; z_i)|. \end{aligned}$$

By construction, $f_S(w_1^*) = f_S(u) = 0$, so $|\ell(w_1^*; z_i) - \ell(v; z_i)| = 0$. By the Lipschitz property and the QG condition we have

$$\begin{aligned} &|\ell(v; z_i) - \ell(w_2^*; z_i)| \\ &\leq L\|v - w_2^*\| \\ &\leq \frac{2L}{\sqrt{\mu}}\sqrt{|f_S(w_2^*) - f_S(v)|} \\ &\leq \frac{2L}{\sqrt{\mu}}\sqrt{|f_S(w_2^*) - f_S(w_1^*)| + |f_S(w_1^*) - f_S(v)|}. \end{aligned} \tag{6}$$

Note that $|f_S(w_1^*) - f_S(v)| = 0$ by our realizability assumption. By assumption on w_1^*, w_2^*, we know that $f_S(w_1^*) \leq f_S(w_2^*)$ and $f_{S^i}(w_2^*) \leq f_{S^i}(w_1^*)$. Some simple analysis shows

$$\begin{aligned} nf_S(w_2^*) &= nf_{S^i}(w_2^*) + \ell(w_2^*; z_i) \\ &\leq nf_{S^i}(w_1^*) + \ell(w_2^*; z_i) \\ &= nf_S(w_1^*) + \ell(w_2^*; z_i) - \ell(w_1^*; z_i). \end{aligned}$$

Since $\ell(w_2^*; z_i) \leq c$, this implies that $n|f_S(w_2^*) - f_S(w_1^*)| \leq c$. Plugging this bound into (6), we get

$$|\ell(v; z_i) - \ell(w_2^*; z_i)| \leq 2L\sqrt{\frac{c}{\mu n}}.$$

This proves the desired result. □

A.4. Proof of Theorem 7

Let $S_1 = \{z_1, \ldots, z_n\}, S_2 = \{z_1, \ldots, z_{n-1}, z'_n\}$ be data sets of size n differing only in one entry. Let w_i denote the output of $\mathcal{A}$ on data set S_i and let w_i^* denote $w_{S_i}^*$. Let $f_i(w) = f_{S_i}(w)$.

Proof. Using the fact that $\ell(\cdot; z)$ is L-Lipschitz we get

$$\begin{aligned} &|\ell(w_1; z) - \ell(w_2; z)| \\ &\leq L\|w_1 - w_2\| \\ &\leq L\|w_1 - w_1^*\| + L\|w_1^* - w_2^*\| + L\|w_2^* - w_2\|. \end{aligned} \tag{7}$$

Note that by **A1**, we know that w_1^* is the closest optimal point of f_1 to w_2^*. By the PL condition,

$$\begin{aligned} &\|w_1^* - w_2^*\| \\ &\leq \frac{1}{\mu}\|\nabla f_1(w_2^*)\| \\ &= \frac{1}{\mu}\|\nabla f_2(w_2^*) - \frac{1}{n}\nabla\ell(w_2^*; z'_n) + \frac{1}{n}\nabla\ell(w_2^*; z_n)\| \\ &\leq \frac{1}{\mu n}(\|\nabla\ell(w_2^*; z'_n)\| + \|\nabla\ell(w_2^*; z_n)\|) \\ &\leq \frac{2L}{\mu n}. \end{aligned}$$

This bounds the second term of (7). The first and third terms must be bounded differently depending on the case. These can be bounded analogously via the method in the proof of Theorem 3, completing the proof. □

A.5. Proof of Theorem 8

We will use the following lemma.

Lemma 18. *Let f_S be QG and assume that $\ell(w; z) \leq c$ for all z and $w \in \mathcal{X}$. We assume* **A1** *as above. Then for S_1, S_2 differing in at most one place,*

$$\|w_1^* - w_2^*\| \leq 2\sqrt{\frac{c}{\mu n}}.$$

Proof. By the QG property:

$$\begin{aligned}&\frac{\mu}{2}\|w_1^* - w_2^*\|^2\\ &\leq |f_1(w_2^*) - f_1(w_1^*)|\\ &\leq |f_1(w_2^*) - f_2(w_2^*)| + |f_2(w_2^*) - f_1(w_1^*)|.\end{aligned}$$

Note that for all w, $|f_1(w) - f_2(w)| = \frac{1}{n}|\ell(w; z_n) - \ell(w; z_n')|$, so this is bounded by $\frac{c}{n}$.

By this same reasoning we get:

$$f_2(w_2^*) \leq f_2(w_1^*) \leq f_1(w_1^*) + \frac{c}{n}.$$

The desired result follows. □

Proof of Theorem 8. Using the fact that $\ell(\cdot; z)$ is L-Lipschitz we get

$$\begin{aligned}&|\ell(w_1; z) - \ell(w_2; z)|\\ &\leq L\|w_1 - w_2\|\\ &\leq L\|w_1 - w_1^*\| + L\|w_1^* - w_2^*\| + L\|w_2^* - w_2\|. \qquad (8)\end{aligned}$$

Note that by **A1**, we know that w_1^* is the closest optimal point of f_1 to w_2^*. By Lemma 18,

$$\|w_1^* - w_2^*\| \leq 2\sqrt{\frac{c}{\mu n}}.$$

This bounds the second term of (7). The first and third terms can be bounded analogously to the methods used in the proof of Theorem 5. This proves the desired result. □

A.6. Proof of Theorem 9

Proof. For almost all w, we can write $\sigma(Xw)$ as $\text{diag}(b)Xw$ for a vector b where $b_i(Xw)_i = \sigma((Xw)_i)$ (this only excludes points w such that $(Xw)_i$ is on a cusp of the piecewise-linear function). Then in an open neighborhood of such an w, we find

$$f(w) = g(\sigma(Xw)) = g(\text{diag}(b)Xw).$$

For a given w, let w_p be the closest global minima of f (*i.e.,* the closest point such that $f^* = f(w_p)$). By strong convexity of g, we find:

$$\begin{aligned}&g(\text{diag}(b)Xw_p) - g(\text{diag}(b)Xw)\\ &\geq \langle \nabla g(\text{diag}(b)Xw), \text{diag}(b)X(w_p - w)\rangle\\ &\quad + \frac{\lambda}{2}\|\text{diag}(b)X(w_p - w)\|^2\\ &\geq \langle X^T \text{diag}(b)^T \nabla g(\text{diag}(b)Xw), w_p - w\rangle\\ &\quad + \frac{\lambda}{2}\|\text{diag}(b)X(w_p - w)\|^2.\end{aligned}$$

Noting that $f(w) = g(\text{diag}(b)Xw)$, this implies

$$\begin{aligned}f(w_p) - f(w) \geq{}& \langle \nabla f(w), w_p - w\rangle\\ &+ \frac{\lambda \sigma_{\min}(X)^2 \sigma_{\min}(\text{diag}(b))^2}{2}\|w_p - w\|^2.\end{aligned}$$

Note that the minimum singular value of $\text{diag}(b)$ is the square root of the minimum eigenvalue of $\text{diag}(b)^2$. Since $\text{diag}(b)^2$ has entries c_i^2 on the diagonal, we know that the minimum singular value is at least $c = \min_i\{|c_i|\}$. Therefore we get:

$$\begin{aligned}&f(w_p) - f(w)\\ &\geq \langle f(w), w_p - w\rangle + \frac{\lambda \sigma_{\min}(X)^2 c^2}{2}\|w_p - w\|^2\\ &\geq \min_y \left[\langle \nabla f(w), y - w\rangle + \frac{\lambda \sigma_{\min}(X)^2 c^2}{2}\|y - w\|^2\right]\\ &= -\frac{1}{2\lambda \sigma_{\min}(X)^2 c^2}\|\nabla f(w)\|^2.\end{aligned}$$

□

A.7. Proof of Lemma 10

Proof. Using basic properties of the Frobenius norm and the definition of the pseudo-inverse, we have

$$\begin{aligned}&\|(WX - Y)X^T\|_F^2 \|(XX^T)^{-1}X\|_F^2\\ &\geq \|(WX - Y)X^T(XX^T)^{-1}X\|_F^2\\ &= \|(WX - Y)X^+X\|_F^2\\ &= \|WXX^+X - YX^+X\|_F^2\\ &= \|WX - YX^+X\|_F^2.\end{aligned}$$

This last step follows by basic properties of the pseudo-inverse. By the triangle inequality,

$$\begin{aligned}\|WX - Y\|_F^2 &= \|WX - YX^+X + YX^+X - Y\|_F^2\\ &\leq \|YX^+X - Y\|_F^2 + \|WX - YX^+X\|_F^2.\end{aligned}$$

Note that $YX^+X - Y$ is the component of Y that is orthogonal to the row-space of X, while YX^+X is the projection of

Y on to this row space. Therefore, $YX^+X - Y$ is orthogonal to $WX - YX^+X$ with respect to the trace inner product. Therefore, the inequality above is actually an equality, that is

$$\|WX - Y\|_F^2 = \|YX^+X - Y\|_F^2 + \|WX - YX^+X\|_F^2.$$

Putting this all together, we find

$$\begin{aligned}&\|(XX^T)^{-1}X\|_F^2 \|(WX - Y)X^T\|_F^2 \\ &\geq \|WX - Y\|_F^2 - \|YX^+X - Y\|_F^2.\end{aligned}$$

□

A.8. Proof of Lemma 11

Proof. Our proof uses similar techniques to that in (Hardt & Ma, 2016). This result can be viewed as a parallel version of their results under the lens of the PL condition. We wish to compute the gradient of f with respect to a matrix W_j. One can show the following:

$$\frac{\partial f}{\partial W_j} = W_{j+1}^T \dots W_\ell^T (WX - Y)X^T W_1^T \dots W_{j-1}^T.$$

Using the fact that for a matrix $A \in \mathbb{R}^{d\times d}$ and another matrix $B \in \mathbb{R}^{d\times k}$, we have $\|AB\|_F \geq \sigma_{\min}(A)\|B\|_F$, we find:

$$\left|\left|\frac{\partial f}{\partial W_j}\right|\right|_F \geq \prod_{i\neq j} \sigma_{\min}(W_i)\|(WX - Y)X^T\|_F.$$

By assumption, $\sigma_{\min}(W_j) \geq \tau$. Therefore:

$$\left|\left|\frac{\partial f}{\partial W_j}\right|\right|_F^2 \geq \tau^{2\ell-2}\|(WX - Y)X^T\|_F^2.$$

Taking the gradient with respect to all W_i we get:

$$\begin{aligned}\left|\left|\frac{\partial f}{\partial (W_1,\dots,W_\ell)}\right|\right|_F^2 &= \sum_{j=1}^{\ell}\left|\left|\frac{\partial f}{\partial W_j}\right|\right|_F^2 \\ &\geq \ell\tau^{2\ell-2}\|(WX - Y)X^T\|_F^2.\end{aligned}$$

□

A.9. Proof of Theorem 12

Proof. Since each W_i has full rank, we know that $\tau = \min_i \sigma_{\min}(W_i) > 0$. Using Lemma 11 and the fact that $\nabla f(W_1,\dots,W_\ell) = 0$, we get $\|(WX - Y)X^T\|_F^2 = 0$. Therefore, $WXX^T = YX^T$. Assuming that (XX^T) is invertible, we find that $W = YX^+$, which equals W^*. □

A.10. Proof of Theorem 13

Proof. By Lemma 11 and Lemma 10 we find:

$$\begin{aligned}&\frac{1}{2}\left|\left|\frac{\partial f}{\partial (W_1,\dots,W_\ell)}\right|\right|_F^2 \\ &\geq \ell\tau^{2\ell-2}\frac{1}{2}\|(WX - Y)X^T\|_F^2 \\ &\geq \frac{\ell\tau^{2\ell-2}}{\|(XX^T)^{-1}X\|_F^2}\frac{1}{2}(\|WX - Y\|_F^2 - \|YX^+X - Y\|_F^2) \\ &= \frac{\ell\tau^{2\ell-2}}{\|(XX^T)^{-1}X\|_F^2}(f(W_1,\dots,W_\ell) - f^*).\end{aligned}$$

□

B. Stability Properties of SGD

B.1. Stability of Gradient Descent for Convex Loss Functions

To prove the stability of gradient descent, we will assume that the underlying loss function is smooth.

Definition 6. *A function $f : \Omega \to \mathbb{R}$ is β-smooth if for all $u, v \in \Omega$, we have*

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

Hardt et al. (2016) show the following theorem.

Theorem 19 ((Hardt et al., 2016)). *Let $\ell(\cdot; z)$ be L-Lipschitz, β-smooth, and convex for all z. Say we perform T iterations of SGD with a constant step size $\gamma \leq \frac{2}{\beta}$ to train iterates w_t on S and $\hat{w}_t$ on S'. Then for all such S, S' with $|S| = |S'| = n$ such that S, S' differ in at most one example,*

$$\mathbb{E}_{\mathcal{A}}[\|w_T - \hat{w}_T\|] \leq \frac{2\gamma LT}{n}$$

If $\ell(\cdot; z)$ is λ-strongly convex for all z, then

$$\mathbb{E}_{\mathcal{A}}[\|w_T - \hat{w}_T\|] \leq \frac{2L}{\lambda n}$$

Performing similar analysis for gradient descent, we obtain the following theorem.

Theorem 20. *Assume that for all z, $\ell(\cdot; z)$ is convex, β-smooth, and L-Lipschitz. Say we run GD for T iterations with step sizes γ_t such that $\gamma_t \leq \frac{2}{\beta}$. Then GD is uniformly stable with*

$$\epsilon_{stab} \leq \frac{2L^2}{n}\sum_{t=0}^{T}\gamma_t.$$

If $\ell(\cdot; z)$ is λ-strongly convex for all z, then GD is uniformly stable with

$$\epsilon_{stab} \leq \frac{2L}{\lambda n}.$$

To prove this theorem, we use similar techniques to those in (Hardt et al., 2016). We first consider the convex case.

Proof. By direct computation, we have:

$$\begin{aligned}
&\|w_T - \hat{w}_T\| \\
&= \Bigg\| w_{T-1} - \hat{w}_{T-1} \\
&\quad - \frac{\gamma_T}{n}\sum_{i=1}^{n-1}\big(\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\big) \\
&\quad - \frac{\gamma_T}{n}\nabla\ell(w_{T-1};z_n) + \frac{\gamma_T}{n}\nabla\ell(\hat{w}_{T-1};z_n')\Bigg\| \\
&\le \Bigg\| w_{T-1} - \hat{w}_{T-1} \\
&\quad - \frac{\gamma_T}{n}\sum_{i=1}^{n-1}\big(\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\big)\Bigg\| \\
&\quad + \Bigg\|\frac{\gamma_T}{n}\nabla\ell(w_{T-1};z_n) - \frac{\gamma_T}{n}\nabla\ell(\hat{w}_{T-1};z_n')\Bigg\|.
\end{aligned}$$

Note that since $\ell(\cdot;z)$ is L-Lipschitz, the second part of this summand is bounded by $\dfrac{2\gamma_T L}{n}$. We now wish to bound the first part. Using the triangle inequality and co-coercivity of $\nabla\ell(\cdot,z_i)$, we have

$$\begin{aligned}
&\Bigg\| w_{T-1} - \hat{w}_{T-1} \\
&\quad - \frac{\gamma_T}{n}\sum_{i=1}^{n-1}\big(\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\big)\Bigg\|^2 \\
&\le \|w_{T-1} - \hat{w}_{T-1}\|^2 \\
&\quad + \sum_{i=1}^{n-1}\Big(\frac{\gamma_T^2}{n^2} - \frac{2\gamma_T}{n^2\beta}\Big)\|\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\|^2.
\end{aligned}$$

Note that in particular, if $\gamma_T \le \frac{2}{\beta}$, each of the $n-1$ summands on the right will be non-positive. Therefore for such γ_T, we get:

$$\begin{aligned}
&\Bigg\| w_{T-1} - \hat{w}_{T-1} \\
&\quad - \frac{\gamma_T}{n}\sum_{i=1}^{n-1}\Big(\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\Big)\Bigg\| \\
&\le \|w_{T-1} - \hat{w}_{T-1}\|.
\end{aligned}$$

So, if $\frac{\gamma_t}{n} \le \frac{2}{\beta}$ for all t, we get:

$$\begin{aligned}
\|w_T - \hat{w}_T\| &\le \|w_{T-1} - \hat{w}_{T-1}\| + \frac{2\gamma_T L}{n} \\
&\le \|w_{T-2} - \hat{w}_{T-2}\| + \frac{2L}{n}(\gamma_{T-1} + \gamma_T) \\
&\ \vdots \\
&\le \|w_0 - \hat{w}_0\| + \frac{2L}{n}\sum_{t=1}^{T}\gamma_t \\
&= \frac{2L}{n}\sum_{t=1}^{T}\gamma_t.
\end{aligned}$$

This last step follows from the fact that $w_0 = \hat{w}_0$ if we initialize at the same point. Using the fact that $f(w)$ is L-Lipschitz (since ℓ is), we get:

$$|f(w_T;z) - f(\hat{w}_T;z)| \le \frac{2L^2}{n}\sum_{t=0}^{n}\gamma_t.$$

□

We now move to the λ-strongly convex case. For simplicity of analysis, we assume that we use a constant step size γ such that $\gamma \le 1/\beta$.

Proof. The proof remains the same, except when using co-coercivity. Under this assumption, some plug and play in an analogous fashion will show:

$$\begin{aligned}
&\Bigg\| w_{T-1} - \hat{w}_{T-1} \\
&\quad - \frac{\gamma_T}{n}\sum_{i=1}^{n-1}\big(\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\big)\Bigg\|^2 \\
&\le \Bigg(1 - \frac{2\gamma\lambda\beta}{\lambda+\beta}\Bigg)\|w_{T-1} - \hat{w}_{T-1}\|^2 \\
&\quad + \sum_{i=1}^{n-1}\Big(\frac{\gamma_T^2}{n^2} - \frac{2\gamma_T}{n\beta}\Big)\|\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\|^2.
\end{aligned}$$

Note that if $\gamma \le \frac{1}{\beta}$ then the second term is nonnegative and one can show that this implies:

$$\begin{aligned}
&\Bigg\| w_{T-1} - \hat{w}_{T-1} \\
&\quad - \frac{\gamma_T}{n}\sum_{i=1}^{n-1}\big(\nabla\ell(w_{T-1};z_i) - \nabla\ell(\hat{w}_{T-1};z_i)\big)\Bigg\| \\
&\le \Bigg(1 - \gamma\lambda\Bigg)\|w_{T-1} - \hat{w}_{T-1}\|.
\end{aligned}$$

Combining, this shows:

$$\begin{aligned}\|w_T - \hat{w}_{T-1}\| &\leq (1-\gamma\lambda)\|w_{T-1} - \hat{w}_{T-1}\| + \frac{2L\gamma}{n}\\ &\vdots\\ &\leq \frac{2L\gamma}{n}\sum_{t=0}^{T}(1-\gamma\lambda)^t\\ &\leq \frac{2L}{\lambda n}.\end{aligned}$$

This implies uniform stability with parameter $\frac{2L^2}{\lambda n}$. □

B.2. Proof of Theorem 15

Suppose we run SGD on the above $f_S, f_{S'}$ with step size 1 and initialize near $\hat{w}$ (we will be more concrete later about where we initialize). With probability $\frac{n-1}{n}$, the first iteration of SGD will use the same example for both S and S', either $z = (-1, 1) = \alpha$ or $z = (-\frac{1}{2}, 1) = \beta$. Computing derivatives at $\hat{w}$ shows

$$\frac{d}{dw}\ell(w;\alpha) \approx -0.486254, \quad \frac{d}{dw}\ell(w;\beta) \approx 0.486254.$$

In both cases, the slope is at least 0.4. Therefore, there is some η such that for all $w \in [\hat{w}-\eta, \hat{w}+\eta]$,

$$\frac{d}{dw}\ell(w;\alpha) < -0.4, \quad \frac{d}{dw}\ell(w;\beta) > 0.4.$$

Since we are taking $\gamma = 1$ and $\hat{w} \approx 0.598004$, this implies that with probability $\frac{n-1}{n}$, after one step of SGD we move outside of the open interval $(0.5, 1)$. This is important since this is the interval where $\ell(w;\alpha)$ and $\ell(w;\beta)$ have slopes that point them towards distinct basins. Similarly, outside of this interval we also have that the slopes of $\ell(w;(z_\pm, 0))$ point towards the same basin.

Continuing to run SGD in this setting, even if we now decrease the step size, will eventually lead us to the same basins of $f_S(w), f_{S'}(w)$. Let $w_S, w_{S'}$ denote the outputs of SGD in this setting after enough steps so that we get convergence to within $\frac{1}{n}$ of a local minima. If our first sample z was $(-1, 1)$, we will end up in the right basin, while if our first sample z was $(-\frac{1}{2}, 1)$, we will end up in the left basin. In particular, for ϵ small, $z_\pm$ are close enough that the minima of $f_S, f_{S'}$ are within $\frac{1}{n}$ of each other. Note that the minima w_1, w_2 that $w_S, w_{S'}$ are converging to are different. However, because they are in the same basin we know that for ϵ small, $z_\pm$ are close enough that $\|w_S - w_{S'}\| \leq O(\frac{1}{n})$. Therefore, $\|w_S - w_{S'}\| \leq O(\frac{1}{n})$.

In our proof of the instability of gradient descent, we only needed to look at z satisfying $\|z\| \leq 2$ to see the instability. However, for SGD if we restrict to $\|z\| \leq 2$, then by compactness we know that $\ell(w;z)$ will be Lipschitz. Therefore, with probability $\frac{n-1}{n}$, $\|\ell(w_S;z) - \ell(w_{S'};z)\| \leq L\|w_S - w_{S'}\| \leq O(\frac{1}{n})$.

With probability $\frac{1}{n}$, SGD first sees the example on which S, S' differ. In this case, $w_S, w_{S'}$ may end up in different basins of the right-most graph in Figure 2. Restricting to $\|z\| \leq 2$, by compactness we have $|\ell(w_S;z) - \ell(w_{S'};z)| \leq C$ for some constant C. Therefore, $|\ell(w_S;z) - \ell(w_{S'};z)| = O(\frac{1}{n})$ with probability $1 - \frac{1}{n}$ and $|\ell(w_S;z) - \ell(w_{S'};z)| = O(1)$ with probability $\frac{1}{n}$. This proves the desired theorem.

B.3. Convergence of SGD with Step Sizes $\gamma_t = c/t$

In this subsection we show that a provable rate for SGD on smooth functions with learning rate proportional to $O(1/t)$ might require a large number of iterations. While (Hardt et al., 2016) show stability of SGD in non-convex settings with such a step size, their stability bounds grow close to linearly with the number of iterations. When exponentially many steps are taken, this no longer implies useful generalization bounds on SGD.

When a function $f(x) = \sum_{i=1}^n f_i(x)$ is β smooth on its domain, then the following holds:

$$f(x) - f(y) \leq \langle \nabla f(y), x - y\rangle + \frac{\beta}{2}\|x-y\|^2.$$

Fix some t and let $x = x_{t+1} = x_t - \gamma_t \nabla f_{s_t}(x_t)$ and let $y = x_t$. Here, x_0 is set to some initial vector value, s_t is a uniform i.i.d. sample from $\{1, \ldots, n\}$, and $\gamma_t = c/t$ for some constant $c > 0$. Then, due to the β-smoothness of f we have the following:

$$\begin{aligned}&f(x_{t+1}) - f(x_t)\\ &\leq \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{\beta\gamma_t^2}{2}\|\nabla f_{s_t}(x_t)\|^2\\ &\leq -\gamma_t\langle \nabla f(x_t), \nabla f_{s_t}(x_t)\rangle + \frac{\beta\gamma_t^2}{2}\|\nabla f_{s_t}(x_t)\|^2.\end{aligned}$$

Taking expectation with respect to all random samples s_t and rearranging yields

$$\left(\gamma_t - \frac{\beta\gamma_t^2}{2}\right)\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \mathbb{E}\left[f(x_t) - f(x_{t+1})\right].$$

Summing the above inequality for all t terms from 0 to T we get

$$\begin{aligned}&\sum_{t=1}^{T}\left(\gamma_t - \frac{\beta\gamma_t^2}{2}\right)\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \mathbb{E}[f(x_0) - f(x_T)]\\ \Rightarrow &\sum_{t=1}^{T}\left(\gamma_t - \frac{\beta\gamma_t^2}{2}\right)\mathbb{E}[\|\nabla f(x_t)\|^2] \leq f(x_0)\\ \Rightarrow &\min_{t=1,\ldots,T}\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \frac{f(x_0)}{\sum_{t=1}^{T}\left(\gamma_t - \frac{\beta\gamma_t^2}{2}\right)}.\end{aligned}$$

Assuming that $\gamma_t = c/t$, we have

$$\begin{aligned}
&\min_{t=1,\dots,T} \mathbb{E}[\|\nabla f(x_t)\|^2] \le \frac{f(x_0)}{\sum_{t=1}^{T} c\left(\gamma_t - \frac{\beta c^2}{2t^2}\right)} \\
\Rightarrow &\min_{t=1,\dots,T} \mathbb{E}[\|\nabla f(x_t)\|^2] \le \frac{f(x_0)}{c\sum_{t=1}^{T} t^{-1}} \\
\Rightarrow &\min_{t=1,\dots,T} \mathbb{E}[\|\nabla f(x_t)\|^2] \le \frac{f(x_0)}{C_1 \log(T)}.
\end{aligned}$$

Here C_1 is a universal constant that depends only on c. Observe that even if $C_2 = 0$, using the above simple bounding technique (a simplified version of the non-convex convergence bounds of (Ghadimi & Lan, 2013)), requires $O(e^{-\epsilon})$ steps to reach error ϵ, an exponentially large number of steps.

We note that the above bound does not imply that there does not exist a smooth function for which $1/t$ step sizes suffice for polynomial-time convergence (in fact there are several convex problems for which $1/t$ suffices for fast convergence). However, the above implies that when we are only assuming smoothness on a non-convex function, it may be the case that there exist non-convex problems where $1/t$ implies exponentially slow convergence.

Moreover, assuming that a function has bounded stochastic gradients, *e.g.,* $\mathbb{E}\|f_{s_t}(x_t)\| \le M$, it is also easy to show that after T steps each with step size $\gamma_t = c/t$, then the distance of the current SGD model x_T from the initial iterate x_0 satisfies

$$\mathbb{E}\|x_T - x_0\| \le 2M \log(T).$$

This implies that if the optimal model is at distance $\Omega(Md)$ from x_0, we would require at least $O(e^{M\cdot d})$ iterations to reach it in expectation.