SHORT OPTIMIZATION PATHS LEAD TO GOOD GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimization and generalization are two essential aspects of machine learning. In this paper, we propose a framework to connect optimization with generalization by analyzing the generalization error based on the length of optimization trajectory under the gradient flow algorithm after convergence. Through our approach, we show that, with a proper initialization, gradient flow converges following a short path with an explicit length estimate. Such an estimate induces a length-based generalization bound, showing that short optimization paths after convergence indicate good generalization. Our framework can be applied to broad settings. For example, we use it to obtain generalization estimates on three distinct machine learning models: underdetermined ℓ_p linear regression, kernel regression, and overparameterized two-layer ReLU neural networks.

1 INTRODUCTION

From the perspective of statistical learning theory, the goal of machine learning is to find a predictive function that can give accurate predictions on new data. For supervised learning problems, empirical risk minimization (ERM) is a common practice to achieve this goal. The idea of ERM is to minimize a cost function on observed data by an optimization algorithm. Therefore, a fundamental question is *whether an optimization algorithm produces a solution with good generalization*.

Recent works have shed light on providing theoretical explanations to this question from different angles. One line of works considered the case when gradient methods converge to minimal norm solutions (Bartlett et al., 2020; Tsigler & Bartlett, 2020; Liang & Rakhlin, 2020; Liang et al., 2020) on kernel regression, and then analyzed the generalization of those minimal norm solutions. However, the phenomenon of norm minimization has been known to happen only for the quadratic loss with an appropriate initialization. Another line of works focused on overparameterized models, e.g., neural networks under the Neural Tangent Kernel (NTK) regime (Allen-Zhu et al., 2019; Arora et al., 2019; Cao & Gu, 2020; Ji & Telgarsky, 2020; Chen et al., 2021), proving that overparameterized neural networks trained by (stochastic) gradient descent ((S)GD) have good generalization performance on certain target functions (for example, polynomial functions).

Although existing works have made significant progress on the interplay of optimization and generalization, they focused on studying specific models, such as the NTK and models possessing minimal norm solutions. In this paper, instead, we study general loss function conditions which induce interesting connections between optimization and generalization.

We start with a simple observation, as shown in Figure 1, that for an underdetermined linear regression model with different random initializations, the generalization performance is closely related to the length of the optimization path[1] after convergence. In particular, the shorter the optimization path is, the better will the model generalize. Note that here the shortness is not due to early stopping, but is in terms of convergence. This empirical investigation motivates us to use the length of the optimization path to connect optimization and generalization. In other words, a general condition that guarantees a short optimization path can be used to induce good generalization performance. Thus, inspired by the theory in (Bolte et al., 2007) that Łojasiewicz gradient inequality (LGI) induces an explicit bound for the gradient flow path, we consider *Uniform-LGI* (Definition 1). This is a modified version of

[1]For discrete iterations, we use the sum of the distance between every two consecutive iterations to represent the length of the optimization path (See details in Appendix A).

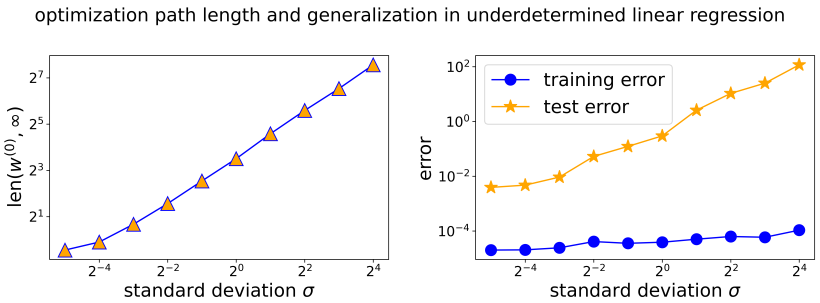

Figure 1: Illustrations of the relationship between optimization path lengths and generalization. **(Left)** Training an underdetermined linear regression model by gradient descent. We initialize the weights from Gaussian distribution $\mathcal{N}(0, \sigma^2)$ and record the gradient descent trajectory lengths $\mathsf{len}(w^{(0)}, \infty)$ from time 0 to ∞ (after convergence) for different σ. We observe that the trajectory length increases as the scale of initialization increases. **(Right)** We observe that short optimization paths corresponds to better generalization than long optimization paths. See more experiment settings in Appendix A.

LGI and plays a critical role in obtaining a length estimate for the gradient flow trajectory. Once the length is estimated, we can use our length-based generalization bounds to show the generalization performance.

Contributions. We summarize our contributions as follows:

- We focus on the gradient flow algorithm and propose a framework for combining optimization and generalization. This framework is based on the length of the gradient flow trajectory, and its key component is the *Uniform-LGI* property.
- We then prove that under appropriate conditions, gradient flow returns a global minimum and gives an explicit length-estimate for the gradient flow trajectory (Theorem 1). For such a global minimum, we derive a length-based generalization bound (Theorem 2).
- We further show applications of Theorem 1 and Theorem 2 to obtain generalization bounds on underdetermined ℓ_p linear regression (Theorem 3), kernel regression (Theorem 4), and overparameterized two-layer ReLU neural networks (Theorem 5). These bounds match or expand the type of scenarios where we can rigorously establish the phenomenon of benign overfitting.

2 Related Works

Optimization. Theoretically analyzing the training process of most machine learning models is a challenging problem as the loss landscapes are highly non-convex. One approach to studying non-convex optimization problems is to use the Polyak-Łojasiewicz (PL) condition (Polyak, 1963), which characterizes the local geometry of loss landscapes and ensures the existence of global minima. It is shown in (Karimi et al., 2016) that GD admits linear convergence for a class of optimization objective functions under the PL condition. In this paper, we consider the Łojasiewicz gradient inequality (Lojasiewicz, 1965), an extended version of the PL condition that can be applied to more cases. In particular, we modify the original LGI to get length estimates of the optimization paths, which will be used to derive length-based generalization bounds.

Generalization. Traditional VC dimension-based generalization bounds depend on the number of parameters and will be vacuous for huge models such as overparameterized neural networks. To overcome this limitation, several non-vacuous generalization bounds are proposed. For example, the norm/margin-based generalization bounds (Neyshabur et al., 2015; Bartlett et al., 2017; Golowich et al., 2018; Neyshabur et al., 2019), and the PAC-Bayes-based bounds (Dziugaite & Roy, 2017; Neyshabur et al., 2018; Zhou et al., 2019; Rivasplata et al., 2020). However, these bounds tend to focus less on optimization, e.g., norm-based generalization bounds may not discuss how small-norm solutions are obtained through practical training. In this paper, we connect the optimization and generalization by deriving generalization bounds based on the optimization trajectory length.

Interface between optimization and generalization. Implicit bias builds the bridge between optimization and generalization, which has been widely studied to explain the generalization ability of machine learning models. Recent works (Soudry et al., 2018a;b; Nacson et al., 2019a;b; Lyu & Li, 2020) showed that linear classifiers or deep neural networks trained by GD/SGD maximizes the margin of the separating hyperplanes and therefore generalizes well. Other works (Arora et al., 2019; Zou & Gu, 2019; Cao & Gu, 2020; Ji & Telgarsky, 2020; Chen et al., 2021) concentrated on overparameterized neural networks, showing that GD/SGD trajectory falls within the NTK regime, where the minimizer has good generalization due to the small complexity of the parameter space. In this work, we focus on specific conditions on loss functions under which we can connect optimization and generalization based on the lengths of optimization paths.

3 MAIN RESULTS

In this section, we aim to address optimization and generalization together by proving results based on the *Uniform-LGI* property. We begin with describing notations and the problem setting in Section 3.1. Then we give an explicit length estimate for the gradient flow trajectory in Section 3.2. Lastly, we show a length-based generalization bound in Section 3.3.

3.1 SETUP AND NOTATIONS

Consider a supervised learning problem on a hypothesis space $\mathcal{F} = \{f(w, \cdot) : \mathbb{R}^d \to \mathbb{R} \mid w \in \mathcal{W}\}$, where $\mathcal{W}$ is a parameter set in Euclidean space. Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and a training set $S = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$ with n independent and identically distributed (i.i.d.) samples from a joint distribution $\mathcal{D}$, the goal of ERM is to optimize the empirical loss function $\mathcal{L}_n(w)$ on S:

$$\arg\min_w \mathcal{L}_n(w) := \frac{1}{n}\sum_{i=1}^n \ell\left(f(w, x_i), y_i\right).$$

We assume that each input vector and label have bounded norms. Specifically, we assume that, for any (x, y) following $\mathcal{D}$, $\|x\| = 1, |y| \leq 1$. This can be achieved by data normalization. To simplify the analysis, we optimize the empirical loss by gradient flow[2]:

$$\frac{dw^{(t)}}{dt} = -\nabla \mathcal{L}_n(w^{(t)}),\ t \in [0, +\infty), \tag{1}$$

where $w^{(t)}$ is the parameter value at time t, $w^{(0)}$ is the initialization parameter set, and $\nabla \mathcal{L}_n(w)$ is the gradient of $\mathcal{L}_n(w)$ with respect to w. Given the gradient flow, we may define the length of the gradient flow from $w^{(0)}$ to $w^{(t)}$ as $\mathsf{len}(w^{(0)}, t) := \int_0^t \left\|\frac{dw^{(s)}}{ds}\right\| ds$.

Notations. We use $\|\cdot\|$ to denote the ℓ_2 norm of a vector or the spectral norm of a matrix, and use $\|\cdot\|_F$ to denote the Fronenius norm of a matrix. For a set $\mathcal{S} \in \mathbb{R}^n$, we use $\partial\mathcal{S}$ to denote its boundary. We use $d(\mathcal{S}_1, \mathcal{S}_2)$ to represent the Euclidean distance between two sets $\mathcal{S}_1, \mathcal{S}_2$, which is defined as $d(\mathcal{S}_1, \mathcal{S}_2) = \{\inf \|s_1 - s_2\| : s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2\}$. For two vectors, we use $\langle , \rangle$ to denote their inner product. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of a symmetric matrix A. For any positive integer n, we denote $[n] = \{1, 2, \ldots, n\}$. We use $\mathcal{O}(\cdot)$ to stand for Big-O notation.

The classical LGI gives a lower bound of the gradient of a differentiable function based on its value above its minimum. Many functions, e.g., real analytic functions and subanalytic functions, satisfy this property, at least locally (Bolte et al., 2007). Here, we require a global version of this inequality as a condition to control the gradient flow trajectory. Let us define this notion below.

Definition 1 (Uniform-LGI)**.** *A function $\mathcal{L}(w)$ satisfies Uniform-LGI on a set $\mathcal{S}$ with two constants $c > 0$ and $\theta \in [\frac{1}{2}, 1)$, if*

$$\|\nabla \mathcal{L}(w)\| \geq c\left(\mathcal{L}(w) - \min_{v \in \mathcal{W}} \mathcal{L}(v)\right)^\theta, \quad \forall w \in \mathcal{S}.$$

[2]Here we assume the existence of the gradient flow and the limit $\lim_{t\to\infty} w^{(t)}$.

In the special case when $\theta = 1/2$ and $c = \sqrt{2\mu}$, the Uniform-LGI corresponds to the μ-PL condition (Karimi et al., 2016). Here we give examples of functions satisfying the Uniform-LGI but not the PL condition: $\mathcal{L}(w) = w^{2k}, k \in \mathbb{Z}^+$. When $k \geq 2$, $\mathcal{L}(w)$ satisfies the Uniform-LGI on $\mathbb{R}$ with $c = 2k$ and $\theta = 1 - 1/2k$.

3.2 Gradient flow trajectory length estimate

Our first result shows that under the uniform-LGI condition, gradient flow returns a global minimum with an explicit estimate of the gradient flow trajectory length as in the following theorem.

Theorem 1. *For any initialization $w^{(0)}$, if $\mathcal{L}_n(w)$ satisfies Uniform-LGI on a closed set $\mathcal{S}_n$ with constants c_n, θ_n such that $\mathcal{S}_n \supseteq B\left(w^{(0)}, r_n(w^{(0)})\right)$, where*

$$r_n(w^{(0)}) = \frac{\left(\mathcal{L}_n(w^{(0)}) - \min_w \mathcal{L}_n(w)\right)^{1-\theta_n}}{c_n(1-\theta_n)}, \tag{2}$$

then $w^{(t)}$ converges to a global minimum $w^{(\infty)}$ with the trajectory length upper bounded by:

$$\mathsf{len}(w^{(0)}, \infty) \leq r_n(w^{(0)}).$$

And the convergence rate of $\mathcal{L}_n(w^{(t)})$ is given by:

When $\theta_n = \frac{1}{2}$,

$$\mathcal{L}_n(w^{(t)}) - \min_w \mathcal{L}_n(w) \leq e^{-c_n^2 t}\left(\mathcal{L}_n(w^{(0)}) - \min_w \mathcal{L}_n(w)\right);$$

When $\frac{1}{2} < \theta_n < 1$,

$$\mathcal{L}_n(w^{(t)}) - \min_w \mathcal{L}_n(w) \leq (1 + M_n t)^{-1/(2\theta_n - 1)}\left(\mathcal{L}_n(w^{(0)}) - \min_w \mathcal{L}_n(w)\right),$$

where $M_n = c_n^2(2\theta_n - 1)\left(\mathcal{L}_n(w^{(0)}) - \min_w \mathcal{L}_n(w)\right)^{2\theta_n - 1}$.

The proof of Theorem 1 is given in Appendix B.

This theorem yields that once the loss function satisfies the Uniform-LGI around the initialization, gradient flow returns a global minimum with an explicit trajectory length estimate. Moreover, for any fixed n, $\mathsf{len}(w^{(0)}, \infty)$ decreases as the initial loss value decreases. The calculation of c_n, θ_n for different n should be analyzed case by case based on the loss landscape, as shown in Section 4.

3.3 How does the optimization path length affect generalization?

The Uniform-LGI property allows us to have an upper bound for the gradient flow trajectory length and a fortiori, the distance between the initialization and the convergence point. Thus, in this subsection, we use the length estimate we get in Theorem 1 to study generalization.

Throughout this subsection, we assume that there exists an almost everywhere differentiable function $\Psi : \mathbb{R}^{p+q} \to \mathbb{R}$ such that $f(w, \cdot)$ can be represented in the following form,

$$\forall x \in \mathbb{R}^d, \quad f(w, x) = \Psi\left(\alpha_1^\top x, \ldots, \alpha_p^\top x, \beta_1, \ldots, \beta_q\right) \tag{3}$$

with $\alpha_1, \ldots, \alpha_p \in \mathbb{R}^d$, $\beta_1, \ldots, \beta_q \in \mathbb{R}$, and $w = \mathrm{vec}\left(\{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q\}\right) \in \mathbb{R}^{pd+q}$. Here vec is the vectorization operator that concatenates all elements into a column vector.

A wide class of functions can be represented in the form (3). Examples include linear functions $f(w, x) = w^\top x$ and two-layer neural networks $f(w, x) = a^\top \phi(Wx)$ with $a \in \mathbb{R}^m, W \in \mathbb{R}^{m \times d}$.

Additional notations. For the loss function ℓ, we use $L_\ell(\mathcal{S})$ to denote its Lipschitz constant (the maximal gradient norm) on $\mathcal{S}$ with respect to its first argument. For Ψ in (3), we define $L_\Psi(\mathcal{S}) := \left(L_\Psi^{(1)}(\mathcal{S}), \cdots, L_\Psi^{(p)}(\mathcal{S}), L_\Psi^{(p+1)}(\mathcal{S}), \cdots, L_\Psi^{(p+q)}(\mathcal{S})\right)^\top$, where $L_\Psi^{(i)}(\mathcal{S})$ is the Lipschitz

constant of Ψ on $\mathcal{S}$ with respect to the i-th variable. We use $\mathrm{vec}\left(\left\{\alpha_1^{(0)}, \ldots, \alpha_p^{(0)}, \beta_1^{(0)}, \ldots, \beta_q^{(0)}\right\}\right)$ to represent the initialization $w^{(0)}$ and use $\mathcal{L}_{\mathcal{D}}(w)$ to denote the expected loss $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f(w,x),y)\right]$.

Now we are ready to state our second main result with regard to generalization based on the gradient flow trajectory length estimate. In the following theorem, for $a = (a_1, \ldots, a_p)^\top$ and $b = (b_1, \ldots, b_q)^\top$, we define $\mathcal{S}_{a,b} = \left\{w : \forall i \in [p], j \in [q], \|\alpha_i - \alpha_i^{(0)}\| \leq a_i, |\beta_j - \beta_j^{(0)}| \leq b_j\right\}$, and $M_R = \sup_{\|a\|^2+\|b\|^2\leq R^2} \sup_{w\in\mathcal{S}_{a,b}, \|x\|\leq 1, |y|\leq 1} \ell\left(f(w,x),y\right)$.

Theorem 2. *Under the same conditions and notations as in Theorem 1, with probability at least $1-\delta$ over the random choice of training set S, we have*

$$\mathcal{L}_{\mathcal{D}}(w^{(\infty)}) \leq \min_w \mathcal{L}_n(w) + \sup_{\|a\|^2+\|b\|^2\leq R^2} \frac{2RL_\ell(\mathcal{S}_{a,b})\left\|L_\Psi(\mathcal{S}_{a,b})\right\|}{\sqrt{n}} + 3M_R\sqrt{\frac{3(p+q)+\log(2/\delta)}{2n}}, \tag{4}$$

where $R = \sqrt{2}r_n(w^{(0)})$ specified in (2).

The proof of Theorem 2 is given in Appendix C.

There are four key terms in (4). To make the bound easier to understand, we discuss them in sequence:

- $r_n(w^{(0)})$: It depends on the initial loss value and the Uniform-LGI constants c_n, θ_n. The asymptotic analysis of c_n, θ_n is problem-dependent, and we will provide several examples in Section 4.
- $\sup_{\|a\|^2+\|b\|^2\leq R^2} L_\ell(\mathcal{S}_{a,b})$ and M_R: These two terms depend on the structure of ℓ. For example, for any loss function $\ell : \mathbb{R}\times\mathbb{R} \to [0,1]$ that is 1-Lipschitz in the first argument, $L_\ell(\mathcal{S}_{a,b}) = M_R = 1$.
- $\sup_{\|a\|^2+\|b\|^2\leq R^2} \|L_\Psi(\mathcal{S}_{a,b})\|$: This term relies on $\mathsf{len}(w^{(0)}, \infty)$ and the structure of Ψ. For example, if $\Psi(x,y) = x+y$, then $\sup_{\|a\|^2+\|b\|^2\leq R^2} \|L_\Psi(\mathcal{S}_{a,b})\| = \sqrt{2}$.

In many practical cases, e.g. a linear function Ψ with a bounded and Lipschitz loss function ℓ, the dominant term of the length-based generalization gap is $\mathcal{O}(r_n(w^{(0)})/\sqrt{n})$ for a large n. Therefore, once $r_n(w^{(0)})$ does not grow faster than $\sqrt{n}$, the generalization gap will decrease to zero as n goes to infinity. Moreover, We will prove in the next section that, for several machine learning models, the term $r_n(w^{(0)})/\sqrt{n}$ does not suffer from the curse of dimensionality when the number of parameters is large enough, indicating that our length-based generalization bound is non-vacuous.

4 Applications

In this section, we apply the framework to three models. To obtain clean expressions of the generalization bound in terms of the sample size n, it is helpful to consider a range of n relating to the dimension d. In particular, we consider underdetermined systems where the ratio n/d remains finite unless stated otherwise:

$$\exists\, \gamma_0, \gamma_1 \in (0,\infty),\ s.t.\ \forall d,\ \gamma_0 d \leq n = n(d) \leq \gamma_1 d.$$

Note that this setting is non-asymptotic since we do not require d to be infinite. For each application, to simplify the analysis, we evaluate the test error on a loss function $\tilde{\ell} : \mathbb{R}\times\mathbb{R} \to [0,1]$ is 1-Lipschitz (on the first argument) with $\tilde{\ell}(y,y) = 0$. The following steps are taken in our framework:

Step 1. Establish the Uniform-LGI property and find the Uniform-LGI constants c_n and θ_n.

Step 2. Apply Theorem 1 to get optimization results.

Step 3. Apply Theorem 2 to get generalization results.

4.1 Underdetermined ℓ_p linear regression

We begin with an underdetermined linear regression model $f(w,x) = w^\top x$ with an ℓ_p loss function ($p \geq 2$ is an even positive integer):

$$\arg\min_{w\in\mathbb{R}^d} \mathcal{L}_n(w) := \frac{1}{pn}\sum_{i=1}^n \left(w^\top x_i - y_i\right)^p, \tag{5}$$

where the input data matrix $\mathcal{X} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times d}$ has full row rank. Then the above regression model has at least one global minimum with zero loss.

Target function. Suppose the training data is generated from an underlying function $g : \mathbb{R}^d \to \mathbb{R}$ with $y_i = g(x_i),\ \forall i \in [n]$. Let $\mathcal{Y} = (y_1, \cdots, y_n)^\top$, and assume that there exits $c^* > 0$ such that

$$\|\mathcal{Y}\| \le c^* \sqrt{\lambda_{\max}(\mathcal{X}\mathcal{X}^\top)}. \tag{6}$$

Functions satisfying (6) include linear and non-linear functions. For example, $g(x) = \phi(x^\top w^*)$, where $w^* \in \mathbb{R}^d$ with $\|w^*\|_2 \le c^*$ for some constant c^*, and $\phi(\cdot)$ is Lipschitz with $\phi(0) = 0$.

Assumption 1. *The entries of $\mathcal{X}$ are i.i.d. subgaussian random variables with zero mean, unit variance, and subgaussian moments*[3] *bounded by* 1.

This assumption allows us to study the spectral properties of the sample matrix by random matrix theory. Especially, when $n(d)/d$ converges to some constant $\gamma \in (0, 1)$, the Marchenko–Pastur law (Marčenko & Pastur, 1967) shows that $\lambda_{\min}(\mathcal{X}\mathcal{X}^\top)/n$ converges to $(1 - \sqrt{\gamma})^2$ almost surely. The non-asymptotic results are provided in (Rudelson & Vershynin, 2010).

Performing our three-step analysis, we get the following results:

Theorem 3. *Consider the undertermined ℓ_p linear regression model (5). Suppose that there exists a universal constant $c_0 \ge 1$ such that $\forall d$, $\left\|w^{(0)}\right\|_2 \le c_0$.*

Step 1. *$\mathcal{L}_n(w)$ satisfies the Uniform-LGI globally on $\mathbb{R}^d$ with*

$$c_n = p^{1-1/p}\sqrt{\frac{\lambda_{\min}(\mathcal{X}\mathcal{X}^\top)}{n}}, \quad \theta_n = 1 - 1/p.$$

Step 2. *$\mathcal{L}_n(w^{(t)})$ converges to zero linearly for $p = 2$ and sublinearly for $p \ge 4$, i.e.,*

$$\begin{aligned} p = 2, &\quad \mathcal{L}_n(w^{(t)}) \le \exp\left(-2\lambda_{\min}(\mathcal{X}\mathcal{X}^\top)t/n\right)\mathcal{L}_n(w^{(0)}); \\ p \ge 4, &\quad \mathcal{L}_n(w^{(t)}) \le (1 + Mt)^{-\frac{p}{p-2}}\,\mathcal{L}_n(w^{(0)}), \end{aligned}$$

where $M = p^{1-\frac{2}{p}}(p-2)\frac{\lambda_{\min}(\mathcal{X}\mathcal{X}^\top)}{n}\mathcal{L}_n(w^{(0)})^{1-\frac{2}{p}}$.

Step 3. *Under Assumption 1, for any target function that satisfies (6), we have with probability at least $1 - \delta - \tau^{d-n+1} - \tau^d$ over the samples,*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\tilde{\ell}\left(f(w^{(\infty)}, x), y\right)\right] \le \mathcal{O}\left(n^{-1/p}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $\tau \in (0, 1)$ depends only on the subgaussian moment of the entries.

The proof of Theorem 3 is given in Appendix D.1.

This theorem shows that compared to the PL condition that corresponds to $p = 2$, the uniform-LGI is more general and can be applied to more cases.

Comparison. This result is related to (Bartlett et al., 2020) that studied the phenomenon of benign overfitting in high-dimensional ℓ_2 linear regression. Our result coincides with theirs as both results uncover some scenarios for benign overfitting in linear regression. In particular, (Bartlett et al., 2020) focused on the minimum ℓ_2 norm estimator. They showed that, if the eigenvalue sequence of the covariance operator $\Sigma := \mathbb{E}[xx^\top]$ have suitable decay rates, then the generalization error will decrease to zero as n increases. In contrast, our result differs from theirs in the problem settings. Specifically, we do not assume the minimum norm property and consider the optimization process that is neglected in (Bartlett et al., 2020). Also, we evaluate the generalization error of the convergence point on a globally Lipschitz loss function. In our settings, entries of each input x are i.i.d. random

[3]The subgaussian moment of X is defined as $\inf\left\{\mathcal{M} \ge 0 \mid \mathbb{E}e^{tX} \le e^{\mathcal{M}^2t^2/2},\ \forall t \in \mathbb{R}\right\}$.

variables, meaning that $\Sigma = \mathbb{I}_{d\times d}$ without decaying eigenvalues. The requirement of i.i.d. entries for each input is a limitation due to the lack of non-asymptotic results of the spectral properties for random matrices with non-i.i.d. entries. However, this limitation can be overcome if one considers the asymptotic results. Moreover, our result works for ℓ_p loss functions with any even positive integer p. This expands the understanding of benign overfitting in the high-dimensional setting.

4.2 Kernel Regression

Consider a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with its corresponding feature map $\varphi : \mathbb{R}^d \to \mathcal{F}$ satisfying $\langle \varphi(x), \varphi(y)\rangle_{\mathcal{F}} = k(x, y)$. We assume that $|k(x, x)| \leq 1, \forall x \in \mathcal{X}$. Let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) with respect to k. If $\mathcal{F} = \mathbb{R}^s$, then the kernel regression model is to solve the following problem

$$\arg\min_{w\in\mathbb{R}^s} \mathcal{L}_n(w) := \frac{1}{2n}\sum_{i=1}^{n}\left(w^\top\varphi(x_i) - y_i\right)^2. \tag{7}$$

Similar to the ℓ_2 linear regression case, we consider the following target function class:

Target function. Suppose the training data is generated from an underlying function $g : \mathbb{R}^d \to \mathbb{R}$ with $y_i = g(x_i),\ \forall i \in [n]$. We further assume that there exits a constant $c^* > 0$ such that

$$\|\mathcal{Y}\| \leq c^* \cdot \sqrt{\lambda_{\max}(k(\mathcal{X}, \mathcal{X}))}, \tag{8}$$

where $k(\mathcal{X}, \mathcal{X})$ is the $n \times n$ kernel matrix on $\mathcal{X}$ with $k(\mathcal{X}, \mathcal{X})_{ij} = k(x_i, x_j)$.

Here, we list an example of class of functions that satisfies (8): $g(x) = \phi(\varphi(x)^\top w^*)$ where $w^* \in \mathbb{R}^s$ with $(\forall s)\ \|w^*\|_2 \leq c^*$ for some constant c^*, and $\phi(\cdot)$ is Lipschitz with $\phi(0) = 0$.

To get the generalization results of kernel regression, we will discuss two types of kernels separately: radial basis function (RBF) kernel and inner product kernel.

RBF kernel. We study the RBF kernel of the form $k(x, y) = \varrho\left(\|y - x\|\right)$ for a certain RBF ϱ. For the input data, we define the *separation distance* of $\mathcal{X}$ as

$$\mathsf{SD} := \frac{1}{2}\min_{i\neq j}\|x_i - x_j\|, \quad \forall i, j \in [n].$$

Inner product kernel. For the inner product kernel, we consider $k(x, y) = \varrho\left(\frac{x^\top y}{d}\right)$.

Following (El Karoui et al., 2010), we make the following assumption on the function ϱ:

Assumption 2. *ϱ is C^3 in a neighborhood of 0 with $\varrho(0) = 0$, $\varrho(1) > \varrho'(0) \geq 0$, $\varrho''(0) \geq 0$.*

Now we are ready to apply our three-step analysis to get optimization and generalization results. For the RBF kernel, the generalization result depends on the separation distance of the samples. For the inner product kernel, we study the high-dimensional random kernel matrix.

Theorem 4. *Consider the kernel regression model (7). Suppose that there exists a universal constant $c_0 \geq 1$ such that $\forall s$, $\left\|w^{(0)}\right\|_2 \leq c_0$.*

Step 1. *$\mathcal{L}_n(w)$ satisfies the Uniform-LGI globally on $\mathbb{R}^s$ with*

$$c_n = \sqrt{\frac{2\lambda_{\min}(k(\mathcal{X}, \mathcal{X}))}{n}}, \quad \theta_n = 1/2,$$

where c_n is controlled by the kernel and input samples.

Step 2. *$\mathcal{L}_n(w^{(t)})$ converges to zero linearly:*

$$\mathcal{L}_n(w^{(t)}) \leq \exp\left(-2\lambda_{\min}(k(\mathcal{X}, \mathcal{X}))t/n\right) \cdot \mathcal{L}_n(w^{(0)}).$$

Step 3. *For any target function that satisfies (8) we have:*

- *For the RBF kernel*[4]*, suppose that* $\varrho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *is a decreasing function and* $\varrho(\|x\|) \in L^1(\mathbb{R}^d)$*. If there exists two positive constants* $q_{\min}$ *and* $q_{\max}$ *such that* $\mathsf{SD} \in [q_{\min}, q_{\max}]$ *for all* n*, then with probability at least* $1-\delta$ *over the samples,*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\tilde{\ell}\left(f(w^{(\infty)}, x), y\right)\right] \leq \mathcal{O}\left(n^{-1/2}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

- *For the inner product kernel, under Assumption 1 and Assumption 2, if* d *is large enough and* $\delta > 0$ *is small enough such that* $d^{-1/2}\left(\sqrt{3}\delta^{-1/2} + \log^{0.51} d\right) \leq 0.5(\varrho(1) - \varrho'(0))$*, then with probability at least* $1 - \delta - d^{-2}$ *over the samples, we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\tilde{\ell}\left(f(w^{(\infty)}, x), y\right)\right] \leq \mathcal{O}\left(n^{-1/2}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

The proof of Theorem 4 is given in Appendix D.2.

Example 1. *RBF kernels satisfying the conditions and assumptions in Theorem 4 include*

- *Gaussian:* $\varrho(r) = e^{-\rho r^2},\ \rho > 0$*;*
- *(inverse) Multiquadrics:* $\varrho(r) = (\rho + r^2)^{\beta/2},\ \rho > 0,\ \beta \in \mathbb{R}\backslash 2\mathbb{N},\ \beta < -d$*.*

Inner product kernels satisfying the conditions and assumptions in Theorem 4 include

- *Polynomial kernel:* $\varrho(r) = r^\beta,\ \beta \in \mathbb{Z}^+,\ \beta \geq 2$*;*
- *NTK corresponding to Two-layer ReLU neural networks on* $\mathbb{S}^{d-1}(\sqrt{d})$*:* $\varrho(r) = \frac{r(\pi - \arccos(r))}{2\pi}$*.*

Comparison. This result of the inner product kernel is related to (Liang & Rakhlin, 2020) who derived generalization bounds for the minimum RKHS norm estimator. They showed that when the data covariance matrix and the kernel matrix enjoy certain decay of the eigenvalues, the generalization bound vanishes as n goes to infinity. For example, for exponential kernel and the covariance matrix $\Sigma := \mathbb{E}[xx^\top]$ with the j-th eigenvalue $\lambda_j(\Sigma) = j^{-\alpha}$, the ℓ_2 generalization bound becomes $\mathcal{O}(n^{-\frac{\alpha}{2\alpha+1}})$ when $\alpha \in (0,1)$ and $n > d$. In comparison, we get an optimal generalization bound $\mathcal{O}\left(n^{-1/2}\right)$ for identity covariance matrix on a globally Lipschitz loss function. Again, we emphasize that the i.i.d. assumption can be relaxed if more new results of random matrix theory are available. Further, we extend the works in (Liang & Rakhlin, 2020) by proving a new result of the RBF kernel. Note that the result of the RBF kernel is not under the high-dimensional setting; thus it is not a direct adaptation of (Liang & Rakhlin, 2020), and the proof itself is of independent interest.

4.3 Overparameterized two-layer Neural Networks

The first two applications are on traditional machine learning models. Indeed, our framework can be applied not only to linear/kernel regression but also to neural network models. In this subsection, we use our framework to study shallow neural networks in an overparameterization regime.

First, define a two-layer ReLU neural network under the *NTK parameterization* (Lee et al., 2019):

$$f(w, x) = \frac{1}{\sqrt{m}} w_2^\top \phi(W_1 x),$$

where $\phi(x) = \max\{0, x\}$, $x \in \mathbb{R}^d$ is the input, $W_1 \in \mathbb{R}^{m\times d}$, $w_2 \in \mathbb{R}^m$ are parameters, $w = \mathrm{vec}\left(\{W_1, w_2\}\right) \in \mathbb{R}^{m(d+1)}$, and m is the width (number of hidden units).

We consider minimizing the quadratic loss by gradient flow:

$$\arg\min_w \mathcal{L}_n(w) := \frac{1}{2n}\sum_{i=1}^n (f(w, x_i) - y_i)^2. \tag{9}$$

[4]Here d is fixed and n is varied.

Random initialization. $W_1^{(0)}$ is drawn from Gaussian $\mathcal{N}(0, \frac{1}{d}\mathbb{I}_{m\times d})$ and $w_{2,i}^{(0)}$ are drawn i.i.d. from uniform distribution $U\{-1,1\}, \forall i \in [m]$.

Following the setting in (Du et al., 2019), we only train the hidden layer W_1 and leave the output layer w_2 as random initialization to simplify analyses.

NTK matrix. The NTK matrix $\Theta(t)$ is defined as: $\Theta_{ij}(t) = \left\langle \nabla_{W_1} f(w^{(t)}, x_i), \nabla_{W_1} f(w^{(t)}, x_j) \right\rangle$, and denote $\widehat{\Theta}$ by the limiting matrix[5]: $\widehat{\Theta}_{ij} = x_i^\top x_j \mathbb{E}_{w\sim\mathcal{N}(0,\frac{1}{d}\mathbb{I}_d)} \left[\phi'(w^\top x_i)\phi'(w^\top x_j)\right], \forall i,j \in [n]$.

Similarly, we apply our three-step analysis to derive the following results.

Theorem 5. *Consider the two-layer ReLU neural network model (9). For any $\delta \in (0,1)$, if $m \geq \mathrm{poly}\left(n, \lambda_{\min}^{-1}(\widehat{\Theta}), \delta^{-1}\right)$, then*

Step 1. *With probability at least $1-\delta$ over training samples and random initialization, $\mathcal{L}_n(w^{(t)})$ satisfies the Uniform-LGI on $\left\{w^{(t)} : t \geq 0\right\}$ with*

$$c_n = \sqrt{\lambda_{\min}(\widehat{\Theta})/n}, \quad \theta_n = 1/2.$$

Step 2. *$\mathcal{L}_n(w^{(t)})$ converges to zero with a linear convergence rate:*

$$\mathcal{L}_n(w^{(t)}) \leq \exp\left(-\lambda_{\min}(\widehat{\Theta})t/n\right)\mathcal{L}_n(w^{(0)}).$$

Step 3. *Under Assumption 1, for any target function that satisfies (6), if $\gamma_1 \in (0,1)$, then with probability at least $1-\delta-\tau^{d-n+1}-\tau^d$ over the samples and random initialization,*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\tilde{\ell}\left(f(w^{(\infty)}, x), y\right)\right] \leq \mathcal{O}\left(\sqrt{\frac{\log(n/\delta)}{n}}\right),$$

where $\tau \in (0,1)$ depends only on the subgaussian moment of the entries.

The proof of Theorem 5 is deferred to Appendix D.3.

Comparison. Our result is related to (Arora et al., 2019), which gave an NTK-based generalization bound for overparameterized two-layer ReLU neural networks. This result matches with theirs in the sense that we discover some underlying functions that are provably learnable. Examples of learnable target functions in (Arora et al., 2019) include polynomials $y = (\beta^\top x)^p$, non-linear activations $y = \cos(\beta^\top x) - 1$, $y = \tilde{\phi}(\beta^\top x)$ with $\tilde{\phi}(z) = z \cdot \arctan(z/2)$, $\|\beta\| \leq 1$. Our result, furthermore, expands the target function class that is provably learnable since we only require $\tilde{\phi}$ to be Lipshcitz. In addition, they set the standard deviation of the initialization to be at most $\mathcal{O}(1/n)$, whereas we use a different initialization with order $\mathcal{O}(1/\sqrt{d})$ that is more often applied in practice. Our result is also related to (Liu et al., 2020), which proved that overparameterized deep neural networks satisfy the PL condition. Further, we extend this work by analyzing the generalization based on the length of optimization trajectories.

5 Conclusion

In this work, we address when and why does an optimization algorithm finds a minimum with good generalization performance. For this problem, we propose a framework to bridge optimization and generalization based on the trajectory length. The pivotal component is the Uniform-LGI property: a condition on the loss function, by which we show that gradient flow returns a global minimum with an explicit length estimate. Further, we give a length-based generalization bound given such a length estimate. Finally, we apply the framework to three machine learning models with certain target functions. By estimating the trajectory lengths, we get non-vacuous generalization bounds that do not suffer from the curse of dimensionality. This framework is not a direct variant of the NTK method, and the results show that our framework is favorable for inducing connections between optimization and generalization.

[5]Here $\lambda_{\min}(\widehat{\Theta})$ changes with n.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6158–6169, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3349–3356, 2020.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep re{lu} networks? In *International Conference on Learning Representations*, 2021.

Benedikt Diederichs and Armin Iske. Improved estimates for condition numbers of radial basis function interpolation matrices. *Journal of Approximation Theory*, 238:38–51, 2019.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Noureddine El Karoui et al. The spectrum of kernel random matrices. *Annals of statistics*, 38(1): 1–50, 2010.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pp. 2683–2711. PMLR, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *arXiv preprint arXiv:2003.00307*, 2020.

S. Lojasiewicz. Ensembles semi-analytiques. *IHES notes*, 1965.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, apr 1967.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Stephen J Montgomery-Smith. The distribution of rademacher sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990.

Mor Shpigel Nacson, Suriya Gunasekar, Jason D Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. *arXiv preprint arXiv:1905.07325*, 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*, pp. 3051–3059. PMLR, 2019b.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

B. Polyak. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3:864–878, 1963.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. Pac-bayes analysis beyond the usual bounds. *arXiv preprint arXiv:2006.13057*, 2020.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pp. 1576–1602. World Scientific, 2010.

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, 2018a.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018b.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach. In *International Conference on Learning Representations*, 2019.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.