ON THE MARGIN THEORY OF FEEDFORWARD NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

Abstract

Past works have shown that, somewhat surprisingly, over-parametrization can help generalization in neural networks. Towards explaining this phenomenon, we adopt a margin-based perspective. We establish: 1) for multi-layer feedforward relu networks, the global minimizer of a weakly-regularized cross-entropy loss has the maximum normalized margin among all networks, 2) as a result, increasing the over-parametrization improves the normalized margin and generalization error bounds for two-layer networks. In particular, an infinite-size neural network enjoys the best generalization guarantees. The typical infinite feature methods are kernel methods; we compare the neural net margin with that of kernel methods and construct natural instances where kernel methods have much weaker generalization guarantees. We validate this gap between the two approaches empirically. Finally, this infinite-neuron viewpoint is also fruitful for analyzing optimization. We show that a perturbed gradient flow on infinite-size networks finds a global optimizer in polynomial time.

1 INTRODUCTION

In deep learning, over-parametrization refers to the widely-adopted technique of using more parameters than necessary (Krizhevsky et al., 2012; Livni et al., 2014). Both computationally and statistically, over-parametrization is crucial for learning neural nets. Controlled experiments demonstrate that over-parametrization eases optimization by smoothing the non-convex loss surface (Livni et al., 2014; Sagun et al., 2017). Statistically, increasing model size without any regularization still improves generalization even after the model interpolates the data perfectly (Neyshabur et al., 2017b). This is surprising given the conventional wisdom on the trade-off between model capacity and generalization.

In the absence of an explicit regularizer, algorithmic regularization is likely the key contributor to good generalization. Recent works have shown that gradient descent finds the minimum norm solution fitting the data for problems including logistic regression, linearized neural networks, and matrix factorization (Soudry et al., 2018; Gunasekar et al., 2018b; Li et al., 2018; Gunasekar et al., 2018a; Ji & Telgarsky, 2018). Many of these proofs require a delicate analysis of the algorithm’s dynamics, and some are not fully rigorous due to assumptions on the iterates. To the best of our knowledge, it is an open question to prove analogous results for even two-layer relu networks. (For example, the technique of Li et al. (2018) on two-layer neural nets with quadratic activations still falls within the realm of linear algebraic tools, which apparently do not suffice for other activations.)

We propose a different route towards understanding generalization: **making the regularization explicit.** The motivations are: 1) with an explicit regularizer, we can analyze generalization without fully understanding optimization; 2) it is unknown whether gradient descent provides additional implicit regularization beyond what \( \ell_2 \) regularization already offers; 3) on the other hand, with a sufficiently weak \( \ell_2 \) regularizer, we can prove stronger results that apply to multi-layer neural nets with relu activations. Additionally, explicit regularization is perhaps more relevant because \( \ell_2 \) regularization is typically used in practice.

Concretely, we add a norm-based regularizer to the cross entropy loss of a multi-layer feedforward neural network with relu activations. We show that the global minimizer of the regularized objective achieves the maximum normalized margin among all the models with the same architecture, if the regularizer is sufficiently weak (Theorem 2.1). Informally, for models with norm 1 that perfectly classify the data, the margin is the smallest difference across all datapoints between the classifier
score for the true label and the next best score. We are interested in normalized margin because its inverse bounds the generalization error (see recent work [Bartlett et al., 2017; Neyshabur et al., 2017a, 2018] and our Theorem 3.1). Our work explains why optimizing the training loss can lead to parameters with a large margin and thus, better generalization error.

At a first glance, it might seem counterintuitive that decreasing the regularizer is the right approach. At a high level, we show that the regularizer only serves as a tiebreaker to steer the model towards choosing the largest normalized margin. Our proofs are simple, oblivious to the optimization procedure, and apply to any norm-based regularizer. We also show that an exact global minimum is unnecessary: if we approximate the minimum loss within a constant, we obtain the max-margin within a constant (Theorem 2.2).

We further study the margin of two-layer networks: let $\gamma^{*\cdot m}$ be the max normalized margin of a neural net with $m$ hidden units (formally defined in Section 3.1). Let $\gamma^{*,\infty} \triangleq \sup_m \gamma^{*\cdot m}$ be the largest possible margin of an infinite two-layer network. We will show three properties of the margins:

1. In Theorem 3.2 we show that the optimal normalized margin of two-layer networks is non-decreasing as the width of the architecture grows, so the generalization error bound only improves with a wider network. Thus, even if the dataset is already separable, it could still be useful to increase the width to achieve larger margin and better generalization. More formally, let $n$ be the number of training examples. We additionally approach the maximum possible margin $\gamma^{*,\infty}$ after over-parameterizing with $m \geq n$ neurons: $\forall m \geq n, \gamma^{*\cdot m} = \gamma^{*,\infty}$.

2. The max-margin of infinite-size nets, $\gamma^{*,\infty}$, equals half the margin of the $\ell_1$-norm SVM (Zhu et al., 2004) over the lifted feature space defined by the activation function applied to all possible hidden units. (See Theorem 3.3.)

3. We compare the neural net margin $\gamma^{*,\infty}$ to the standard margin for the kernel SVM on the same features. We design a simple data distribution (Figure 1) where neural net margin $\gamma^{*,\infty}$ is large but the kernel margin is small. This translates to an $\Omega(\sqrt{d})$ factor gap between the generalization error bounds for the two approaches and demonstrates the power of neural nets compared to kernel methods. We experimentally confirm that a gap does indeed exist.

In the context of bullet 2, our work is closely related to that of [Rosset et al., 2007] and [Neyshabur et al., 2014], who show that optimizing the loss over the parameters of a two-layer relu network is equivalent to optimizing the loss of a “convex neural net” parametrized by a distribution over hidden units. We go one step further and connect the weakly regularized training loss to the $\ell_1$ SVM.

We will also adopt this view of infinite-size neural networks to study how over-parametrization helps optimization. Prior works [Mei et al., 2018; Chizat & Bach, 2018] show that gradient descent on two-layer networks becomes Wasserstein gradient flow over parameter distributions in the limit of infinite neurons. For this setting, we prove that perturbed Wasserstein gradient flow finds a global optimizer in polynomial time.

Finally, we empirically validate several of the claims made in this paper. First, we train a two-layer network on a one-dimensional classification task that is simple to visualize. In one dimension, it is possible to brute-force approximate the maximum neural network margin and we show that training with a progressively smaller regularizer results in convergence to this margin. Second, we compare the generalization performance of neural networks and kernel methods and confirm that neural networks do achieve better generalization, as our theory predicts.

1.1 Additional Related Work

Zhang et al. (2016) and Neyshabur et al. (2017b) show that neural network generalization defies conventional explanations and requires new ones. One proposed explanation is the inductive bias of the training algorithm. Recent papers [Hardt et al., 2015; Brutzkus et al., 2017; Chaudhari et al., 2016] study inductive bias through training time and sharpness of local minima. Neyshabur et al. (2015a) propose a new steepest descent algorithm in a geometry invariant to weight rescaling and show that this improves generalization. Morcos et al. (2018) relate generalization in deep nets to the number of “directions” in the neurons. Other papers [Gunasekar et al., 2017; Soudry et al., 2018; Gunasekar et al., 2018b; Li et al., 2018; Gunasekar et al., 2018a] study implicit regularization towards a specific solution. Ma et al. (2017) show that implicit regularization can help gradient descent avoid
We will use the notation $\| \cdot \|$ when applicable, as a matrix, and interpret over-parametrization as a means of implicit acceleration during optimization. Mei (2018) investigate the Rademacher complexity of two-layer networks. Liang & Rakhlin (2018) and (Neyshabur et al., 2015b; Bartlett et al., 2017; Neyshabur et al., 2017a; Golowich et al., 2017) and throughout this paper, we reserve the symbol $\| \cdot \|$ for some general norm $a > 0$ prediction functions $f$. As a concrete example, feedforward relu networks are positive-homogeneous. For a positive-homogeneous prediction function, the normalized margin of the optimum converges to some $\lambda$ and $\alpha$-th power of the absolute value is Lebesgue integrable. For $\alpha$, $\alpha_2 \in L^2(S^{d-1})$, we can define $\langle \alpha_1, \alpha_2 \rangle \triangleq \int_{S^{d-1}} \alpha_1(\vec{u})\alpha_2(\vec{u})d\vec{u} < \infty$. Furthermore, we will use $\text{Vol}(S^{d-1}) \triangleq \int_{S^{d-1}} 1d\vec{u}$. Throughout this paper, we reserve the symbol $X = [x_1, \ldots, x_n]$ to denote the collection of datapoints (as a matrix), and $Y = [y_1, \ldots, y_n]$ to denote labels. We use $d$ to denote the dimension of our data. We often use $\Theta$ to denote the parameters of a prediction function $f$, and $f(\Theta; x)$ to denote the prediction of $f$ on datapoint $x$.

We will use the notation $\lesssim, \gtrsim$ to mean less than or greater than up to a universal constant, respectively. Unless stated otherwise, we use $O(\cdot), \Omega(\cdot)$ as placeholders for some universal constant in upper and lower bounds, respectively. We will use poly to denote some universal constant-degree polynomial in the arguments.

2 Weak Regularizer Guarantees Max Margin Solutions

In this section, we will show that when we add a weak regularizer to cross-entropy loss with a positive-homogeneous prediction function, the normalized margin of the optimum converges to some max-margin solution. As a concrete example, feedforward relu networks are positive-homogeneous.

Let $l$ be the number of labels, so the $i$-th example has label $y_i \in [l]$. We work with a family $\mathcal{F}$ of prediction functions $f(\Theta; \cdot) : \mathbb{R}^d \to \mathbb{R}^l$ that are $a$-positive-homogeneous in their parameters for some $a > 0$: $f(c \Theta; x) = c^a f(\Theta; x), \forall c > 0$. We additionally require that $f$ is continuous in $\Theta$. For some general norm $\| \cdot \|$, we study the $\lambda$-regularized cross-entropy loss $L_\lambda$, defined as

$$L_\lambda(\Theta) \triangleq \sum_{i=1}^n - \log \frac{\exp(f_{y_i}(\Theta; x_i))}{\sum_{j=1}^l \exp(f_j(\Theta; x_i))} + \lambda \|\Theta\|^r \quad (2.1)$$
for fixed $r > 0$. Let $\Theta_\lambda \in \arg \min L_\lambda(\Theta)$. We define the normalized margin of $\Theta_\lambda$ as:

$$\gamma_\lambda \triangleq \min_i \left( f_{y_i}(\bar{\Theta}_\lambda; x_i) - \max_{j \neq y_i} f_j(\bar{\Theta}_\lambda; x_i) \right)$$

(2.2)

Define the $\| \cdot \|_{\text{max}}$-max normalized margin as

$$\gamma^* \triangleq \max_{\| \Theta \| \leq 1} \left[ \min_i \left( f_{y_i}(\Theta; x_i) - \max_{j \neq y_i} f_j(\Theta; x_i) \right) \right]$$

and let $\Theta^*$ be a parameter achieving this maximum. We show that with sufficiently small regularization level $\lambda$, the normalized margin $\gamma_\lambda$ approaches the maximum margin $\gamma^*$.

**Theorem 2.1.** Assume the training data is separable by a network $f(\Theta^*; \cdot) \in F$ with an optimal normalized margin $\gamma^* > 0$. Then, the normalized margin of the global optimum of the weakly-regularized objective (equation 2.1) converges to $\gamma^*$ as the strength of the regularizer goes to zero. Mathematically, let $\gamma_\lambda$ be defined in equation 2.2. Then

$$\gamma_\lambda \to \gamma^*$$

as $\lambda \to 0$.

An intuitive explanation for our result is as follows: because of the homogeneity, the loss $L_\lambda(\Theta)$ roughly satisfies the following (for small $\lambda$, and ignoring problem parameters such as $n$):

$$L_\lambda(\Theta) \approx \exp(-\|\Theta\|\alpha \gamma_\lambda) + \lambda \|\Theta\|^r$$

Thus, the loss focuses on choosing parameters with larger margin, and the regularization term biases the loss to select parameters with a smaller norm. The full proof of the theorem is deferred to Section A.1.

We can also provide an analogue of Theorem 2.1 for the binary classification setting. For this setting, our prediction is now a single real output and we train using logistic loss. We provide formal definitions and results in Section A.2. Our theory for two-layer neural networks (see Section 3) is based in this setting.

### 2.1 Optimization Accuracy

Since $L_\lambda$ is typically hard to optimize exactly for neural nets, it would be ideal to relax the condition that $\Theta_\lambda$ minimizes $L_\lambda$. Thus, we ask, how accurately do we need to optimize $L_\lambda$ to obtain a margin that approximates $\gamma^*$ up to a constant? The following theorem shows that if suffices to find $\Theta'$ achieving a constant factor multiplicative approximation of $L_\lambda(\Theta_\lambda)$, where $\lambda$ is some sufficiently small polynomial in $n, l, \gamma^*$. Though our theorem is stated for the general multi-class setting, our result applies for binary classification as well. We provide the proof in Section A.3.

**Theorem 2.2.** In the setting of Theorem 2.1, suppose that we choose $\lambda = \frac{(\gamma^*)^{r/a}}{n^{l-1}}$ for sufficiently large $c$ (that only depends on $r/a$). Let $\Theta'$ denote a $2$-approximate minimizer of $L_\lambda$, so $L_\lambda(\Theta') \leq 2L_\lambda(\Theta_\lambda)$. Denote the normalized margin of $\Theta'$ by $\gamma'$. Then

$$\gamma' \geq \frac{\gamma^*}{2 \cdot 4^{a/r}}$$

### 3 Margins of Over-parameterized Two-layer Homogeneous Neural Nets

In Section 2, we showed that a weakly-regularized logistic loss leads to the maximum normalized margin. In this section, we analyze the properties of the max-margin of neural nets more closely. We will contrast neural networks with kernel methods, for which margins have already been extensively studied. Towards a first-cut understanding, we focus on two-layer networks for binary classification.

---

1 We formally show that $L_\lambda$ has a minimizer in Claim A.1 of Section A.

2 The exact approximation constant is not important, so we choose 2 for simplicity.
First, in Section 3.1 we provide a bound stating that the generalization error is roughly linear in the inverse of the margin, establishing that a larger margin implies better generalization. In Section 3.2 we show that the maximum normalized margin is non-decreasing with the hidden layer size and stays constant as soon as there are more hidden units than data points. This suggests that increasing the size of the network improves the generalization of the solution.

Second, in Section 3.3 we draw an analogy to classical kernel methods by proving that the maximum \( \ell_2 \)-normalized margin of an over-parameterized neural net is equal to half the maximum possible \( \ell_1 \)-normalized margin of linear functionals on a lifted feature space. In other words, we establish an equivalence between neural networks and the 1-norm SVM (Zhu et al., 2004) on the lifted features. These features are constructed by applying the activation function on all possible hidden layer weights.

Third, continuing this analogy, we will compare the generalization power of a two-layer neural network to that of a kernel method on the lifted space. This kernel method corresponds to fixing random weights for the hidden layer and solving a 2-norm max-margin problem on the top layer weights. We demonstrate instances where two layer neural networks give better generalization error guarantees than the kernel method.

### 3.1 Setup and Margin-based Generalization Error

In the rest of the paper, we work with two-layer neural networks with a single output for binary classification. We use \( m \) to denote the number of hidden units, \( w_1, \ldots, w_m \in \mathbb{R}^d \) for the weight vectors on the first layer, and \( u_1, \ldots, u_m \in \mathbb{R} \) for the weights on the second layer. We let \( \theta_j \triangleq (w_j, u_j) \), and we use \( \Theta \) to denote the collection of all the parameters. We assume in this section that the activation \( \phi : \mathbb{R} \rightarrow \mathbb{R} \) is 1-homogeneous and 1-Lipschitz. The network thus computes a single score

\[
    f(\Theta; x) \triangleq \sum_{j=1}^{m} w_j \phi(u_j^\top x)
\]

We consider \( \ell_2 \) regularization from here on. The regularized logistic loss of the architecture with \( m \) hidden units is therefore

\[
    L_{\lambda, m} \triangleq \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i f(\Theta; x_i))) + \lambda \|\Theta\|^2
\]

(3.1)

where \( \|\Theta\| \) denotes the Euclidean norm of all the parameters in \( \Theta \). We note that \( f \) and the regularizer are both 2-homogeneous in \( \Theta \), so the results of Section 2 apply to \( L_{\lambda, m} \).

Following our conventions from Section 2, we denote the optimizer of \( L_{\lambda, m} \) by \( \Theta_{\lambda, m} \), the normalized margin of \( \Theta_{\lambda, m} \) by \( \gamma_{\lambda, m} \), the max-margin solution by \( \Theta^{*, m} \), and the max-margin by \( \gamma^{*, m} \). We emphasize the size of the network in our notation. Since our classifier \( f \) now predicts a single real value, we need to redefine

\[
    \gamma_{\lambda, m} \triangleq \min_i y_i f(\Theta_{\lambda, m}; x_i)
\]

\[
    \gamma^{*, m} \triangleq \max_{\|\Theta\| \leq 1} \min_i y_i f(\Theta; x_i)
\]

When the data is not separable by a \( m \)-unit neural net, \( \gamma^{*, m} \) is zero by definition.

Recall that \( X = [x_1, \ldots, x_n] \) denotes the matrix with all the data points as columns, and \( Y = [y_1, \ldots, y_n] \) denotes the labels. We sample \( X \) and \( Y \) i.i.d. from the data generating distribution \( p_{\text{data}} \), which is supported on \( \mathcal{X} \times \{-1, +1\} \). We can define the population 0-1 loss and the training 0-1 loss of the network \( \Theta \) as

\[
    L(\Theta) = \Pr_{(x, y) \sim p_{\text{data}}} [y f(\Theta; x) \leq 0]
\]

We will let \( D \triangleq \frac{\|X\|_F^2}{n} \) be the average norm squared of the data and \( C \triangleq \sup_{x \in \mathcal{X}} \|x\|_2 \) be an upper bound on the norm of a single datapoint. The following theorem shows that the generalization error only depends on the parameters through the inverse of the margin on the training data. We provide a proof in Section C.1.

\[\text{\footnotesize{[Although Theorem 2.1 is written in the language of multi-class prediction where the classifier outputs \( l \geq 2 \) scores, the results translate to single-output binary classification. See Section A.2.]}}\]
Theorem 3.1. Suppose \( \phi \) is 1-Lipschitz and 1-homogeneous. Then for any \( \Theta \) that separates the data with margin \( \gamma \triangleq \min_i y_i f(\Theta; x_i) > 0 \), with probability at least \( 1 - \delta \) over the draw of \( X, Y \),

\[
L(\Theta) \leq \frac{6 \sqrt{D}}{n} + \epsilon(\gamma)
\]

(3.2)

where \( \epsilon(\gamma) \triangleq \sqrt{\frac{\log\log D}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \). Note that \( \epsilon(\gamma) \) is typically small, and thus the above bound mainly scales with \( \frac{1}{\sqrt{n}} \). As a corollary, with probability \( 1 - \delta \),

\[
\lim_{\lambda \to 0} L(\Theta_{\lambda,m}) \leq \frac{6 \gamma_{\star,m}}{n} + \epsilon(\gamma_{\star,m})
\]

(3.3)

Above we implicitly assume \( \gamma_{\star,m} > 0 \), since otherwise the right hand side of the bound is vacuous.

One consequence of the above theorem and Theorem 2.2 is that if \( \lambda \) is polynomially small in \( \gamma_{\star,m} \) and \( n \), we only need to optimize \( L_{\lambda,m} \) up to a constant multiplicative factor to obtain parameters with generalization bounds roughly as good as those for \( \Theta^{\star,m} \).

3.2 The Max Margin Is Non-decreasing in the Hidden Layer Size

Now we show that the maximum normalized margin is non-decreasing with the hidden layer size and stays constant once we have more hidden units than examples.

Theorem 3.2. In the setting of Section 3.1 recall that \( \gamma_{\star,m} \) denotes the max normalized margin of a two-layer neural network with hidden layer size \( m \). Then,

\[
\gamma_{1} \leq \gamma_{2} \ldots \leq \gamma_{n} = \gamma_{n+1} = \gamma_{n+2} = \ldots
\]

(3.4)

We note that \( \gamma_{n} \) will be positive when \( \phi \) is a sufficiently powerful activation such as relu or sigmoid and the data points are not repetitive, so the neural network can fit any function of the data. We prove Theorem 3.2 in Section B. Theorem 3.2 can explain why additional over-parametrization has been observed to improve generalization in two-layer networks [Neyshabur et al., 2017b]. Our margin does not decrease with a larger network size, and therefore Theorem 3.1 gives a better generalization bound. We precisely characterize the value of \( \gamma_{n} \) in the following section.

3.3 The Max Margin of Neural Nets is Equivalent to \( \ell_1 \) SVM in Lifted Space

We link infinite-size neural networks to the \( \ell_1 \) SVM over a lifted space, defined via a lifting function \( \varphi : \mathbb{R}^d \to \mathcal{L}^\infty(S^{d-1}) \) mapping data to an infinite feature vector:

\[
x \in \mathbb{R}^d \to \varphi(x) \in \mathcal{L}^\infty(S^{d-1}) \text{ satisfying } \varphi(x)[\bar{u}] = \phi(\bar{u}^\top x)
\]

(3.5)

We look at the margin of linear functionals corresponding to \( \alpha \in \mathcal{L}^1(S^{d-1}) \). The 1-norm SVM over the lifted feature \( \varphi(x) \) solves for the maximum margin:

\[
\gamma_{\ell_1} \triangleq \max_{\alpha} \min_{i \in [n]} y_i \langle \alpha, \varphi(x_i) \rangle
\]

subject to \( \|\alpha\|_1 \leq 1 \)

(3.6)

where we rely on the inner product and 1-norm defined in Section 1.2. A priori, it is unclear how to optimize this since the kernel trick does not work for \( \ell_1 \) norm. Here we will show that optimizing two-layer neural networks with weak regularization is equivalent to solving equation 3.6.

Theorem 3.3. Let \( \gamma_{\ell_1} \) be defined in equation 3.6 and \( \gamma_{\star,m} \) be defined in Section 3.1. For any \( m \geq n \),

\[
\gamma_{\star,m} = \frac{\gamma_{\ell_1}}{2}
\]

(3.7)

Rosset et al. (2007) and Neyshabur et al. (2014) show a similar equivalence, but between a lifted logistic regression problem and equation 3.1. In contrast, the above theorem, proved in Section B, shows the equivalence between equation 3.1 and the 1-norm SVM when the regularizer is small.

3.4 Comparison to Kernel Methods

The quantity \( \lim_{\lambda \to 0} L(\Theta_{\lambda,m}) \) does not necessarily exist, but here we take it to mean \( \forall \delta > 0, \exists \lambda_0 > 0 \) with \( L(\Theta_{\lambda,m}) \) at most the RHS of equation 3.3 plus \( \delta \) for all \( \lambda < \lambda_0 \).
We compare the ℓ₁ SVM margin, attainable by a finite neural network, to the ℓ₂ margin attainable via kernel methods. Following the setup of Section 3.3, we define the kernel problem over α ∈ L²(Sd−1):  

\[
\gamma_{\ell_2} \triangleq \max_{\alpha} \min_{i \in [n]} y_i \langle \alpha, \varphi(x_i) \rangle 
\]

subject to \( \sqrt{\kappa} \| \alpha \|_2 \leq 1 \) \hspace{1cm} (3.8)

where \( \kappa \triangleq \text{Vol}(S^{d-1}) \). (We scale \( \| \alpha \|_2 \) by \( \sqrt{\kappa} \) to make the lemma statement below cleaner.) First, \( \gamma_{\ell_2} \) can be used to obtain a standard upper bound on the generalization error of the kernel SVM. Following the notation of Section 3.1, we will let \( L_{\ell_2-\text{svm}} \) denote the 0-1 population classification error for the optimizer of equation 3.8.

**Lemma 3.4.** In the setting of Theorem 3.1 with probability at least \( 1 - \delta \), the generalization error of the standard kernel SVM with relu feature (defined in equation 3.8) is bounded by

\[
L_{\ell_2-\text{svm}} \lesssim \frac{1}{\gamma_{\ell_2}} \sqrt{\frac{D}{dn}} + \epsilon_{\ell_2} \quad (3.9)
\]

where \( \epsilon_{\ell_2} \triangleq \frac{\log \max \left\{ \log \frac{\sqrt{\kappa} \gamma_{\ell_2}^2}{n}, 2 \right\}}{n} + \sqrt{\frac{\log(1/\delta)}{n}} \) is typically a lower-order term.

The bound above follows from standard techniques (Bartlett & Mendelson, 2002), and we provide a full proof in Section C.1. We construct a data distribution for which this lemma does not give a good bound for kernel methods, but Theorem 3.1 does imply good generalization for two-layer networks.

**Theorem 3.5.** There exists a data distribution \( p_{\text{data}} \) such that the ℓ₁ SVM with relu features has a good margin:

\[
\gamma_{\ell_1} \gtrsim 1
\]

and with probability \( 1 - \delta \) over the choice of i.i.d. samples from \( p_{\text{data}} \), obtains generalization error

\[
L_{\ell_1-\text{svm}} \lesssim \sqrt{\frac{d}{n}} + \epsilon_{\ell_1}
\]

where \( \epsilon_{\ell_1} \triangleq \sqrt{\frac{\log \log(d \log n)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \) is typically a lower order term. Meanwhile, with high probability the ℓ₂ SVM has a small margin:

\[
\gamma_{\ell_2} \lesssim \max \left\{ \sqrt{\frac{\log n}{n}}, 1/d \right\}
\]

and therefore the generalization upper bound from Lemma 3.4 is at least

\[
\Omega \left( \min \left\{ \frac{1}{\log n}, \frac{d}{\sqrt{n}} \right\} \right)
\]

We briefly overview the construction of \( p_{\text{data}} \) here and defer the full proof of Theorem 3.5 to Section D.1.

**Proof sketch for Theorem 3.5.** We base \( p_{\text{data}} \) on the distribution \( D \) of examples \((x, y)\) described below. Here \( e_i \) is the \( i \)-th standard basis vector and we use \( x^\top e_i \) to represent the \( i \)-coordinate of \( x \) (since the subscript is reserved to index training examples).

\[
\begin{bmatrix}
e_1^\top x \\
e_2^\top x \\
\vdots \\
e_{d-2}^\top x \\
e_d^\top x
\end{bmatrix} \sim \mathcal{N}(0, I_{d-2}), \quad \begin{cases} y = +1, & x^\top e_1 = +1, & x^\top e_2 = +1 \quad \text{w/ prob. } 1/4 \\
y = +1, & x^\top e_1 = -1, & x^\top e_2 = -1 \quad \text{w/ prob. } 1/4 \\
y = -1, & x^\top e_1 = +1, & x^\top e_2 = -1 \quad \text{w/ prob. } 1/4 \\
y = -1, & x^\top e_1 = -1, & x^\top e_2 = +1 \quad \text{w/ prob. } 1/4 
\end{cases}
\]
Figure 1 shows samples from $D$ when there are 3 dimensions. From the visualization, it is clear that there is no linear separator for $D$. As Lemma D.1 shows, a relu network with four neurons can fit this relatively complicated decision boundary. On the other hand, for kernel methods, we prove that the symmetries in $D$ induce cancellation in feature space. The following lemmas, proved in Section D.1, formalize this cancellation and show that it results in a small margin for kernel methods.

Lemma 3.6 (Margin upper bound tool). In the setting of Theorem 3.5, we have

$$\gamma_{\ell_2} \leq \frac{1}{\sqrt{\kappa}} \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)y_i \right\|_2$$

Lemma 3.7. In the setting of Theorem 3.5, let $(x_i, y_i)_{i=1}^{n}$ be n i.i.d samples and corresponding labels from $D$. Let $\varphi$ be defined in equation 3.5 with $\phi = \text{relu}$. With high probability (at least $1 - dn^{-10}$), we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)y_i \right\|_2 \leq \sqrt{\kappa/n} \log n + \sqrt{\kappa/d}$$

Combining these lemmas gives us the desired bound on $\gamma_{\ell_2}$.

Gap in regression setting: We are able to prove an even larger $\Omega(\sqrt{n/d})$ gap between neural networks and kernel methods in the regression setting where we wish to interpolate continuous labels. Analogously to the classification setting, optimizing a regularized squared error loss on neural networks is equivalent to solving a minimum 1-norm regression problem (see Theorem D.3). Furthermore, kernel methods correspond to a minimum 2-norm problem. We construct distributions $p_{\text{data}}$ where the 1-norm solution will have a generalization error bound of $O(\sqrt{d/n})$, whereas the 2-norm solution will have a generalization error bound that is $\Omega(1)$ and thus vacuous. In Section D.2, we define the 1-norm and 2-norm regression problems. In Theorem D.6, we formalize our construction.

4 Perturbed Wasserstein gradient flow finds global optimizers in polynomial time

In the prior section, we studied the limiting behavior of the generalization of a two-layer network as its width goes to infinity. In this section, we will now study the limiting behavior of the optimization algorithm, gradient descent. Prior work [Mei et al., 2018; Chizat & Bach, 2018] has shown that as the hidden layer size grows to infinity, gradient descent for a finite neural network approaches the Wasserstein gradient flow over distributions of hidden units (defined in equation 4.1). Chizat & Bach (2018) also prove that Wasserstein gradient flow converges to a global optimizer in this setting but do not specify a convergence rate.

We show that a perturbed version of Wasserstein gradient flow converges in polynomial time. The informal take-away of this section is that a perturbed version of gradient descent converges in polynomial time on infinite-size neural networks (for the right notion of infinite-size.)

Formally, we optimize the following functional over distributions $\rho$ on $\mathbb{R}^{d+1}$:

$$L[\rho] \triangleq R \left( \int \Phi d\rho \right) + \int Vd\rho$$

where $\Phi : \mathbb{R}^{d+1} \to \mathbb{R}^k$, $R : \mathbb{R}^k \to \mathbb{R}$, and $V : \mathbb{R}^{d+1} \to \mathbb{R}$. In this work, we consider 2-homogeneous $\Phi$ and $V$. We will additionally require that $R$ is nonnegative and $V$ is positive on the unit sphere. Finally, we need standard regularity assumptions on $R$, $\Phi$, and $V$:

Assumption 4.1 (Regularity conditions on $\Phi, R, V$). $\Phi$ and $V$ are differentiable as well as upper bounded and Lipschitz on the unit sphere. $R$ is Lipschitz and its Hessian has bounded operator norm.

We provide more details on the specific parameters (for boundedness, Lipschitzness, etc.) in Section E.1. We note that relu networks satisfy every condition but differentiability of $\Phi$. We can fit a neural network under our framework as follows:

The relu activation is non-differentiable at 0 and hence the gradient flow is not well-defined. Chizat & Bach (2018) acknowledge this same difficulty with relu.
We first verify the normalized margin convergence on a two-layer network with one-dimensional input. A single hidden unit computes the following: \( x \mapsto a_j \text{relu}(w_j x + b_j) \). We add \( \| \cdot \|_2^2 \) regularization to \( a, w, b \) and compare the resulting normalized margin to that of an approximate solution of the \( \ell_1 \) SVM problem with features \( \text{relu}(wx_\star + b) \) for \( w^2 + b^2 = 1 \). Writing this feature vector is intractable, so we solve an approximate version by choosing 1000 evenly spaced values of \( \{w, b\} \). Our theory predicts that with decreasing regularization, the margin of the neural network converges to the \( \ell_1 \) SVM objective. In Figure 2, we plot this margin convergence and visualize the final networks and ground truth labels. The network margin approaches the ideal one as \( \lambda \to 0 \), and the visualization shows that the network and \( \ell_1 \) SVM functions are extremely similar.

Next, we experiment on synthetic data in a higher-dimensional setting. For classification and regression, we compare the generalization error and predicted generalization upper bounds\(^6\) from Chizat & Bach (2018), for instance).
Figure 2: Neural network with input dimension 1. **Left:** Normalized margin as we decrease $\lambda$. **Right:** Visualization of the normalized functions computed by the neural network and $\ell_1$ SVM solution for $\lambda \approx 10^{-14}$.

Figure 3: Comparing neural networks and kernel methods. **Left:** Classification. **Right:** Regression.

Theorem 3.1 and Lemmas 3.4, D.4, and D.5) of a trained neural network against a $\ell_2$ kernel SVM with relu features as we vary $n$. For classification we plot 0-1 error, whereas for regression we plot squared error. Our ground truth comes from a random neural network with 6 hidden units. For classification, we used rejection sampling to obtain datapoints with unnormalized margin of at least 0.1 on the ground truth network. We use a fixed dimension of $d = 20$. For all experiments, we train the network for 20000 steps with $\lambda = 10^{-8}$ and average over 100 trials for each plot point.

The plots in Figure 3 show that two-layer networks clearly outperform kernel methods in test error as $n$ grows. However, there seems to be looseness in the upper bounds for kernel methods: the kernel generalization bound appears to stay constant with $n$ (as predicted by our theory for regression), but the kernel test error decreases. There is also some variance in the neural network generalization bound for classification. This occurred likely because we did not tune learning rate and training time, so the optimization failed to find the best margin.

In Section F, we include additional experiments training modified WideResNet architectures on CIFAR10 and CIFAR100. Although ResNet is not homogeneous, we still report interesting increases in generalization performance from annealing the weight decay during training, versus staying at a fixed decay rate.

6 CONCLUSION

We have made the case that maximizing margin is one of the inductive biases of relu networks with cross-entropy loss. We show that we can obtain a maximum normalized margin by training with a weak regularizer. We also prove that larger $\ell_2$-normalized margin indicates better generalization for two-layer nets. Our work leaves open the question of how the $\ell_2$-normalized margin relates to generalization in much deeper neural networks. This is a fascinating theoretical and empirical question for future work. On the optimization side, we make progress towards understanding over-parametrized gradient descent by analyzing infinite-size neural networks. A natural direction for future work is to apply our theory to optimize the margin of finite-sized neural networks.
REFERENCES


We first show that $L_\lambda$ does indeed have a global minimizer.

**Claim A.1.** *In the setting of Theorems 2.1 and A.3, \( \arg\min_{\Theta} L_\lambda(\Theta) \) exists.*

**Proof.** We will argue in the setting of Theorem 2.1 where $L_\lambda$ is the multi-class cross entropy loss, because the logistic loss case is analogous. We first note that $L_\lambda$ is continuous in $\Theta$ because $f$ is continuous in $\Theta$ and the term inside the logarithm is always positive. Next, define $b \equiv \inf_{\Theta} L_\lambda(\Theta) > 0$. Then we note that for $\|\Theta\| > (b/\lambda)^{1/r} \leq M$, we must have $L_\lambda(\Theta) > b$. It follows that $\inf_{\|\Theta\| \leq M} L_\lambda(\Theta) = \inf_{\Theta} L_\lambda(\Theta)$. However, there must be a value $\Theta_\lambda$ which attains $\inf_{\|\Theta\| \leq M} L_\lambda(\Theta)$, because $\{\Theta : \|\Theta\| \leq M\}$ is a compact set and $L_\lambda$ is continuous. Thus, $\inf_{\Theta} L_\lambda(\Theta)$ is attained by some $\Theta_\lambda$. \( \square \)

### A.1 Missing Proofs for Multi-class Setting

Towards proving Theorem 2.1, we first show as we decrease $\lambda$, the norm of the solution $\|\Theta_\lambda\|$ grows.

**Lemma A.2.** *In the setting of Theorem 2.1 as $\lambda \to 0$, we have $\|\Theta_\lambda\| \to \infty$.*

To prove Theorem 2.1, we rely on the exponential scaling of the cross entropy: $L_\lambda$ can be lower bounded roughly by $\exp(-\|\Theta_\lambda\|/\gamma_\lambda)$, but also has an upper bound that scales with $\exp(-\|\Theta_\lambda\|/\gamma^*)$. By Lemma A.2, we can take large $\|\Theta_\lambda\|$ so the gap $\gamma^* - \gamma_\lambda$ vanishes. This proof technique is inspired by that of Rosset et al. (2004).
Proof of Theorem 2.1. For any $M > 0$ and $\Theta$ with $\gamma_\Theta \triangleq \min_i \{ f(\bar{\Theta}; x_i) - \max_{j \neq y_i} f(\bar{\Theta}; x_i) \}$,

\[
L_\lambda(M\Theta) = \frac{1}{n} \sum_{i=1}^{n} -\log \frac{\exp(M^af_{y_i}(\Theta; x_i))}{\sum_{j=1}^{n} \exp(M^af_j(\Theta; x_i))} + \lambda M^r\|\Theta\|^r \quad \text{(by the homogeneity of $f$)}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} -\log \frac{1}{1 + \sum_{j \neq y_i} \exp(M^a(f_j(\Theta; x_i) - f_{y_i}(\Theta; x_i)))} + \lambda M^r\|\Theta\|^r \quad \text{(A.1)}
\]

We can also apply $\sum_{j \neq y_i} \exp(M^a(f_j(\Theta; x_i) - f_{y_i}(\Theta; x_i))) \geq \max \exp(M^a(f_j(\Theta; x_i) - f_{y_i}(\Theta; x_i))) = \exp \gamma_\Theta$ in order to lower bound equation (A.1) and obtain

\[
L_\lambda(M\Theta) \geq \frac{1}{n} \log(1 + \exp(-M^a\gamma_\Theta)) + \lambda M^r\|\Theta\|^r \quad \text{(A.3)}
\]

Applying equation (A.2) with $M = \|\Theta_\lambda\|$ and $\Theta = \Theta^*$, noting that $\|\Theta^*\| \leq 1$, we have:

\[
L_\lambda(\Theta^*\|\Theta_\lambda\|) \leq \log(1 + (l - 1) \exp(-\|\Theta_\lambda\|^{a\gamma^*})) + \lambda\|\Theta_\lambda\|^r \quad \text{(A.4)}
\]

Next we lower bound $L_\lambda(\Theta_\lambda)$ by applying equation (A.3)

\[
L_\lambda(\Theta_\lambda) \geq \frac{1}{n} \log(1 + \exp(-\|\Theta_\lambda\|^{a\gamma_\lambda})) + \lambda\|\Theta_\lambda\|^r \quad \text{(A.5)}
\]

Combining equation (A.4) and equation (A.5) with the fact that $L_\lambda(\Theta_\lambda) \leq L_\lambda(\Theta^*\|\Theta_\lambda\|)$ (by the global optimality of $\Theta_\lambda$), we have

\[
\forall \lambda > 0, \ n \log(1 + (l - 1) \exp(-\|\Theta_\lambda\|^{a\gamma^*})) \geq \log(1 + \exp(-\|\Theta_\lambda\|^{a\gamma_\lambda}))
\]

Recall that by Lemma (A.2) as $\lambda \to 0$, we have $\|\Theta_\lambda\| \to \infty$. Therefore, $\exp(-\|\Theta_\lambda\|^{a\gamma^*}), \exp(-\|\Theta_\lambda\|^{a\gamma_\lambda}) \to 0$. Thus, we can apply Taylor expansion to the equation above with respect to $\exp(-\|\Theta_\lambda\|^{a\gamma^*})$ and $\exp(-\|\Theta_\lambda\|^{a\gamma_\lambda})$. If $\max\{\exp(-\|\Theta_\lambda\|^{a\gamma^*}), \exp(-\|\Theta_\lambda\|^{a\gamma_\lambda})\} < 1$, then we obtain

\[
n(l - 1) \exp(-\|\Theta_\lambda\|^{a\gamma^*}) \geq \exp(-\|\Theta_\lambda\|^{a\gamma_\lambda}) - O(\max\{\exp(-\|\Theta_\lambda\|^{a\gamma^*})^2, \exp(-\|\Theta_\lambda\|^{a\gamma_\lambda})^2\})
\]

We claim this implies that $\gamma^* \leq \lim \inf_{\lambda \to 0} \gamma_\lambda$. If not, we have $\lim \inf_{\lambda \to 0} \gamma_\lambda > \gamma^*$, which implies that the equation above is violated with sufficiently large $\|\Theta_\lambda\| (\|\Theta_\lambda\| \gg \log(2(l - 1)n)^{1/a}$ would suffice). By Lemma (A.2) $\|\Theta_\lambda\| \to \infty$ as $\lambda \to 0$ and therefore we get a contradiction.

Finally, we have $\gamma_\lambda \leq \gamma^*$ by definition of $\gamma^*$. Hence, $\lim \inf_{\lambda \to 0} \gamma_\lambda$ exists and equals $\gamma^*$. \qed

Now we fill in the proof of Lemma (A.2)

Proof of Lemma (A.2) For the sake of contradiction, we assume that $\exists C > 0$ such that for any $\lambda_0 > 0$, there exists $0 < \lambda < \lambda_0$ with $\|\Theta_\lambda\| \leq C$. We will determine the choice of $\lambda_0$ later and pick $\lambda$ such that $\|\Theta_\lambda\| \leq C$. Then the logits (the prediction $f_j(\Theta; x_i)$ before softmax) are bounded in absolute value by some constant (that depends on $C$), and therefore the loss function $-\log \frac{\exp(f_{y_i}(\Theta; x_i))}{\sum_j \exp(f_j(\Theta; x_i))}$ for every example is bounded from below by some constant $D > 0$ (depending on $C$ but not $\lambda$).

Let $M = \lambda^{-1/(r+1)}$, we have that

\[
0 < D \leq L_\lambda(M\Theta^*) \leq L_\lambda(M\Theta) \quad \text{(by the optimality of $\Theta_\lambda$)}
\]

\[
\leq -\log \frac{1}{1 + (l - 1) \exp(-M^a\gamma^*)} + \lambda M^r \quad \text{(by equation (A.2))}
\]

\[
= \log(1 + (l - 1) \exp(-\lambda^{-a/(r+1)}\gamma^*)) + \lambda^1/(r+1)
\]

\[
\leq \log(1 + (l - 1) \exp(-\lambda_0^{-a/(r+1)}\gamma^*)) + \lambda_0^1/(r+1)
\]

Taking a sufficiently small $\lambda_0$, we obtain a contradiction and complete the proof. \qed
A.2 Full Binary Classification Setting

For completeness, we state and prove our max-margin results for the setting where we fit binary labels \( y_i \in \{-1, +1\} \) (as opposed to indices in \([l]\)) and redefining \( f(\Theta; \cdot) \) to assign a single real-valued score (as opposed to a score for each label). This lets us work with the simpler \( \lambda \)-regularized logistic loss:

\[
L_\lambda(\Theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_if(\Theta; x_i))) + \lambda\|\Theta\|^r
\]

As before, let \( \Theta_\lambda \in \arg \min L_\lambda(\Theta) \), and define the normalized margin \( \gamma_\lambda \) by \( \gamma_\lambda \triangleq \min_i y_if(\Theta_\lambda; x_i) \). Define the maximum possible normalized margin

\[
\gamma^* \triangleq \max_{\|\Theta\| \leq 1} \min_i y_if(\Theta; x_i)
\]

**Theorem A.3.** Assume \( \gamma^* > 0 \) in the binary classification setting with logistic loss. Then as \( \lambda \to 0 \),

\( \gamma_\lambda \to \gamma^* \).

The proof follows via simple reduction to the multi-class case.

**Proof of Theorem A.3.** We prove this theorem via reduction to the multi-class case with \( l = 2 \). Construct \( \tilde{f} : \mathbb{R}^d \to \mathbb{R}^2 \) with \( \tilde{f}_1(\Theta; x_i) = -\frac{1}{2} f(\Theta; x_i) \) and \( \tilde{f}_2(\Theta; x_i) = \frac{1}{2} f(\Theta; x_i) \). Define new labels \( \tilde{y}_i = 1 \) if \( y_i = -1 \) and \( \tilde{y}_i = 2 \) if \( y_i = 1 \). Note that \( \tilde{f}_{\tilde{y}_i}(\Theta; x_i) - \tilde{f}_{\tilde{y}_j \neq \tilde{y}_i}(\Theta; x_i) = y_i f(\Theta; x_i) \), so the multi-class margin for \( \Theta \) under \( \tilde{f} \) is the same as binary margin for \( \Theta \) under \( f \). Furthermore, defining

\[
\tilde{L}_\lambda(\Theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \left( -\log \frac{\exp(\tilde{f}_{\tilde{y}_i}(\Theta; x_i))}{\sum_j \exp(\tilde{f}_j(\Theta; x_i))} \right) + \lambda\|\Theta\|^r
\]

we get that \( \tilde{L}_\lambda(\Theta) = L_\lambda(\Theta) \), and in particular, \( \tilde{L}_\lambda \) and \( L_\lambda \) have the same set of minimizers. Therefore we can apply Theorem 2.1 for the multi-class setting and conclude \( \gamma_\lambda \to \gamma^* \) in the binary classification setting.

A.3 Missing Proof for Optimization Accuracy

**Proof of Theorem 2.2.** Choose \( B \triangleq \left( \frac{1}{\gamma^*} \log \left( \frac{(l-1)(\gamma^*)^r/a}{\lambda} \right) \right)^{1/a} \). We can upper bound \( L_\lambda(\Theta') \) by computing

\[
L_\lambda(\Theta') \leq 2L_\lambda(\Theta) \leq 2L_\lambda(B\Theta^*) \\
\leq 2 \log(1 + (l-1)\exp(-B^a\gamma^*)) + 2\lambda B^r \\
\leq 2(l-1)\exp(-B^a\gamma^*) + 2\lambda B^r \\
\leq 2\frac{\lambda}{(\gamma^*)^r/a} + 2\lambda \left( \frac{1}{\gamma^*} \log \frac{(l-1)(\gamma^*)^{r/a}}{\lambda} \right)^{r/a} \\
\leq 4\lambda \left( \frac{1}{\gamma^*} \log \frac{(l-1)(\gamma^*)^{r/a}}{\lambda} \right)^{r/a} \triangleq L(U^B)
\]

Furthermore, it holds that \( \|\Theta'^*\|^r \leq \frac{L(U^B)}{\lambda} \). Now we note that

\[
L_\lambda(\Theta') \leq L(U^B) \leq 4 \left( \frac{(c+1)\log(n(l-1))} {n^c(l-1)^c} \right)^{r/a} \leq \frac{1}{2n}
\]

for sufficiently large \( c \). Now using the fact that \( \log(x) \geq \frac{x}{1+x} \forall x \geq -1 \), we additionally have the lower bound \( L_\lambda(\Theta') \geq \frac{1}{n} \log(1 + \exp(-\gamma'\|\Theta'^*\|^a)) \geq \frac{1}{n} \frac{1}{1+\exp(-\gamma'\|\Theta'^*\|^a)} \). Since \( L(U^B) \leq 1 \), we can rearrange to get

\[
\gamma' \geq -\log \frac{nL_\lambda(\Theta')}{\|\Theta'^*\|^a} \geq -\log \frac{nL(U^B)}{1-nL(U^B)} \geq -\log \left( \frac{2nL(U^B)}{\|\Theta'^*\|^a} \right)
\]
The middle inequality followed because $\frac{x}{1-x}$ is increasing in $x$ for $0 \leq x < 1$, and the last because $L^{(UB)} \leq \frac{1}{2n}$. Since $-\log 2nL^{(UB)} > 0$ we can also apply the bound $\|\Theta\|^r \leq \frac{L^{(UB)}}{x}$ to get

$$\gamma' \geq \frac{-\lambda^{a/r} \log 2nL^{(UB)}}{(L^{(UB)})^{a/r}} = -\log \left( 8n \lambda \left( \frac{1}{\gamma} \log \frac{(l-1)(\gamma^{r/a})^{r/a}}{x} \right) \right)$$

(by definition of $L^{(UB)}$)

$$\geq \frac{\gamma^*}{4^{a/r}} \left( \frac{\log (\gamma^{r/a})}{\log (l-1)(\gamma^{r/a})} - \frac{\log (\gamma^{r/a})}{\log (l-1)(\gamma^{r/a})} \right)$$

By assumption, $\log (\gamma^{r/a}) \geq c \log (l-1)$, and so $\log (\gamma^{r/a}) \geq \frac{e-2}{c+1} \log (l-1)(\gamma^{r/a})$. Thus, if $c \geq 4$,

$$\gamma' \geq \frac{\gamma^*}{4^{a/r}} \left( \frac{3}{5} \frac{\log (l-1)(\gamma^{r/a})}{\log (l-1)(\gamma^{r/a})} \right) \geq \frac{\gamma^*}{2 \cdot 4^{a/r}}$$

for $c$ sufficiently large.

\[ \square \]

B. **MISSING PROOFS IN SECTIONS 3.2 AND 3.3**

We first show Theorem 3.3 and then complete the proof of Theorem 3.2. The proof of Theorem 3.3 will consist of two steps: first, show that equation 3.6 has an optimal solution with sparsity $n$, and second, show that sparse solutions to equation 3.6 can be mapped to a neural network with the same margin, and vice versa. The following lemma and proof are based on Lemma 14 of Tibshirani et al. (2013).

**Lemma B.1.** Let $\text{supp}(\alpha) \triangleq \{u : |\alpha(u)| > 0\}$. There exists an optimal solution $\alpha^*$ to equation 3.6 with $|\text{supp}(\alpha^*)| \leq n$.

For the proof of this lemma, we find it convenient to work with a minimum norm formulation which we show is equivalent to equation 3.6

$$\min_{\alpha} \|\alpha\|_1$$

subject to $y_i \langle \alpha, \varphi(x_i) \rangle \geq 1 \forall i$ \hspace{1cm} (B.1)

**Claim B.2.** Let $S \subset \mathcal{L}^1(S^{d-1})$ be the set of optimizers for equation 3.6 and let $S' \subset \mathcal{L}^1(S^{d-1})$ be the set of optimizers for equation B.1. If equation B.1 is feasible, for any $\alpha \in S$, $\frac{\alpha}{\|\alpha\|_1} \in S'$, and for any $\alpha' \in S'$, $\frac{\alpha'}{\|\alpha'\|_1} \in S$.

**Proof.** Let $\text{opt}'$ denote the optimal objective for equation B.1. We note that $\frac{\alpha'}{\|\alpha'\|_1}$ is feasible for equation 3.6 with objective $\frac{1}{\gamma_{\ell_1}}$, and therefore $\gamma_{\ell_1} \geq \frac{1}{\text{opt}' \gamma_{\ell_1}}$. Furthermore, $\frac{1}{\gamma_{\ell_1}} y_i \int_{u \in S^{d-1}} \alpha(\bar{u}) \phi(u ; x_i) d\bar{u} \geq 1 \forall i$, and so $\frac{\alpha}{\gamma_{\ell_1}}$ is feasible for equation B.1 with objective $\frac{1}{\gamma_{\ell_1}}$. Therefore, $\text{opt}' \leq \frac{1}{\gamma_{\ell_1}}$. As a result, it must hold that $\text{opt}' = \frac{1}{\gamma_{\ell_1}}$, which means that $\frac{\alpha'}{\|\alpha'\|_1}$ is optimal for equation 3.6 and $\frac{\alpha}{\|\alpha\|_1}$ is optimal for equation B.1 as desired. \hspace{1cm} $\square$

First, note that if equation B.1 is not feasible, then $\gamma_{\ell_1} = 0$ and equation 3.6 has a trivial sparse solution, the all zeros function. Thus, it suffices to show that an optimal solution to equation B.1 exists that is $n$-sparse, since by Lemma B.2, equation B.1 and equation 3.6 have equivalent solutions up to a scaling. We begin by taking the dual of equation B.1.
Claim B.3. The dual of equation [B.7] has form
\[
\max_{\lambda \in \mathbb{R}^n} \lambda^T y \\
\text{subject to } \sum_{i=1}^n \lambda_i y_i \phi(u^T x_i) \leq 1 \quad \forall u \in S^{d-1} \\
\lambda_i \geq 0
\]

For any primal optimal solution \( \alpha^* \) and dual optimal solution \( \lambda^* \), it must hold that
\[
\sum_{i=1}^n \lambda_i^* y_i \phi(u^T x_i) = \text{sign}(\alpha^*(\bar{u})) \iff \alpha^*(\bar{u}) \neq 0 
\] (B.2)

**Proof.** The dual form can be solved for by computation. By strong duality, equation [B.2] must follow from the KKT conditions.

Now define the mapping \( v : S^{d-1} \rightarrow \mathbb{R}^n \) with \( v_i(\bar{u}) \equiv y_i \phi(u^T x_i) \). We will show a general result about linearly dependent \( v(\bar{u}) \) for \( \bar{u} \in \text{supp}(\alpha^*) \), after which we can reduce directly to the proof of [Tibshirani et al., 2013].

Claim B.4. Let \( \alpha^* \) be any optimal solution. Suppose that there exists \( S \subseteq \text{supp}(\alpha^*) \) such that \( \{ v(\bar{u}) : \bar{u} \in S \} \) forms a linearly dependent set, i.e.
\[
\sum_{\bar{u} \in S} c_{\bar{u}} v(\bar{u}) = 0 
\] (B.3)

for coefficients \( c \). Then \( \sum_{\bar{u} \in S} c_{\bar{u}} \text{sign}(\alpha^*(\bar{u})) = 0 \).

**Proof.** Let \( \lambda^* \) be any dual optimal solution, then \( \lambda^T v(\bar{u}) = \text{sign}(\alpha^*(\bar{u})) \forall \bar{u} \in \text{supp}(\alpha^*) \) by Claim [B.3]. Thus, we apply \( \lambda^T \) to both sides of equation [B.3] to get the desired statement.

**Proof of Lemma B.4.** The rest of the proof follows Lemma 14 in [Tibshirani et al., 2013]. The lemma argues that if the conclusion of Claim [B.4] holds and an optimal solution \( \alpha^* \) with \( \{ v(\bar{u}) : \bar{u} \in S \} \) linearly dependent, we can construct a new \( \alpha' \) with \( \|\alpha'\|_1 = \|\alpha^*\|_1 \) and \( \text{supp}(\alpha') \subseteq \text{supp}(\alpha^*) \) (where the inclusion is strict). Thus, if we consider an optimal \( \alpha^* \) with minimal support, it must follow that \( \{ v(\bar{u}) : \bar{u} \in \text{supp}(\alpha^*) \} \) is a linearly independent set, and therefore \( \|\text{supp}(\alpha^*)\| \leq n \).

We can now complete the proof of Theorem 3.3.

**Proof of Theorem 3.3.** We first apply Lemma [B.1] to conclude that equation [3.6] admits a \( n \)-sparse optimal solution \( \alpha^* \). Because of sparsity, we can now abuse notation and treat \( \alpha^* \) as a real-valued function such that \( \sum_{\bar{u} \in \text{supp}(\alpha^*)} |\alpha^*(\bar{u})| \leq 1 \). We construct \( \Theta \) with normalized margin at least \( \gamma_t / 2 \). For every \( \bar{u} \in \text{supp}(\alpha) \), we let \( \Theta \) have a corresponding hidden unit \( j \) with \( (w_j, u_j) = \left( \text{sign}(\alpha^*(\bar{u})), \sqrt{\frac{|\alpha^*(\bar{u})|}{2}}, \sqrt{\frac{|\alpha^*(\bar{u})|}{2}} \bar{u} \right) \), and set the remaining hidden units to 0. This is possible because \( m \geq n \). Now
\[
f(\Theta; x) = \sum_{j=1}^m w_j \phi(u_j^T x) = \frac{1}{2} \sum_{\bar{u} \in \text{supp}(\alpha^*)} \alpha^*(\bar{u}) \phi(u^T x)
\]
Furthermore,
\[
\|\Theta\|_2^2 = \sum_{j=1}^m w_j^2 + \|u_j\|_2^2 = \sum_{\bar{u} \in \text{supp}(\alpha)} \frac{|\alpha^*(\bar{u})|}{2} + \frac{|\alpha^*(\bar{u})|}{2} \|\bar{u}\|_2^2 = \sum_{\bar{u} \in \text{supp}(\alpha)} |\alpha^*(\bar{u})| \leq 1
\]
Thus it follows that \( \Theta \) has normalized margin at least \( \gamma_t / 2 \), so \( \gamma_{t + m} \geq \gamma_t / 2 \).
To conclude, we show that $\gamma^*,m \leq \gamma_0/2$. Let $\Theta^*,m$ denote the parameters obtaining optimal $m$-unit margin $\gamma^*,m$ with hidden units $(w_j^*,m, u_j^*,m)$ for $j \in [m]$. We can construct $\alpha$ to put a scaled delta mass of $2u_j^*,m \|u_j^*,m\|_2$ on $u_j^*,m$ for $j \in [m]$. It follows that

$$\|\alpha\|_1 = \sum_{j=1}^m 2|w_j^*,m|\|u_j^*,m\|_2 \leq \sum_{j=1}^m w_j^*,m^2 + \|u_j^*,m\|^2_2 = \|\Theta^*,m\|^2_2 \leq 1$$

Furthermore,

$$\int S_{\delta^{-1}} \alpha(u)\phi(\hat{u}^\top x) = 2 \sum_{j=1}^m w_j^*,m \|u_j^*,m\|_2 \phi((u_j^*,m)^\top x) = 2 \sum_{j=1}^m w_j^*,m \phi(u_j^*,m^\top x) = 2 f(\Theta^*,m; x)$$

Thus, $\alpha$ is a feasible solution to equation $3.6$ with objective value at least $2\gamma^*,m$. Therefore, $\gamma_0/2 \leq 2\gamma^*,m$, so $\gamma^*,m = \gamma_0/2$.

Theorem 3.1 follows almost immediately.

**Proof of Theorem 3.1.** Let $\Theta^*,m$ denote the parameters obtaining optimal $m$-unit margin $\gamma^*,m$. It is simple to see that $\gamma^*,m+1 \geq \gamma^*,m$: we can set $m$ units of our $m + 1$-hidden unit network to $\Theta^*,m$, and set the remaining unit to $\hat{u}$. Then this network will compute $f(\Theta^*,m; \cdot)$ using $m + 1$ units and have the same parameter norm. Finally, by Theorem 3.3 we have $\gamma_0/2 = \gamma^*,n = \gamma^*,n+1 = \cdots$.

**C Rademacher Complexity and Generalization Error**

We prove the generalization error bounds stated in Theorem 3.1 and Lemma 3.4 via Rademacher complexity. In this section, we will introduce the margin-based generalization error bounds for Rademacher complexity and prove Rademacher complexity bounds for the $\ell_1$ and $\ell_2$ norms.

Assume that our data $X, Y$ are drawn i.i.d. from ground truth distribution $p_{\text{data}}$ supported on $\mathcal{X} \times \mathcal{Y}$. For some hypothesis class $\mathcal{F}$ of real-valued functions, we define the empirical Rademacher complexity $\hat{\mathcal{R}}(\mathcal{F})$ as follows:

$$\hat{\mathcal{R}}(\mathcal{F}) \triangleq \frac{1}{n} \mathbb{E}_{\epsilon} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i) \right]$$

where $\epsilon_i$ are independent Rademacher random variables. Specifically, we consider the hypothesis classes of linear functionals in lifted feature space with bounded norm: for activation $\phi$, define $f^\phi(\alpha; \cdot) : \mathbb{R}^d \to \mathbb{R}$ with $f^\phi(\alpha; x) \triangleq (\phi(x), \alpha)$, and let $\mathcal{F}^\phi_B \triangleq \{ f^\phi(\alpha; \cdot) : \alpha \in L^p, \|\alpha\|_p \leq B \}$. Recall that we defined $\phi(x) \in L^\infty(\mathbb{S}^{d-1})$ in equation 3.5.

For the binary classification setting, we will use the following ramp loss, which can be thought of as a Lipschitz version of the standard 0-1 loss:

$$L_\gamma(a) \triangleq \begin{cases} 1 & a \leq 0 \\ 1 - a/\gamma & 0 < a \leq \gamma \\ 0 & a > \gamma \end{cases}$$

Following Section 3.1, we will use $L(\alpha) \triangleq \Pr_{(x,y) \sim p_{\text{data}}} (y f^\phi(\alpha; x) \leq 0)$ to denote the population 0-1 loss of the classifier $f^\phi(\alpha; \cdot)$. The following theorem follows from classical results in the literature (Koltchinskii et al. 2002, Kakade et al. 2009).

**Theorem C.1.** Let $(x_i, y_i)_{i=1}^n$ be drawn i.i.d. from $p_{\text{data}}$. We work in the binary classification setting, so $\mathcal{Y} = \{-1, 1\}$. Let $q \triangleq \frac{1}{p}, C \triangleq \sup_{x \in X} \|\phi(x)\|_q$. Then with probability at least $1 - \delta$ over the random draws of the data, every $f^\phi(\alpha; \cdot) \in \mathcal{F}^\phi_B$ satisfies

$$L(\alpha) \leq \frac{1}{n} \sum_{i=1}^n L_\gamma(y_i f^\phi(\alpha; x_i)) + \frac{4\hat{\mathcal{R}}(\mathcal{F}^\phi_B)}{\gamma} + \frac{\log \log 2 + C}{\gamma n} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{C.1}$$
In particular, for any $\alpha$ that classifies all training data correctly with $p$-norm normalized margin $\gamma_\alpha \triangleq \min_i y_i f^\phi(\frac{\langle \alpha, \phi(x_i) \rangle}{\|\alpha\|_p}; x_i) > 0$, with probability at least $1 - \delta$,

$$L(\alpha) \leq \frac{4\hat{R}(F_1^{p,\phi})}{\gamma_\alpha} + \sqrt{\frac{\log \log_2 \frac{4C}{\gamma_\alpha}}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{C.2}$$

**Proof.** Our starting point is Theorem 2 of \cite{Kakade2009}, which states the same claim for $C \triangleq \sup_{f^\phi(\alpha; \cdot) \in F_1^{p,\phi}} \sup_{x \in X} f^\phi(\alpha; x)$. We note that

$$\sup_{f^\phi(\alpha; \cdot) \in F_1^{p,\phi}} \sup_{x \in X} f^\phi(\alpha; x) = \sup_{f^\phi(\alpha; \cdot) \in F_1^{p,\phi}} \sup_{x \in X} \langle \alpha, \phi(x) \rangle \leq \sup_{f^\phi(\alpha; \cdot) \in F_1^{p,\phi}} \sup_{x \in X} \| \alpha \|_p \| \phi(x) \|_q \quad \text{(by the duality of } p \text{ and } q \text{ norms)}$$

$$\leq C$$

For the second statement, we simply apply equation [C.1] to $\alpha \triangleq \alpha/\|\alpha\|_p$ and use the fact that $L(\alpha) = L(\tilde{\alpha})$. \hfill \Box

Next, we focus on analyzing the Rademacher complexity $\hat{R}(F_B^{p,\phi})$, specifically for the cases $p = 1, 2$. Since our general analysis only involves duality of norms, similar derivations have been done in the past \cite{Bartlett2002}. We include our derivations here for completeness.

**Lemma C.2.** $\hat{R}(F_B^{1,\phi}) \leq \frac{1}{n} B E_{e_i} \left[ \| \sum_{i=1}^n \epsilon_i \phi(x_i) \|_{\infty} \right]$.

**Lemma C.3.** $\hat{R}(F_B^{2,\phi}) \leq \frac{1}{n} B \sqrt{\sum_{i=1}^n \| \phi(x_i) \|_2^2}$.

**Proof of Lemmas C.2 and C.3.** For $p = 1, 2$, let $\| \cdot \|_{p,*}$ denote the dual norm of $\| \cdot \|_p$, so $\| \cdot \|_{1,*} = \| \cdot \|_{\infty}$ and $\| \cdot \|_{2,*} = \| \cdot \|_2$. We write

$$\hat{R}(F_B^{p,\phi}) = \frac{1}{n} E_{e_i} \left[ \sup_{\alpha \in F_B^{p,\phi}} \left\langle \alpha, \sum_{i=1}^n \epsilon_i \phi(x_i) \right\rangle \right]$$

$$\leq \frac{1}{n} E_{e_i} \left[ \sup_{\alpha \in F_B^{p,\phi}} \| \alpha \|_k \left\| \sum_{i=1}^n \epsilon_i \phi(x_i) \right\|_{k,*} \right]$$

$$\leq \frac{1}{n} B \cdot E_{e_i} \left[ \left\| \sum_{i=1}^n \epsilon_i \phi(x_i) \right\|_{k,*} \right]$$

Lemma C.2 immediately follows. To obtain Lemma C.3 we have

$$\hat{R}(F_B^{2,\phi}) \leq \frac{1}{n} B E_{e_i} \left[ \left\| \sum_{i=1}^n \epsilon_i \phi(x_i) \right\|_2 \right]$$

$$\leq \frac{1}{n} B \sqrt{E_{e_i} \left[ \left\| \sum_{i=1}^n \epsilon_i \phi(x_i) \right\|_2^2 \right]} \quad \text{(via Jensen’s inequality)}$$

$$\leq \frac{1}{n} B \sqrt{E_{e_i} \left[ \sum_{i=1}^n \sum_{j=1}^n \epsilon_i \epsilon_j \langle \phi(x_i), \phi(x_j) \rangle \right]}$$

$$\leq \frac{1}{n} B \sqrt{\sum_{i=1}^n \| \phi(x_i) \|_2^2} \quad \text{(terms where } i \neq j \text{ cancel out)}$$

\hfill \Box
We will now complete the bound on $\hat{\mathcal{R}}(F_B^{1,\phi})$ for general Lipschitz activations $\phi$.

**Lemma C.4.** Suppose that our activation $\phi$ is $M$-Lipschitz. Then

$$\hat{\mathcal{R}}(F_B^{1,\phi}) \leq B(\sqrt{\sum_{i=1}^{n} \|\varphi(x_i)\|_\infty^2} + 2M \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2})$$

**Proof.** We will show that

$$E_{\epsilon_i} \left[ \left\| \sum_{i=1}^{n} \epsilon_i \varphi(x_i) \right\|_\infty \right] \leq \sqrt{\sum_{i=1}^{n} \|\varphi(x_i)\|_\infty^2} + 2M \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2}$$

from which the statement of the lemma follows via Lemma C.2. Fix any $\bar{u} \in S^{d-1}$. Then we get the decomposition

$$E_{\epsilon_i} \left[ \sup_{\bar{u} \in S^{d-1}} \left| \sum_{i=1}^{n} \epsilon_i \phi(\bar{u}^T x_i) \right| \right] \leq E_{\epsilon_i} \left[ \left| \sum_{i=1}^{n} \epsilon_i \phi(\bar{u}^T x_i) \right|^2 \right]$$

$$= E_{\epsilon_i} \left[ \left( \sum_{i=1}^{n} \epsilon_i \phi(\bar{u}^T x_i) \right)^2 \right] \leq \sum_{i=1}^{n} \phi(\bar{u}^T x_i)^2$$

We can bound the second term of equation (C.3) by

$$E_{\epsilon_i} \left[ \sup_{\bar{u} \in S^{d-1}} \left| \sum_{i=1}^{n} \epsilon_i \phi(\bar{u}^T x_i) \right| - \inf_{\bar{u} \in S^{d-1}} \left| \sum_{i=1}^{n} \epsilon_i \phi(\bar{u}^T x_i) \right| \right]$$

$$\leq 2E_{\epsilon_i} \left[ \sup_{\bar{u} \in S^{d-1}} \left| \sum_{i=1}^{n} \epsilon_i \phi(\bar{u}^T x_i) \right| \right]$$

This simply gives an empirical Rademacher complexity of the hypothesis class $\mathcal{F} \triangleq \{ x \mapsto \phi(\bar{u}^T x) : \bar{u} \in S^{d-1} \} \text{ scaled by } n$. By the Lipschitz contraction property of Rademacher complexity, using the fact that $\phi$ is $M$-Lipschitz, we can therefore bound equation (C.5) by

$$2E_{\epsilon_i} \left[ \sup_{\bar{u} \in S^{d-1}} \left| \sum_{i=1}^{n} \epsilon_i \phi(\bar{u}^T x_i) \right| \right] \leq 2M \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2}$$

Plugging equation (C.4) and equation (C.6) back into equation (C.3) gives the desired bound. \qed

As an example, we can apply this bound to relu features:

**Corollary C.5.** Suppose that $\phi$ is the relu activation. Let $\kappa \triangleq \text{Vol}(S^{d-1})$. Then $\hat{\mathcal{R}}(F_B^{1,\phi}) \leq \frac{3B\|X\|_{FP}}{n}$, and $\hat{\mathcal{R}}(F_B^{2,\phi}) \lesssim \frac{B\sqrt{\kappa\|X\|_F}}{n\sqrt{d}}$. 

20
Proof. In the setting of Lemma \ref{lem:c4}, we have $\|\varphi(x_i)\|_\infty \leq \|x_i\|_2$ and $M = 1$. Applying this in Lemma \ref{lem:c4} gives the desired 1-norm Rademacher bound.

For the 2-norm case, we first show that $\|\varphi(x_i)\|_2 = \Theta \left( \frac{\kappa}{\sqrt{\eta}} \|x_i\|_2 \right)$. We can compute

$$\|\varphi(x_i)\|^2_2 = \text{Vol}(S^d) \mathbb{E}_{\bar{u} \sim S^{d-1}} [\text{relu}(\bar{u}^\top x_i)^2]$$

$$= \frac{k}{d} \mathbb{E}_{\bar{u} \sim S^{d-1}} [\text{relu}(\sqrt{d} \bar{u}^\top x_i)^2]$$

$$= \frac{k}{d} \frac{1}{M_2} \mathbb{E}_{u \sim \mathcal{N}(0,I_{d \times d})} [\text{relu}(u^T x_i)^2] \quad (M_2 \text{ is the second moment of } \mathcal{N}(0,1))$$

$$= \Theta \left( \frac{\kappa}{\sqrt{\eta}} \|x_i\|^2_2 \right) \quad (C.7)$$

where the last line uses the computation provided in Lemma A.1 by Du et al. (2017). Now we plug this into Lemma C.3 to get the desired bound. \hfill $\Box$

C.1 Proof of Network and Kernel Generalization Bounds

We use our Rademacher complexity-based generalization bounds to provide a proof of Theorem 3.1 and Lemma 3.4.

Proof of Theorem 3.1. Since $\phi$ is 1-homogeneous and 1-Lipschitz, it must follow that $|\phi(\bar{u}^\top x)| \leq |\bar{u}^\top x|$, and so $\|\varphi(x)\|_\infty \leq \|x\|_2$. Thus, $\sup_{x \in \mathcal{X}} \|\varphi(x)\|_\infty \leq \sup_{x \in \mathcal{X}} \|x\|_2$. Furthermore, $\hat{R}(\mathcal{F}_1, \phi) \leq \frac{4 \|x\|_F}{n \gamma}$. Finally, we note that the mapping given in the proof of Theorem 3.3 between neural network and linear functional on $\varphi(\cdot)$ applies here as well: we construct $\alpha$ to put a mass of $2\gamma \|\bar{u}^\top x\|_2$ on $\bar{u}^\top x$ for each $(\bar{u}_j, u_j) \in \Theta$. Then $\|\alpha\|_1 \leq \|\Theta\|_2^2$, and $\langle \alpha, \varphi(x) \rangle = 2f(\Theta; x_i)$. Thus, following the notation of Theorem C.1, we can apply equation C.2 of Theorem C.1 to get with probability $1 - \delta$,

$$L(\Theta) = L(\alpha)$$

$$\leq \frac{4 \hat{R}(\mathcal{F}_1, \phi)}{\gamma \alpha} + \sqrt{\frac{\log \log_2 \sup_{x \in \mathcal{X}} \|x\|_2 \|x\|_2}{n \gamma}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

$$\leq 6 \frac{\|x\|_F}{n \gamma} + \sqrt{\frac{\log \log_2 \frac{4C}{\gamma \alpha}}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \quad \text{(plugging in our bound for } \hat{R}(\mathcal{F}_1, \phi))$$

which gives us equation 3.2. To conclude equation 3.3, we apply the above on $\Theta_{\lambda_m}$ and use Theorem A.3. \hfill $\Box$

We will now prove Lemma 3.4.

Proof of Lemma 3.4. From equation C.7, we first obtain $\sup_{x \in \mathcal{X}} \|\varphi(x)\|_2 \lesssim C \sqrt{\gamma}$. Denote the optimizer for equation 3.8 by $\alpha_{\ell_2}$. Since $\alpha_{\ell_2}$ has normalized margin $\sqrt{\gamma_{\ell_2}}$, we apply Theorem C.1 to get with probability $1 - \delta$,

$$L_{\ell_2, \text{svm}} = L(\alpha_{\ell_2}) \leq \frac{4 \hat{R}(\mathcal{F}_2, \phi)}{\sqrt{\gamma_{\ell_2}}} + \sqrt{\frac{\log \log_2 \sup_{x \in \mathcal{X}} \|\varphi(x)\|_2}{n \gamma_{\ell_2}} \|x\|_2} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

$$\lesssim \frac{\|x\|_F}{n \gamma_{\ell_2} \sqrt{d}} + \sqrt{\frac{\log \max \left\{ \log_2 \frac{C \sqrt{d}}{\gamma_{\ell_2}}, 2 \right\}}{n \gamma_{\ell_2}}} + \sqrt{\frac{\log(1/\delta)}{n \gamma_{\ell_2}}} \quad \text{(applying Corollary C.5)}$$

\hfill $\Box$
D Missing Proofs for Comparison to Kernel Methods

D.1 Classification

In this section we will complete a proof of Theorem 3.5. Recall the construction of the distribution \( D \) provided in Section 3.4. We first provide a classifier of this data with small \( \ell_1 \) norm.

**Lemma D.1.** In the setting of Theorem 3.5, we have that

\[
\gamma_{\ell_1} \geq \frac{\sqrt{2}}{4}.
\]

**Proof.** Consider the network \( f(x) = \frac{1}{2} \left( (x^T (e_1 + e_2) + (x^T (-e_1 - e_2) + (x^T (-e_1 + e_2))/\sqrt{2})_+ - (x^T (-e_1 - e_2)/\sqrt{2})_+ \right) \). The attained margin \( \gamma = \sqrt{2} \), so \( \gamma_{\ell_1} \geq \frac{\sqrt{2}}{4} \). \( \square \)

Now we will upper bound the margin attainable by the \( \ell_2 \) SVM.

**Proof of Lemma 3.6.** By the definition of \( \gamma_{\ell_2} \), we have that for any \( \alpha \) with \( \sqrt{n} \| \alpha \|_2 \leq 1 \), we have

\[
\gamma_{\ell_2} \leq \max_{\sqrt{n} \| \alpha \|_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} \langle \alpha, y_i \varphi(x_i) \rangle
\]

Setting \( \alpha = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(x_i) y_i / \| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(x_i) y_i \|_2 \) completes the proof. (Attentive readers may realize that this is equivalent to setting the dual variable of the convex program 3.8 to all 1’s function.) \( \square \)

**Proof of Lemma 3.7.** Let \( W_i = \varphi(x_i) y_i \). We will bound several quantities regarding \( W_i \)'s. In the rest of the proof, we will condition on the event \( E \) that \( \forall i, \| x_i \|_2^2 \lesssim d \log n \). Note that \( E \) is a high probability event and conditioned on \( E \), \( x_i \)'s are still independent. We omit the condition on \( E \) in the rest of the proof for simplicity.

We first show that assuming the following three inequalities that the conclusion of the Lemma follows.

1. \( \forall i, \| W_i \|_2^2 \lesssim \kappa \log n \).
2. \( \sigma^2 = \text{Var}[\sum_{i} W_i] \triangleq \sum_{i=1}^{n} \mathbb{E}[\| W_i - \mathbb{E} W_i \|_2^2] \lesssim n\kappa \log n \)
3. \( \| \mathbb{E} [\sum W_i] \|_2 \lesssim \sqrt{\kappa n} / d \).

By bullets 1, 2, and Bernstein inequality, we have that with probability at least \( 1 - dn^{-10} \) over the randomness of the data \( (X,Y) \),

\[
\left\| \sum_{i=1}^{n} W_i - \mathbb{E} \left[ \sum_{i=1}^{n} W_i \right] \right\|_2 \leq \sqrt{\kappa \log 1.5 n} + \sqrt{n\kappa \log^2 n} \lesssim \sqrt{\kappa \log^2 n}
\]

By bullet 3 and equation above, we complete the proof with triangle inequality:

\[
\left\| \sum_{i=1}^{n} W_i \right\|_2 \leq \left\| \mathbb{E} \left[ \sum_{i=1}^{n} W_i \right] \right\|_2 + \sqrt{n\kappa \log^2 n} \lesssim \sqrt{n\kappa \log^2 n} + \sqrt{\kappa n} / d
\]

Therefore, it suffices to prove bullets 1, 2 and 3. Note that 2 is a direct corollary of 1 so we will only prove 1 and 3. We start with 3:

By the definition of the \( \ell_2 \) norm in \( L^2(\mathbb{S}^{d-1}) \) and the independence of \( (x_i, y_i) \)'s, we can rewrite

\[
\left\| \mathbb{E} \left[ \sum_{i=1}^{n} W_i \right] \right\|_2^2 = \kappa \cdot n^2 \mathbb{E}_{\tilde{u} \sim \mathbb{S}^{d-1}} \left[ \mathbb{E}_{(x,y) \sim D} \varphi(x)[\tilde{u}] \cdot y \right]^2 \tag{D.1}
\]
Let \( \tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_d) \) and \( \tilde{u}_{-2} = (\tilde{u}_3, \ldots, \tilde{u}_d) \in \mathbb{R}^{d-2} \), and define \( \tau \triangleq \|\tilde{u}_{-2}\|_2 \). Let \( x_{-2} = (x^T e_3, \ldots, x^T e_d) \). Note that \( \varphi(x)[u]y = y^T [u^T e_1 + \tilde{u}_2 \cdot x^T e_2 + \tilde{u}_{-2}^T x_{-2}] \) and \( \tilde{u}_{-2}^T x_{-2} \) has distribution \( \|\tilde{u}_{-2}\|_2 \cdot \mathcal{N}(0,1) = \tau \cdot \mathcal{N}(0,1) \). Let \( z = \tilde{u}_{-2}^T x_{-2}/\tau \), and therefore \( z \) has standard normal distribution. With this change of the variables, by the definition of the distribution \( D \), we have

\[
\mathbb{E}_{(x,y) \sim D} \varphi(x)[u] \cdot y = \frac{1}{4} \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[ (\tilde{u}_1 + \tilde{u}_2 + \tau z)_+ + \frac{1}{2} \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[ (\tilde{u}_1 - \tilde{u}_2 + \tau z)_+ \right] - \frac{1}{4} \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[ (\tilde{u}_1 - \tilde{u}_2 + \tau z)_+ \right] \right]
\]

By claim \([D.2]\) and the 1-homogeneity of \( \text{relu} \), we can simplify the above equation to

\[
\mathbb{E}_{(x,y) \sim D} \varphi(x)[u] \cdot y = \frac{1}{4} \tau \cdot (2c_1 + O(\min\{|\tilde{u}_1 + \tilde{u}_2|/\tau, |\tilde{u}_1 + \tilde{u}_2|^2/\tau^2\}))
\]

\[
\leq \min\{|\tilde{u}_1| + |\tilde{u}_2|, (|\tilde{u}_1| + |\tilde{u}_2|)^2/\tau\}
\]

It follows that

\[
\mathbb{E}_{\tilde{u} \sim \mathcal{S}^{d-1}} \left[ \mathbb{E}_{(x,y) \sim D} \varphi(x)[u] \cdot y \right]^2 \leq \mathbb{E}_{\tilde{u}} \left[ \min\{|\tilde{u}_1| + |\tilde{u}_2|^2, (|\tilde{u}_1| + |\tilde{u}_2|^2)/|\tilde{u}_{-2}\|_2 \} \right] \leq \mathbb{E}_{\tilde{u}} \left[ (|\tilde{u}_1| + |\tilde{u}_2|^2)^4/|\tilde{u}_{-2}\|_2 \} \right] \leq \exp(-d) \leq d^{2}
\]

Combining equation \([D.1]\) and equation \([D.2]\) we complete the proof of bullet 3. Next we prove bullet 1. Note that \( \varphi(x)[u]y \) is bounded by \( |\tilde{u}_1| + |\tilde{u}_2| + ||\tilde{u}_{-2}^T x_{-2}||_2 \). Therefore, conditioned on \( ||x_{-2}||_2 \leq d \log n \)

\[
||W_1||_2^2 \leq \mathbb{E}_{\tilde{u} \sim \mathcal{S}^{d-1}} \left[ (|\tilde{u}_1| + |\tilde{u}_2| + ||\tilde{u}_{-2}^T x_{-2}||_2)^2 \right] \leq \mathbb{E}_{\tilde{u}} \left[ ||\tilde{u}_1||^2 + \mathbb{E}_{\tilde{u}} \left[ ||\tilde{u}_2||^2 \right] + \mathbb{E}_{\tilde{u}} \left[ ||\tilde{u}_{-2}^T x_{-2}||_2 \right] \right] \leq 1/d \log n
\]

Hence we complete the proof. \( \square \)

**Claim D.2.** Let \( Z \sim \mathcal{N}(0,1) \) and \( a \in \mathbb{R} \). Then, there exists a universal constant \( c_1 \) and \( c_2 \) such that

\[
|\mathbb{E}[(a + Z)_+ + (-a + Z)_+] - 2c_1| \leq c_2 \min\{|a|, a^2\}.
\]

**Proof.** Without loss of generality we can assume \( a \geq 0 \). Then,

\[
\mathbb{E}[(a + Z)_+ + (-a + Z)_+] = \mathbb{E}[(a + Z)1[Z \geq -a]] + \mathbb{E}[(Z - a)1[Z \geq a]]
\]

\[
= \mathbb{E}[a \cdot 1[Z \geq -a]] + \mathbb{E}[Z \cdot 1[Z \geq -a]] - \mathbb{E}[a \cdot 1[Z \geq a]] + \mathbb{E}[Z \cdot 1[Z \geq a]]
\]

\[
= \mathbb{E}[a1[\cdot a \leq Z \leq a]] + 2 \mathbb{E}[Z \cdot 1[Z \geq a]]
\]

\[
= \mathbb{E}[a1[\cdot a \leq Z \leq a]] + 2 \mathbb{E}[Z \cdot 1[Z \geq 0]] - 2 \mathbb{E}[Z \cdot 1[a \geq Z \geq 0]]
\]

\[
= 2c_1 + O(\min\{|a|, a^2\})
\]

where the last equality uses the fact that \( c_1 \triangleq \mathbb{E}[Z \cdot 1[Z \geq 0]] \) and \( \mathbb{E}[a1[\cdot a \leq Z \leq a]] \leq a \mathbb{E}[1[a \leq Z \leq a]] \leq a \min\{1, a\}. \) \( \square \)

Now we will prove Theorem 3.5.

**Proof of Theorem 3.5.** To circumvent the technical issue of bounded support in Theorem 3.1 and Lemma 3.4, we construct \( p_{data} \) to be a slightly modified version of \( D \): perform rejection sampling of \( (x, y) \sim D \) until we obtain a sample with \( ||x||_2^2 \leq d \log n \). Since this occurs with very high
probability, the high probability result of Lemma\textsuperscript{3.7} still translates to $p_{\text{data}}$. Now apply Lemma\textsuperscript{3.6} to conclude that $\gamma_{\ell_1} \lesssim \frac{\log n}{\sqrt{d}} + \frac{1}{d}$. Furthermore, Lemma\textsuperscript{D.1} allows us to conclude that $\gamma_{\ell_1} \gtrsim 1$.

With high probability over the draws of the samples, we will also get that $\|X\|_F / \sqrt{n} \lesssim \sqrt{d}$. We can therefore apply Theorem\textsuperscript{3.1} and conclude that with probability $1 - \delta$,

$$L_{\ell_1, \text{svm}} \lesssim \sqrt{\frac{d}{n}} + \sqrt{\frac{\log \log (d n)}{n}} + \sqrt{\frac{\log (1/\delta)}{n}}$$

Furthermore, plugging $\gamma_{\ell_1}$ into the bound of Lemma\textsuperscript{3.4} gives us

$$\min \left\{ \frac{1}{\log n} \frac{d}{\sqrt{n}} \right\} + \sqrt{\frac{\log \log (d n)}{n}} + \frac{\log (1/\delta)}{n}$$

\[ \Box \]

D.2 Regression

We will first define the 1-norm and 2-norm regression problems. The regression equivalent of equation\textsuperscript{3.6} for $\alpha \in L^1(S^{d-1})$ is as follows:

$$\alpha_{\ell_1} \in \arg \min_{\alpha} \| \alpha \|_1$$

subject to $\langle \alpha, \varphi(x_i) \rangle = y_i$ \hfill (D.3)

Next we define the regression version of equation\textsuperscript{3.8}

$$\alpha_{\ell_2} \in \arg \min_{\alpha} \| \alpha \|_2$$

subject to $\langle \alpha, \varphi(x_i) \rangle = y_i$ \hfill (D.4)

where $\alpha \in L^2(S^{d-1})$.

We will briefly motivate our study of the regression setting by connecting the minimum 1-norm solution to neural networks. To compare, in the classification setting, optimizing the weakly regularized loss over neural networks is equivalent to solving the $\ell_1$ SVM. In the regression setting, solving the weakly regularized squared error loss is equivalent to finding the minimum 1-norm solution that fits the datapoints exactly.

**Theorem D.3.** Let $f(\Theta; \cdot)$ be some two-layer neural network with $n \geq n$ hidden units parametrized by $\Theta$, as in Section\textsuperscript{3.1}. Define the $\lambda$-regularized squared error loss

$$L_{\lambda,m}(\Theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} (f(\Theta; x_i) - y_i)^2 + \lambda \| \Theta \|_2^2$$

with $\Theta_{\lambda,m} \in \arg \min_{\Theta} L_{\lambda,m}(\Theta)$. Suppose that equation\textsuperscript{D.3} is feasible with optimal solution $\alpha_{\ell_1}$. Then as $\lambda \to 0$, $L_{\lambda,m}(\Theta_{\lambda,m}) \to 0$ and $\| \Theta_{\lambda,m} \|_2 \to 2\| \alpha_{\ell_1} \|_1$.

**Proof.** We can see that equation\textsuperscript{D.3} will have a $n$-sparse solution $\alpha^*$ using the same reasoning as the proof of Lemma\textsuperscript{B.1}. Furthermore, following the proof of Theorem\textsuperscript{3.3}, the function $x \mapsto \langle \alpha^*, \varphi(x) \rangle$ is implementable by a neural network $\Theta^{*,m}$ with $\| \Theta^{*,m} \|_2^2 = 2\| \alpha^* \|_1 = 2\| \alpha_{\ell_1} \|_1$. Following the same reasoning as before, we can also conclude that $\Theta^{*,m}$ is an optimal solution for:

$$\min_{\Theta} \| \Theta \|_2^2$$

subject to $f(\Theta; x_i) = y_i$ \hfill (D.5)

Now we note that $\lambda \| \Theta_{\lambda,m} \|_2^2 \leq L_{\lambda,m}(\Theta_{\lambda,m}) \leq L_{\lambda,m}(\Theta^{*,m}) = \lambda \| \Theta^{*,m} \|_2^2$, so as $\lambda \to 0$, and also $\| \Theta_{\lambda,m} \|_2 \leq \| \Theta^{*,m} \|_2$. Now assume for the sake of contradiction that $\exists B$ with $\| \Theta_{\lambda,m} \|_2 \leq B < \| \Theta^{*,m} \|_2$ for arbitrarily small $\lambda$. We define

$$r^* \triangleq \min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} (f(\Theta; x_i) - y_i)^2$$

subject to $\| \Theta \|_2 \leq B$ \hfill (D.5)

Note that $r^* > 0$ since $\Theta^{*,m}$ is optimal for equation\textsuperscript{D.5}. However, $L_{\lambda,m} \geq r^*$ for arbitrarily small $\lambda$, a contradiction. Thus, $\lim_{\lambda \to 0} \| \Theta^{*,m} \|_2^2 = \| \Theta^{*,m} \|_2^2$. \hfill $\Box$
We can also provide the same generalization error bound for the $2$-norm and relu features: 

**Lemma D.4.** Let $l(\cdot; y) : \mathbb{R} \rightarrow [-c, c]$ be a bounded $M$-Lipschitz loss function. Assume that $\phi$ is a $1$-homogeneous and $1$-Lipschitz activation. Let $(x_i, y_i)_{i=1}^n$ be drawn i.i.d from $p_{\text{data}}$. Then with probability at least $1 - \delta$ over the dataset, every $\alpha \in \mathcal{L}^1(\mathcal{F}^{d-1})$ satisfies 

\[
E_{(x,y) \sim p_{\text{data}}}[l(f^\phi(\alpha; x); y)] \leq \frac{1}{n} \sum_{i=1}^n l(f^\phi(\alpha; x_i); y_i) + 12M \max\left\{1, \|\alpha\|_1 \|X\|_F\right\} + c \sqrt{\log(1/\delta) + \log(\max\{1, 2\|\alpha\|_1 \|X\|_F\})/2n}
\]

**Proof.** Our starting point is Theorem 1 of [Kakade et al. (2009)](citations needed), which states that with probability $1 - \delta$, for any fixed hypothesis class $\mathcal{F}$ and $f \in \mathcal{F}$.

\[
E_{(x,y) \sim p_{\text{data}}}[l(f(x); y)] \leq \frac{1}{n} \sum_{i=1}^n l(f(x_i); y_i) + 2M \hat{\mathcal{R}}(\mathcal{F}) + c \sqrt{\log(1/\delta)/n} \tag{D.6}
\]

We define $B_j \triangleq 2^{2j} \mathcal{F}^{1,\phi}_{B_j}$ for $j \geq 0$. We note that by Lemma C.4, $\hat{\mathcal{R}}(\mathcal{F}^{1,\phi}_{B_j}) \leq \frac{32j}{2^{2j+1}}$, and apply the above on $\mathcal{F}^{1,\phi}_{B_j}$ using $\delta_j \triangleq \frac{\delta}{2^{2j+1}}$. Then using a union bound, with probability $1 - \sum_{j=0}^\infty \delta_j = 1 - \delta$, for all $j \geq 0$ and $f^\phi(\alpha; \cdot) \in \mathcal{F}^{1,\phi}_{B_j}$.

\[
E_{(x,y) \sim p_{\text{data}}}[l(f^\phi(\alpha; x); y)] \leq \frac{1}{n} \sum_{i=1}^n l(f^\phi(\alpha; x_i); y_i) + 2M \hat{\mathcal{R}}(\mathcal{F}^{1,\phi}_{B_j}) + c \sqrt{\log(1/\delta_j)/n}
\]

We define $B_j \triangleq 2^{2j} \mathcal{F}^{2,\phi}_{B_j}$ for $j \geq 0$. We note that by Lemma C.5, $\mathcal{R}(\mathcal{F}^{2,\phi}_{B_j}) \leq \frac{\sqrt{2j}}{n\sqrt{d}}$, from Lemma C.5. Thus, again union bounding over all $j$, equation \text{D.6} gives with probability $1 - \delta$, for all $j \geq 0$ and $f^\phi(\alpha; \cdot) \in \mathcal{F}^{2,\phi}_{B_j}$.

\[
E_{(x,y) \sim p_{\text{data}}}[l(f^\phi(\alpha; x); y)] \leq \frac{1}{n} \sum_{i=1}^n l(f^\phi(\alpha; x_i); y_i) + M \sqrt{n\sqrt{d}} \max\left\{1, \|\alpha\|_2 \|X\|_F\right\} + c \sqrt{\log(1/\delta) + \log(\max\{1, \|\alpha\|_2 \|X\|_F\})/2n}
\]

Now we assign the $\alpha$ to different $j$ as before to obtain the statement in the lemma. \qed

We can also provide the same generalization error bound for the $2$-norm and relu features:

**Lemma D.5.** In the setting of Lemma D.4, choose $\phi$ to be the relu activation. Then with probability $1 - \delta$,

\[
E_{(x,y) \sim p_{\text{data}}}[l(f^\phi(\alpha; x); y)] \leq \frac{1}{n} \sum_{i=1}^n l(f^\phi(\alpha; x_i); y_i) + M \sqrt{n\sqrt{d}} \max\left\{1, \|\alpha\|_2 \|X\|_F\right\} + c \sqrt{\log(1/\delta) + \log(\max\{1, \|\alpha\|_2 \|X\|_F\})/2n}
\]

**Proof.** We proceed the same way as in the proof of Lemma D.4. We define $B_j$ as before, and this time have $\mathcal{R}(\mathcal{F}^{2,\phi}_{B_j}) \leq \frac{\sqrt{2j}}{n\sqrt{d}}$ from Lemma C.5. Thus, again union bounding over all $j$, equation equation \text{D.6} gives with probability $1 - \delta$, for all $j \geq 0$ and $f^\phi(\alpha; \cdot) \in \mathcal{F}^{2,\phi}_{B_j}$.

\[
E_{(x,y) \sim p_{\text{data}}}[l(f^\phi(\alpha; x); y)] \leq \frac{1}{n} \sum_{i=1}^n l(f^\phi(\alpha; x_i); y_i) + M \sqrt{n\sqrt{d}} \max\{1, \|\alpha\|_2 \|X\|_F\} + c \sqrt{\log(1/\delta) + \log(\max\{1, \|\alpha\|_2 \|X\|_F\})/n}
\]

Now we assign the $\alpha$ to different $j$ as before to obtain the statement in the lemma. \qed
Note that if $l$ is some bounded loss such that $l(y; y) = 0$ (for example, truncated squared error), for $\alpha_{\ell_1}$ and $\alpha_{\ell_2}$, the loss terms over the datapoints (in the bounds of Lemmas D.4 and D.5) vanish. For loss $l$, define

$$L_{\ell_1\text{-reg}} \triangleq \mathbb{E}_{x, y \sim p_{\text{data}}} [l(f^\phi(\alpha_{\ell_1}; x); y)]$$

$$L_{\ell_2\text{-reg}} \triangleq \mathbb{E}_{x, y \sim p_{\text{data}}} [l(f^\phi(\alpha_{\ell_2}; x); y)]$$

Next, we will define the kernel matrix $K$ with $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$. Meanwhile, in the case that $\|K\|_{1\to 1} < \infty$, the loss terms over the datapoints (in the bounds of Lemmas D.4 and D.5) vanish. For any dataset\footnote{We sample $\beta \sim A$ as follows: first sample $\bar{u} \sim S^{d-1}$ uniformly. Then set $\beta$ to have a delta mass of 1 at $\bar{u}$ and be 0 everywhere else. Define the vector $v_u \triangleq (\varphi(\bar{u}^\top x_1), \ldots, \varphi(\bar{u}^\top x_n))$; then it follows that we set our labels $y$ to $v_u$. It is immediately clear that $\|\alpha_{\ell_1}\|_1 \leq \|\beta\|_1 \leq 1$. To lower bound $\mathbb{E}_{\beta \sim A}[\|\alpha_{\ell_1}\|_2^2]$, from Claim D.8 we get}

$$\mathbb{E}_{\beta \sim A}[\|\alpha_{\ell_1}\|_2^2] = E_{\bar{u} \sim S^{d-1}}[v_u^\top K^{-1} v_u]$$

$$= E_{\bar{u} \sim S^{d-1}}[\text{trace}(K^{-1}(v_u v_u^\top))]$$

$$= \text{trace}(K^{-1} E_{\bar{u} \sim S^{d-1}}[v_u v_u^\top])$$

$$= \frac{1}{n} \text{trace}(K^{-1})$$

(by definition of $K$)
Proof of Theorem \textbf{D.6} We note that since the distribution \( A \) of Lemma \textbf{D.7} does not depend on the dataset \( X \), it must hold that
\[
\mathbb{E}_{(x_i)_{i=1}^n \sim \text{unif}(0, I_{d \times d})} \left[ \mathbb{E}_{\beta \sim A} \left[ \| \alpha_{V_2 \ell_2} \|_2^2 \right] \right] = \frac{n}{k}
\]
Thus, there exists \( \beta^* \) such that if we sample \( X \) i.i.d. from the standard normal and set \( y_i = \langle \varphi(x_i), \beta^* \rangle \), the expectation of \( \| \alpha_{V_2 \ell_2} \|_2^2 \) is at least \( \frac{n}{k} \). We choose \( \rho_{\text{data}} \) corresponding to this \( \beta^* \), with \( x \) sampled from the standard normal. Now it is clear that \( \rho_{\text{data}} \) will satisfy the norm conditions of Theorem \textbf{D.6}.

For the generalization bounds, with high probability \( \| X \|_F = \Theta(\sqrt{nd}) \) as \( x \) is sampled from the standard normal distribution. Thus, Lemma \textbf{D.4} immediately gives the desired generalization error bounds for \( L_{\ell_1, \text{reg}} \). On the other hand, if \( \| \alpha_{V_2 \ell_2} \|_2 \geq \sqrt{\frac{n}{k}} \), then the bound of Lemma \textbf{D.5} is at least
\[
\frac{\sqrt{R} \| \alpha_{V_2 \ell_2} \|_2 \| X \|_F}{n \sqrt{d}} \geq \Omega(1)
\]

\( \square \)

E Missing Proofs in Section 4

E.1 Detailed Setup

We first write our regularity assumptions on \( \Phi, R, \) and \( V \) in more detail:

**Assumption E.1** (Regularity conditions on \( \Phi, R, V \)). \( R \) is nonnegative, Lipschitz, and smooth: \( \exists M_R, C_R \) such that \( \| \nabla^2 R \|_{\text{op}} \leq C_R \), and \( \| \nabla R \|_2 \leq M_R \).

**Assumption E.2**. \( \Phi \) is differentiable, bounded and Lipschitz on the sphere: \( \exists B_\Phi, M_\Phi \) such that \( \| \Phi(\theta) \| \leq B_\Phi \forall \theta \in S^d \), and \( |\Phi_t(\theta) - \Phi_t(\theta')| \leq M_\Phi \| \theta - \theta' \|_2 \forall \theta, \theta' \in S^d \).

**Assumption E.3**. \( V \) is Lipschitz and upper and lower bounded on the sphere: \( \exists b_V, B_V, M_V \) such that \( 0 < b_V \leq V(\theta) \leq B_V \forall \theta \in S^d \), and \( \| \nabla V(\theta) \|_2 \leq M_V \forall \theta \in S^d \).

We state the version of Theorem \textbf{4.3} that collects these parameters:

**Theorem E.4** (Theorem \textbf{4.3} with problem parameters). Suppose that \( \Phi \) and \( V \) are 2-homogeneous and Assumptions \textbf{D.2}, \textbf{E.1}, and \textbf{E.2} hold. Fix a desired error threshold \( \epsilon > 0 \). Suppose that from a starting distribution \( \rho_0 \), a solution to the dynamics in equation \textbf{4.2} exists. Choose
\[
\sigma \triangleq \exp(-d \log(1/\epsilon) \text{poly}(k, M_V, M_R, M_\Phi, b_V, B_V, C_R, B_\Phi, L[\rho_0] - L^*))
\]
\[
t_* \triangleq \frac{d^2}{\epsilon^2} \text{poly}(\log(1/\epsilon), k, M_V, M_R, M_\Phi, b_V, B_V, C_R, B_\Phi, L[\rho_0] - L^*)
\]
Then it must hold that \( \min_{0 \leq t \leq t_*} L[\rho_t] - \inf_{\rho} L[\rho] \leq 2\epsilon \).

E.2 Proof of Theorem \textbf{E.4}

Throughout the proof, it will be useful to keep track of \( W_t \triangleq \sqrt{\mathbb{E}_{\theta \sim \rho_t} \left[ \| \theta \|_2^2 \right]} \), the second moment of \( \rho_t \). We first introduce a general lemma on integrals over vector field divergences.

**Lemma E.5.** For any \( h_1 : \mathbb{R}^{d+1} \to \mathbb{R}, h_2 : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1} \) and distribution \( \rho \) with \( \rho(\theta) \to 0 \) as \( \| \theta \| \to \infty \),
\[
\int h_1(\theta) \nabla : (h_2(\theta) \rho(\theta)) d\theta = -\mathbb{E}_{\theta \sim \rho} \left[ \langle \nabla h_1(\theta), h_2(\theta) \rangle \right]
\]

**Proof.** The proof follows from integration by parts. \( \square \)

We note that \( \rho_t \) will satisfy the boundedness condition of Lemma \textbf{E.5} during the course of our algorithm - \( \rho_0 \) starts with this property, and Lemma \textbf{E.9} proves that \( \rho_t \) will continue to have this property. We therefore freely apply Lemma \textbf{E.5} in the remaining proofs. We first bound the absolute value of \( L'[\rho_t] \) over the sphere by \( B_L \triangleq M_R B_\Phi + B_V \).
Lemma E.6. For any $\bar{\theta} \in \mathbb{R}^{d-1}, t \geq 0$, $|L'[\rho_t](\bar{\theta})| \leq B_L$.

Proof. We compute
$$
|L'[\rho_t](\bar{\theta})| = \left| \nabla R \left( \int \Phi d\rho_t, \Phi(\bar{\theta}) \right) + V(\bar{\theta}) \right|
\leq \left| \nabla R \left( \int \Phi d\rho_t \right) \right|_2 \|\Phi(\bar{\theta})\|_2 + V(\bar{\theta}) \leq M_R B_\Phi + B_V
$$
\[\□\]

Now we analyze the decrease in $L[\rho_t]$.

Lemma E.7. Under the perturbed Wasserstein gradient flow
$$
\frac{d}{dt} L[\rho_t] = -\sigma \mathbb{E}_{\theta \sim \rho_t} [L'[\rho_t](\theta)] + \sigma \mathbb{E}_{\bar{\theta} \sim U^d} [L'[\rho_t](\bar{\theta})] - \mathbb{E}_{\theta \sim \rho_t} [||v[\rho_t](\theta)||^2_2]
$$

Proof. Applying the chain rule, we can compute
$$
\frac{d}{dt} L[\rho_t] = \nabla \cdot \int \Phi d\rho_t + \int V d\rho_t
\begin{aligned}
= \frac{d}{dt} \mathbb{E}_{\theta \sim \rho_t} [L'[\rho_t](\theta)] \\
= \int L'[\rho_t](\theta) d\rho_t + \int L'[\rho_t]dU^d - \int L'[\rho_t](\theta) \nabla \cdot (v[\rho_t](\theta) \rho_t(\theta)) d\theta \\
= -\sigma \mathbb{E}_{\theta \sim \rho_t} [L'[\rho_t](\theta)] + \sigma \mathbb{E}_{\bar{\theta} \sim U^d} [L'[\rho_t](\bar{\theta})] - \mathbb{E}_{\theta \sim \rho_t} [||v[\rho_t](\theta)||^2_2],
\end{aligned}
$$
where we use Lemma E.5 with $h_1 = L'[\rho_t]$ and $h_2 = v[\rho_t]$. \[\□\]

Now we show that the decrease in objective value is approximately the average velocity of all parameters under $\rho_t$ plus some additional noise on the scale of $\sigma$. At the end, we choose $\sigma$ small enough so that the noise terms essentially do not matter.

Corollary E.8. We can bound $\frac{d}{dt} L[\rho_t]$ by
$$
\frac{d}{dt} L[\rho_t] \leq \sigma B_L (W^2_t + 1) - \mathbb{E}_{\theta \sim \rho_t} [||v[\rho_t](\theta)||^2_2] \tag{E.1}
$$

Proof. By homogeneity, and Lemma E.6 $\mathbb{E}_{\theta \sim \rho_t} [L'[\rho_t](\theta)] = \mathbb{E}_{\bar{\theta} \sim \rho_t} [L'[\rho_t](\bar{\theta})] \leq B_L W^2_t$. We also get $\mathbb{E}_{\bar{\theta} \sim U^d} [L'[\rho_t](\bar{\theta})] \leq B_L$ since $U^d$ is only supported on $\mathbb{S}^d$. Combining these with Lemma E.7 gives the desired statement. \[\□\]

Now we show that if we run the dynamics for a short time, the second moment of $\rho_t$ will grow slowly, again at a rate that is roughly the scale of the noise $\sigma$.

Lemma E.9. For all $0 \leq t' \leq t$, $W^2_{t'} \leq \frac{L[\rho_0] + \sigma B_L}{b_\nu - t' B_L}$.

Proof. Let $t^* \triangleq \arg\max_{t' \in [0, t]} W^2_{t'}$. Integrating both sides of equation E.1 and rearranging, we get
$$
0 \leq \int_0^{t^*} \mathbb{E}_{\theta \sim \rho_s} [||v[\rho_s](\theta)||^2_2] ds \leq L[\rho_0] - L[\rho_{t^*}] + \sigma B_L \int_0^{t^*} (W^2_s + 1) ds \\
\leq L[\rho_0] - L[\rho_{t^*}] + t^* \sigma B_L (W^2_{t^*} + 1)
$$
Now since $R$ is nonnegative, we apply $L[\rho_{t^*}] \geq \mathbb{E}_{\theta \sim \rho_{t^*}} [V(\theta)] \geq \mathbb{E}_{\theta \sim \rho_{t^*}} [V(\theta)] \geq b_\nu W^2_{t^*}$. We now plug this in and rearrange to get $W^2_{t'} \leq W^2_{t^*} \leq \frac{L[\rho_0] + t^* \sigma B_L}{b_\nu - t^* B_L} \leq \frac{L[\rho_0] + t \sigma B_L}{b_\nu - t B_L} \quad \forall 0 \leq t' \leq t$. \[\□\]
Now let $W_t^2 \triangleq \frac{L[\rho_t] + \sigma t B_L}{\rho_t^{-1} - \sigma B_L}$. By Lemma E.9 \(0 \leq t \leq t_c, W_t^2 \leq W_t^2\).

The next statement allows us to argue that our dynamics will never increase the objective by too much.

**Lemma E.10.** For any $t_1, t_2$ with $0 \leq t_1 \leq t_2 \leq t_c$, $L[\rho_{t_2}] - L[\rho_{t_1}] \leq \sigma(t_2 - t_1)B_L(W_{t_c}^2 + 1)$.

**Proof.** From Corollary E.8 \(\forall t \in [t_1, t_2]\) we have

$$\frac{d}{dt}L[\rho_t] \leq \sigma B_L(W_{t_c}^2 + 1)$$

Integrating from $t_1$ to $t_2$ gives the desired result. \(\square\)

The following lemma bounds the change in expectation of a 2-homogeneous function over $\rho_t$. At a high level, we lower bound the decrease in our loss as a function of the change in this expectation.

**Lemma E.11.** Let $h : \mathbb{R}^{d+1} \to \mathbb{R}$ that is 2-homogeneous, with $\|\nabla h(\theta)\| \leq M \forall \theta \in \mathbb{S}^d$ and $|h(\theta)| \leq B \forall \theta \in \mathbb{S}^d$. Then $\forall 0 \leq t \leq t_c$, we have

$$\left| \frac{d}{dt} \int h d\rho_t \right| \leq \sigma B(W_{t_c}^2 + 1) + MW\left(-\frac{d}{dt}L[\rho_t] + \sigma B_L(W_{t_c}^2 + 1)\right)^{1/2} \quad (E.2)$$

**Proof.** Let $Q(t) \triangleq \int h d\rho_t$. We can compute:

$$Q'(t) = \int h(\theta)\frac{d\rho_t}{dt}(\theta)d\theta$$

$$= \int h(\theta)(-\sigma \rho_t(\theta) - \nabla \cdot (v[\rho_t](\theta)\rho_t(\theta)))d\theta + \sigma \int h dU^d$$

$$= -\sigma \int h(\theta)||\theta||_2^2 \rho_t(\theta)d\theta + \sigma \int h dU^d - \int h(\theta)\nabla \cdot (v[\rho_t](\theta)\rho_t(\theta))d\theta \quad (E.3)$$

Note that the first two terms are bounded by $\sigma B(W_{t_c}^2 + 1)$ by the assumptions for the lemma. For the third term, we have from Lemma E.5

$$\left| \int h(\theta)\nabla \cdot (v[\rho_t](\theta)\rho_t(\theta))d\theta \right| = \left| E_{\theta \sim \rho_t}[\langle \nabla h(\theta), v[\rho_t](\theta) \rangle] \right|$$

$$\leq \sqrt{E_{\theta \sim \rho_t}[\|\nabla h(\theta)\|^2] E_{\theta \sim \rho_t}[\|v[\rho_t](\theta)\|^2]} \quad \text{(by Cauchy-Schwarz)}$$

$$\leq \sqrt{E_{\theta \sim \rho_t}[\|\nabla h(\theta)\|^2] E_{\theta \sim \rho_t}[\|v[\rho_t](\theta)\|^2]} \quad \text{(by homogeneity of } \nabla h \text{)}$$

$$\leq MW_e \sqrt{E_{\theta \sim \rho_t}[\|v[\rho_t](\theta)\|^2]} \quad \text{(since } h \text{ is Lipschitz on the sphere)}$$

$$\leq MW_e \left(-\frac{d}{dt}L[\rho_t] + \sigma B_L(W_{t_c}^2 + 1)\right)^{1/2} \quad \text{(by Corollary E.8)}$$

Plugging this into equation E.3 we get that

$$|Q'(t)| \leq \sigma B(W_{t_c}^2 + 1) + MW_e \left(-\frac{d}{dt}L[\rho_t] + \sigma B_L(W_{t_c}^2 + 1)\right)^{1/2} \quad \square$$

We apply this result to bound the change in $L'[\rho_t]$ over time in terms of the change of the objective value. For clarity, we write the bound in terms of $c_1$ that is some polynomial in the problem constants.

**Lemma E.12.** Define $Q(t) \triangleq \int \Phi d\rho_t$. For every $\theta \in \mathbb{S}^d$ and $0 \leq t \leq t + l \leq t_c$, $\exists c_1 \triangleq \text{poly}(k, C_R, B_0, M_\Phi, B_L)$ such that

$$|L'[\rho_t](\theta) - L'[\rho_{t+l}](\theta)| \leq C_R B_\Phi \int_t^{t+l} ||Q'(t)||_1$$

$$\leq \sigma c_1(W_{t_c}^2 + 1) + c_1 W_e \sqrt{L[\rho_t] - L[\rho_{t+l}]} + \sigma c_1(W_{t_c}^2 + 1)^{1/2} \quad (E.5)$$
Proof. Recall that $L'[ho_t](\tilde{\theta}) = \langle \nabla R(\int \Phi d\rho_t), \Phi(\tilde{\theta}) \rangle + V(\tilde{\theta})$. Differentiating with respect to $t$,
\[
\frac{d}{dt} L'[ho_t](\tilde{\theta}) = \left\langle \frac{d}{dt} \nabla R \left( \int \Phi d\rho_t \right), \Phi(\tilde{\theta}) \right\rangle
= \Phi(\tilde{\theta})^\top \nabla^2 R(Q(t)) Q'(t)
\leq C_R B_\Phi \|Q'(t)\|_2
\leq C_R B_\Phi \|Q'(t)\|_1
\tag{E.6}
\]
Integrating and applying the same reasoning to $-L'[ho_t]$ gives us equation [E.4]. Now we apply Lemma [E.11] to get
\[
\|Q'(t)\|_1 = \sum_{i=1}^k \left| \frac{d}{dt} \int \Phi_i d\rho_t \right| \leq \sum_{i=1}^k \left[ \sigma B_\Phi (W_e^2 + 1) + M_\Phi W_e \left( -\frac{d}{dt} L[\rho_t] + \sigma B_L (W_e^2 + 1) \right)^{1/2} \right]
\leq k \sigma B_\Phi (W_e^2 + 1) + k M_\Phi W_e \left( -\frac{d}{dt} L[\rho_t] + \sigma B_L (W_e^2 + 1) \right)^{1/2}
\]
We plug this into equation [E.6] and then integrate both sides to obtain
\[
C_R B_\Phi \int_t^{t+i} \|Q'(t)\|_1 \leq k \sigma C_R B_\Phi^2 (W_e^2 + 1) + k C_R B_\Phi M_\Phi W_e \int_t^{t+i} \left( -\frac{d}{dt} L[\rho_t] + \sigma B_L (W_e^2 + 1) \right)^{1/2}
\]
Now we also show that $L'$ is Lipschitz on the unit ball. For clarity, we let $c_2 \triangleq \sqrt[k]{C_R M_\Phi + M_V}$.

Lemma E.13. For all $\tilde{\theta}, \tilde{\theta}' \in S^d$,
\[
|L'[ho](\tilde{\theta}) - L'[ho](\tilde{\theta}')| \leq c_2 \|\tilde{\theta} - \tilde{\theta}'\|_2
\tag{E.7}
\]
Proof. Using the definition of $L'$ and triangle inequality,
\[
|L'[ho](\tilde{\theta}) - L'[ho](\tilde{\theta}')| \leq \left\| \nabla R \left( \int \Phi d\rho \right) \right\|_2 \|\Phi(\tilde{\theta}) - \Phi(\tilde{\theta}')\|_2 + |V(\tilde{\theta}) - V(\tilde{\theta}')|
\leq (\sqrt[k]{C_R M_\Phi + M_V}) \|\tilde{\theta} - \tilde{\theta}'\|_2 \quad \text{(by definition of $M_\Phi, M_R, M_V$)}
\]
Now the remainder of the proof will proceed as follows: we show that if $\rho_t$ is far from optimality, either the expected velocity of $\theta$ under $\rho_t$ will be large in which case the loss decreases from Corollary [E.8] or there will exist $\tilde{\theta}$ such that $L'[\rho_t](\tilde{\theta}) \leq 0$. We will first show that in the latter case, the $\sigma u$ noise term will grow mass exponentially fast in a descent direction until we make progress. Define $K^{-\tau} \triangleq \{ \tilde{\theta} \in S^d : L'[\rho_t](\tilde{\theta}) \leq -\tau \}$, the $-\tau$-sublevel set of $L'[\rho_t]$, and let $m(S) \triangleq \mathbb{E}_{\theta \sim U} |1(\theta \in S)|$ be the normalized spherical area of the set $S$.

Lemma E.14. If $K^{-\tau}$ is nonempty, for $0 \leq \delta \leq \tau$, $\log m(K^{-\tau+\delta}) \geq -2d \log \frac{c_2}{c_2}$.
Proof. Let $\tilde{\theta} \in K^{-\tau}$. From Lemma [E.13] $L'[\rho](\tilde{\theta}') \leq -\tau + \delta$ for all $\tilde{\theta}'$ with $\|\tilde{\theta}' - \tilde{\theta}\|_2 \leq \frac{\delta}{c_2}$. Thus, we have
\[
m(K^{-\tau+\delta}) \geq \mathbb{E}_{\tilde{\theta} \sim U} \left[ 1(\|\tilde{\theta}' - \tilde{\theta}\|_2 \leq \frac{\delta}{c_2}) \right]
\]
Now the statement follows by Lemma 2.3 of [Ball et al. 1997].
Now we show that if a descent direction exists, the added noise will find it and our function value will decrease. We start with a general lemma about the magnitude of the gradient of a 2-homogeneous function in the radial direction.

**Lemma E.15.** Let \( h : \mathbb{R}^{d+1} \rightarrow \mathbb{R} \) be a 2-homogeneous function. Then for any \( \theta \in \mathbb{R}^{d+1} \), \( \theta^T \nabla h(\theta) = 2\|\theta\|_2 h(\theta) \).

**Proof.** We have \( h(\theta + \alpha \bar{\theta}) = (\|\theta\|_2 + \alpha)^2 h(\bar{\theta}) \). Differentiating both sides with respect to \( \alpha \) and evaluating the derivative at 0, we get \( \theta^T \nabla h(\theta) = 2\|\theta\|_2 h(\bar{\theta}) \), as desired. \( \square \)

We state the lemma claiming that our objective will decrease if \( L'[\rho_t](\bar{\theta}) \leq 0 \) for some \( \bar{\theta} \in \mathbb{S}^d \).

**Lemma E.16.** Choose

\[
\ell \geq \frac{\log(W_e^2/\sigma)}{\tau - \sigma} + 2d \log \frac{2\pi}{\tau} + 1
\]

If \( K_{t^*} \) is nonempty for some \( t^* \) satisfying \( t^* + \ell \leq t_e \), then after \( \ell \) steps, we will have

\[
L_t[\rho_{t+s}] \leq L_t[\rho_t] - \frac{(\tau/4 - \sigma c_1(W_e^2 + 1))^2}{4\sigma^2 c_1} + \sigma c_1(W_e^2 + 1)
\]

(E.8)

We will first show that a descent direction in \( L'[\rho_t] \) will remain for the next \( \ell \) time steps. In the notation of Lemma E.12, we define \( z(s) \equiv C_l B_{c_1} \int_{t^*}^{t^*+s} \|Q(t)\|_1 dt \). Note that from Lemma E.12, for all \( \bar{\theta} \in \mathbb{S}^d \) we have \( L'[ho_{t+s}](\bar{\theta}) = L'[ho_t](\bar{\theta}) + z(s) \). Thus, the following holds:

**Claim E.17.** For all \( s \leq \ell \), \( K_{t^*+s} \) is nonempty.

**Proof.** By assumption, \( \exists \bar{\theta} \in K_{t^*} \). Then \( L'[ho_{t+s}](\bar{\theta}) = L'[ho_t](\bar{\theta}) + z(s) \leq -\tau + z(s) \), so \( K_{t^*+s} \) is nonempty. \( \square \)

Let \( T_s \equiv K_{t^*+2s} \) for \( 0 \leq \ell \leq \ell \). We now argue that this set \( T_s \) does not shrink as \( t \) increases.

**Claim E.18.** For all \( s' > s \), \( T_{s'} \supseteq T_s \).

**Proof.** From equation E.6 and the definition of \( z(s) \), \( |L'[ho_{t+s}](\bar{\theta}) - L'[ho_{t+s}](\bar{\theta})| \leq z(s') - z(s) \). It follows that for \( \bar{\theta} \in T_s \)

\[
L'[ho_{t+s}](\bar{\theta}) \leq L'[ho_{t+s}](\bar{\theta}) + z(s') - z(s)
\]

\[
\leq -\tau/2 + z(s) - z(s) + z(s')
\]

(by definition of \( T_s \))

\[
\leq -\tau/2 + z(s')
\]

which means that \( \bar{\theta} \in T_{s'} \). \( \square \)

Now we show that the weight of the particles in \( T_s \) grows very fast if \( z(k) \) is small.

**Claim E.19.** Suppose that \( z(l) \leq \tau/4 \). Let \( \bar{T}_s \equiv \{ \theta \in \mathbb{R}^{d+1} : \bar{\theta} \in T_s \} \). Define \( N(s) \equiv \frac{1}{\int_{T_s} \|\theta\|^2 d\rho_{t+s}} \) and \( \beta \equiv \exp(-2d \log \frac{2\pi}{\tau}) \). Then \( N(s) \geq (\tau - \sigma) N(s) + \sigma \beta \).

**Proof.** From the assumption \( z(l) \leq \tau/4 \), it holds that \( T_s \subseteq K_{t^*/2} \forall s \leq k \). Since \( T_s \) is defined as a sublevel set, \( v[\rho_{t+s}](\bar{\theta}) \) points inwards on the boundary of \( T_s \) for all \( \bar{\theta} \in T_s \), and by 1-homogeneity of the gradient, the same must hold for all \( u \in T_s \).

Now consider any particle \( \theta \in \bar{T}_s \). We have that \( \theta \) flows to \( \theta + v[\rho_{t+s}](\bar{\theta}) \) at time \( t^* + s + ds \). Furthermore, since the gradient points inwards from the boundary, it also follows that \( u + v[\rho_{t+s}](\bar{\theta}) ds \in T_s \). Now we compute

\[
\int_{\bar{T}_s} \|\theta\|^2 d\rho_{t^*+s} ds = (1 - \sigma ds) \int_{\bar{T}_s} \|\theta + v[\rho_{t^*+s}](\bar{\theta}) ds\|^2 d\rho_{t^*+s} + \sigma ds \int_{\bar{T}_s} 1 dU
\]

\[
\geq (1 - \sigma ds) \int_{\bar{T}_s} \left( \|\theta\|^2 + 2\theta^T v[\rho_{t^*+s}](\bar{\theta}) ds \right) d\rho_{t^*+s} + \sigma m(K_{t^*/2} + z(s)) ds
\]

(E.9)
Now we apply Lemma E.15 using the 2-homogeneity of $F'$ and the fact that $L'[\rho_{t+s}](\theta) \leq -\tau/4 \nexists \theta \in T_s$

$$\|\theta\|^2_2 + 2\theta^T v[\rho_{t+s}](\theta)ds = \|\theta\|^2_2 - 4\|\theta\|^2_2 L'[\rho_{t+s}](\theta)ds$$

$$\geq \|\theta\|^2_2 (1 + \tau ds)$$

(E.10)

Furthermore, since $K_{t+s}^{-\tau/2}(s)$ is nonempty by Claim E.17 we can apply Lemma E.14 and obtain

$$m(K_{t+s}^{-\tau/2}(s)) \geq \beta$$

(E.11)

Plugging equation E.10 and equation E.11 back into equation E.9, we get

$$\int_{F_s} u_2^2 d\rho_{t+s+ds} \geq (1 - \sigma ds)(1 + 2\tau ds)N(s) + \sigma \beta ds$$

Since we also have that $\hat{T}_{t+s+ds} \supseteq \hat{T}_s$, it follows that

$$N(s + ds) = \int_{\hat{T}_{t+s+ds}} u_2^2 d\rho_{t+s+ds} \geq (1 - \sigma ds)(1 + \tau ds)N(s) + \sigma \beta ds$$

and so $N'(s) \geq (\tau - \sigma)N(s) + \sigma \beta$.

Now we are ready to prove Lemma E.16.

**Proof of Lemma E.16.** If $z(l) = C_l B_\phi f_{t+t} \|Q'(t)\|_1 \geq \frac{\tau}{4}$, then by rearranging the conclusion of Lemma E.12 we immediately get equation E.8.

Suppose for the sake of contradiction that $z(l) \leq \tau/4$. From Claim E.19 it follows that $N(1) \geq \sigma \beta$, and $N(l) \geq \exp((\tau - \sigma)(l - 1))N(1)$. Thus, in $\frac{\log(W^2/l\sigma) + 2d log \frac{2d}{\epsilon}}{\tau - \sigma} + 1$ time, $W_{t+1} \geq N(l) \geq W_\epsilon^2$, a contradiction. Therefore, it must be true that $z(l) \geq \tau/4$.

The following lemma will be useful in showing that the objective will decrease fast when $\rho_i$ is very suboptimal.

**Lemma E.20.** For any time $t$ with $0 \leq t \leq t_\epsilon$, we have

$$\frac{d}{dt} L[\rho_t] \leq \sigma B_t(W_\epsilon^2 + 1) - \frac{\mathbb{E}_{\theta \sim \rho_t}[L'[\rho_t](\theta)]^2}{W_\epsilon^2}$$

(E.12)

**Proof.** We can first compute

$$\mathbb{E}_{\theta \sim \rho_t}[L'[\rho_t](\theta)] = \mathbb{E}_{\theta \sim \rho_t}[L'[\rho_t](\theta) \|\theta\|^2]$$

$$= \frac{1}{2} \mathbb{E}_{\theta \sim \rho_t}[\|\theta\|^2 \theta^Tv[\rho_t](\theta)]$$

$$\leq \frac{1}{2} \mathbb{E}_{\theta \sim \rho_t}[\|\theta\|^2] \mathbb{E}_{\theta \sim \rho_t}[\|v[\rho_t](\theta)\|^2]$$

(by Cauchy-Schwarz)

$$\leq \frac{1}{2} W_\epsilon \sqrt{\mathbb{E}_{\theta \sim \rho_t}[\|v[\rho_t](\theta)\|^2]}$$

Rearranging gives $\mathbb{E}_{\theta \sim \rho_t}[\|v[\rho_t](\theta)\|^2] \geq \frac{\mathbb{E}_{\theta \sim \rho_t}[L'[\rho_t](\theta)]^2}{W_\epsilon^2}$, and plugging this into equation E.1 gives the desired result.

**Proof of Theorem E.4.** Let $L^*$ denote the infimum $\inf_{\rho} L[\rho]$, and let $\rho^*$ be an $\epsilon$-approximate global minimizer of $L$: $L[\rho^*] \leq L^* + \epsilon$. (We define $\rho^*$ because a true minimizer of $L$ might not exist.) Let $W^* \triangleq \mathbb{E}_{\theta \sim \rho^*}[\|\theta\|^2]$. We first note that since $bV W^2 \leq L[\rho_*] \leq L[\rho_0], W^2 \leq L[\rho_0]/bV \leq W^*_\epsilon$.

Now we bound the suboptimality of $\rho_i$: since $L$ is convex in $\rho$,

$$L[\rho^*] \geq L[\rho_t] + \mathbb{E}_{\theta \sim \rho_t}[L'[\rho_t](\theta)] - \mathbb{E}_{\theta \sim \rho_t}[L'[\rho_t](\theta)]$$

32
Rearranging gives

\[
L[\rho_t] - L[\rho^*] \leq \mathbb{E}_{\theta \sim p_{\theta}}[L'[\rho_t](\theta)] - \mathbb{E}_{\theta \sim p_{\theta}^*}[L'[\rho_t](\theta)]
\]

\[
\leq \mathbb{E}_{\theta \sim p_{\theta}}[L'[\rho_t](\theta)] - W^2 \min_{\bar{\theta} \in \mathcal{S}^{d-1}} \left\{ \min_{\theta \in \mathcal{S}^{d-1}} L'[\rho_t](\bar{\theta}), 0 \right\}
\]  \tag{E.13}

Now let \( t \equiv \frac{W}{2/\epsilon^2} (2 \log \frac{W^2}{\sigma^2} + 2d \log \frac{4W^2 c^2}{\epsilon^2}) \), which satisfies Lemma E.16 with the value of \( \tau \) later specified. Suppose that there is a \( t \) with \( 0 \leq t \leq t_\varepsilon - 2l \) and \( \forall t' \in [t, t + l] \), \( L[\rho_t] - L[\rho^*] \geq 2\epsilon \). Then \( L[\rho_t] - L[\rho^*] \geq \epsilon \). We will argue that the objective decreases when we are \( \epsilon \)-optimal:

\[
L[\rho_t] - L[\rho_{t+2l}] \geq \epsilon \frac{8W^2 - l \sigma c_1(W^2 + 1)}{c^2_W l} - 3\sigma l c_1(W^2 + 1)
\]  \tag{E.15}

Using equation E.13 and \( W_\varepsilon \geq W^* \), we first note that

\[
\epsilon \leq \mathbb{E}_{\theta \sim p_{\theta}^*}[L'[\rho_t](\theta)] - W^2 \min_{\bar{\theta} \in \mathcal{S}^{d-1}} \left\{ \min_{\theta \in \mathcal{S}^{d-1}} L'[\rho_t](\bar{\theta}), 0 \right\}
\]

Thus, either \( \min_{\bar{\theta} \in \mathcal{S}^{d-1}} L'[\rho_t](\bar{\theta}) \leq -\frac{W^2 \epsilon}{2/\epsilon^2} \leq -\frac{W^2 \epsilon}{4/\epsilon^2} \), or \( \mathbb{E}_{\theta \sim p_{\theta}^*}[L'[\rho_t](\theta)] \geq \frac{\epsilon}{2} \). If \( \exists t' \in [t, t + l] \) such that the former holds, then the \( \tau \equiv \frac{W^2 \epsilon}{4/\epsilon^2} \) sub-level set \( K_{\tau}^{E.10} \) is non-empty. Applying Lemma E.16 gives

\[
L[\rho_t] - L[\rho_{t+l}] \geq \left( \frac{\epsilon}{2} \frac{W^2 - l \sigma c_1(W^2 + 1)}{c^2_W l} \right) - 3\sigma l c_1(W^2 + 1)
\]  \tag{E.16}

In the second case \( \mathbb{E}_{\theta \sim p_{\theta}^*}[L'[\rho_t](\theta)] \geq \frac{\epsilon}{2} \), \( \forall t' \in [t, t + l] \). Therefore, we can integrate equation E.12 from \( t \) to \( t + l \) in order to get

\[
L[\rho_t] - L[\rho_{t+l}] \geq t \frac{\epsilon^2}{4W^2} - \sigma l B_L(W^2 + 1)
\]

Therefore, applying Lemma E.10 again gives

\[
L[\rho_t] - L[\rho_{t+2l}] \geq l \frac{\epsilon^2}{4W^2} - 2\sigma l B_L(W^2 + 1)
\]  \tag{E.17}

Thus equation E.15 follows.

Now recall that we choose

\[
\sigma \equiv \exp(-d \log(1/\epsilon)) \text{poly}(k, M_V, M_R, M_B, b_V, b_R, B_V, B_R, B_{\theta}, L[\rho_0] - L[\rho^*])
\]

For the simplicity, in the remaining notation, we will use \( O(\cdot) \) notation to hide polynomials in the problem parameters besides \( d, \epsilon \). We simply write \( \sigma = \exp(-c_3 d \log(1/\epsilon)) \). Recall our choice \( t_\varepsilon \equiv O\left(\frac{\epsilon}{\sigma} \log^2(1/\epsilon)\right) \). It suffices to show that our objective would have sufficiently decreased in \( t_\varepsilon \) steps. We first note that with \( c_3 \) sufficiently large, \( W_\varepsilon^2 = O(L[\rho_0]/b_\varepsilon) = O(1) \). Simplifying our expression for \( l \), we get that \( l = O\left(\frac{d}{\epsilon} \log \frac{1}{\epsilon}\right) \), so long as \( \sigma W^2 = o(\epsilon) \), which holds for sufficiently large \( c_3 \). Now let

\[
\delta_1 \equiv \left( \frac{\epsilon}{4W^2} - l \sigma c_1(W^2 + 1) \right) - 3\sigma l c_1(W^2 + 1)
\]

\[
\delta_2 \equiv \frac{\epsilon^2}{4W^2} - 2\sigma l B_L(W^2 + 1)
\]

Again, for sufficiently large \( c_3 \), the terms with \( \sigma \) become negligible, and \( \delta_1 = \frac{\epsilon^2}{2W^2} \equiv O\left(\frac{\epsilon^2}{d \log(1/\epsilon)}\right) \). Likewise, \( \delta_2 = O\left(\epsilon d \log(1/\epsilon)\right) \).

Thus, if by time \( t \) we have not encountered \( 2\epsilon \)-optimal \( \rho_t \), then we will decrease the objective by \( O\left(\frac{\epsilon^2}{d \log(1/\epsilon)}\right) \) in \( O\left(\frac{d}{\epsilon} \log \frac{1}{\epsilon}\right) \) time. Therefore, a total of \( O\left(\frac{\epsilon^2}{d \log(1/\epsilon)} \log^2(1/\epsilon)\right) \) time is sufficient to obtain \( \epsilon \) accuracy. \( \square \)
### Table 1: Test error on CIFAR10 and CIFAR100 for initial $\lambda = 0.0005$

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight decay annealing</td>
<td>5.86</td>
<td>26.22</td>
</tr>
<tr>
<td>Fixed weight decay</td>
<td>6.01</td>
<td>27.00</td>
</tr>
</tbody>
</table>

#### E.3 DISCRETE-TIME OPTIMIZATION

To circumvent the technical issue of existence of a solution to the continuous-time dynamics, we also note that polynomial time convergence holds for discrete-time updates.

**Theorem E.21.** Along with Assumptions E.2, E.2, E.3 additionally assume that $\nabla \Phi_i$ and $\nabla V$ are $C_\Phi$ and $C_V$-Lipschitz, respectively. Let $\rho_t$ evolve according to the following discrete-time update:

$$
\rho_{t+1} \triangleq \rho_t + \eta \left( -\sigma \rho_t + \sigma U^d - \nabla \cdot (v(\rho_t) \rho_t) \right)
$$

There exists a choice of

- $\sigma \triangleq \exp(-d \log(1/\epsilon) \text{poly}(k, M_V, M_R, b_V, B_V, C_R, B_\Phi, C_\Phi, C_V, \mathcal{L}[\rho_0] - \mathcal{L}[\rho^*]))$
- $\eta \triangleq \text{poly}(k, M_V, M_R, b_V, B_V, C_R, B_\Phi, C_\Phi, C_V, \mathcal{L}[\rho_0] - \mathcal{L}[\rho^*])$
- $t_\epsilon \triangleq \frac{d^2}{\epsilon^2} \text{poly}(k, M_V, M_R, b_V, B_V, C_R, B_\Phi, C_\Phi, C_V, \mathcal{L}[\rho_0] - \mathcal{L}[\rho^*])$

such that $\min_{0 \leq t \leq t_\epsilon} \mathcal{L}[\rho_t] - L^* \leq \epsilon$.

The proof follows from a standard conversion of the continuous-time proof of Theorem E.4 to discrete time, and we omit it here for simplicity.

#### F ADDITIONAL EXPERIMENTS

We train a modified WideResNet architecture (Zagoruyko & Komodakis, 2016) on CIFAR10 and CIFAR100. Our theory does not entirely apply because the identity mapping prevents ResNet architectures from being homogeneous, but our experiments show that reducing weight decay can still help generalization error in this setting. Because batchnorm can cause the regularizer to have different effects (van Laarhoven, 2017), we remove batchnorm layers and train a 16-layer deep WideResNet. We again compare a network trained with weight decayed annealing to one trained without annealing. We used a fixed learning rate schedule that starts at 0.1 and decreases by a factor of 0.2 at epochs 60, 120, and 160. For CIFAR10, we use an initial weight decay of 0.0002 and decrease the weight decay by 0.2 at epoch 60, and then by 0.5 at epochs 90, 120, 140, 160. For CIFAR100, we initialize weight decay at 0.0005 and decrease it by 0.2 at epochs 60, 120, and 160. We tried different parameters for the initial weight decay and chose the ones that worked best for the model without annealing. We also tried using small weight decays at initialization, but these models failed to generalize well – we believe this is due to an optimization issue where the algorithm fails to find a true global minimum of the regularized loss. We believe that annealing the weight decay directs the optimization algorithm closer towards the global minima for small $\lambda$.

Table 1 shows the test error achieved by models with and without annealing. We see that the simple change of annealing decay can decrease the test error for this architecture.