Generalization Guarantees for Neural Architecture Search with Train-Validation Split

Samet Oymak* Mingchen Li† Mahdi Soltanolkotabi‡

April 28, 2021

Abstract

Neural Architecture Search (NAS) is a popular method for automatically designing optimized architectures for high-performance deep learning. In this approach, it is common to use bilevel optimization where one optimizes the model weights over the training data (lower-level problem) and various hyperparameters such as the configuration of the architecture over the validation data (upper-level problem). This paper explores the statistical aspects of such problems with train-validation splits. In practice, the lower-level problem is often overparameterized and can easily achieve zero loss. Thus, a-priori it seems impossible to distinguish the right hyperparameters based on training loss alone which motivates a better understanding of the role of train-validation split. To this aim this work establishes the following results:
• We show that refined properties of the validation loss such as risk and hyper-gradients are indicative of those of the true test loss. This reveals that the upper-level problem helps select the most generalizable model and prevent overfitting with a near-minimal validation sample size. Importantly, this is established for continuous spaces – which are highly relevant for popular differentiable search schemes.
• We establish generalization bounds for NAS problems with an emphasis on an activation search problem. When optimized with gradient-descent, we show that the train-validation procedure returns the best (model, architecture) pair even if all architectures can perfectly fit the training data to achieve zero error.
• Finally, we highlight rigorous connections between NAS, multiple kernel learning, and low-rank matrix learning. The latter leads to novel algorithmic insights where the solution of the upper problem can be accurately learned via efficient spectral methods to achieve near-minimal risk.

1 Introduction

Hyperparameter optimization (HPO) is a critical component of modern machine learning pipelines. It is particularly important for deep learning applications where there are many possibilities for choosing a variety of hyperparameters to achieve the best test accuracy. A crucial hyperparameter for deep learning is the architecture of the network. The architecture encodes the flow of information from the input to output, which is governed by the network's graph and the set of nonlinear operations that transform hidden feature representations. In this case HPO is often referred to as Neural Architecture Search (NAS). NAS is critical to finding the most suitable architecture in an automated manner without extensive user trial and error.

HPO/NAS problems are often formulated as bilevel optimization problems and critically rely on a train-validation split of the data, where the parameters of the learning model (e.g. weights of the neural network)

*Department of Electrical and Computer Engineering, University of California, Riverside, CA. Email: soymak@ucr.edu
†Department of Computer Science and Engineering, University of California, Riverside, CA. Email: mli176@ucr.edu
‡Ming Hsieh Dept. of Electrical Engineering, University of Southern California, Los Angeles, CA. Email: soltanol@usc.edu

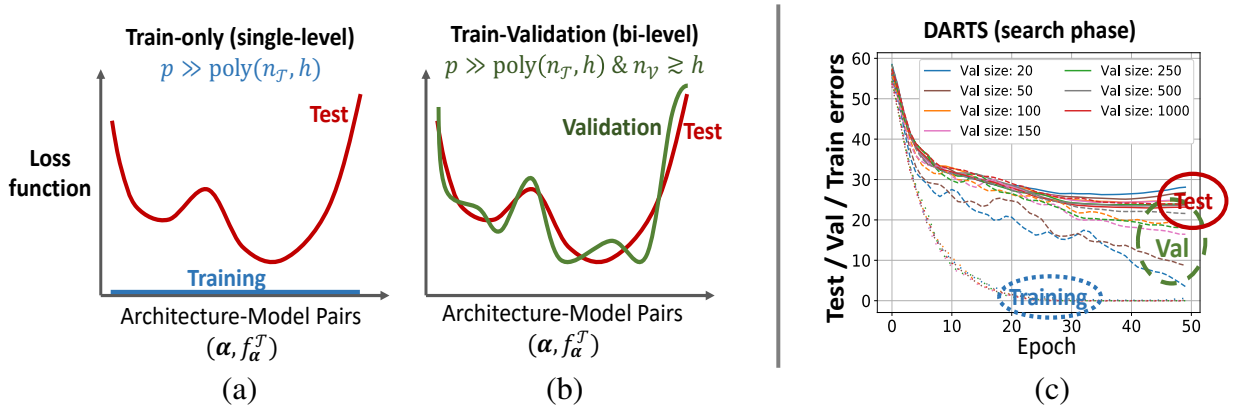


Figure 1: This figure depicts the typical scenario arising in modern hyperparameter optimization problems –such as those arising in NAS– involving p parameters, h hyperparameters and $n_{\mathcal{T}}$ training data. In these modern settings, as depicted in Figure (a), the network is expressive enough so that it can perfectly fit the training data for all choices of (continuously-parameterized) architectures simultaneously. However, the network exhibits very different test performance for different hyperparameter choices. Perhaps surprisingly, as depicted in Figure (b), when we solve the problem via a train-validation split, as long as the amount of validation data $n_{\mathcal{V}}$ is comparable to the number of hyperparameters h (i.e. $n_{\mathcal{V}} \gtrsim h$), the validation loss uniformly concentrates around the test loss and enables the discovery of the optimal model even if the training loss is uniformly zero over all architectures. This paper aims to rigorously establish this phenomena (e.g., see our Theorem 3). Figure (c) shows NAS experiments on DARTS during the search phase. We evaluate train/test/validation losses of the continuously-parameterized supernet. These experiments are consistent with the depictions in Figures (a) and (b) and our theory: Training error is consistently zero after 30 epochs. Validation error is never zero even with extremely few (20 or 50) samples. Validation error almost perfectly tracks test error as soon as validation size is mildly large (e.g. ≥ 250) which is also comparable with the number of architectural parameters ($h = 224$).

are optimized over the training data (lower-level problem), and the hyperparameters are optimized over a validation data (upper-level problem). With an ever growing number of configurations/architecture choices in modern learning problems, there has been a surge of interest in differentiable HPO methods that focus on continuous hyperparameter relaxations. For instance, differentiable architecture search schemes often learn continuously parameterized architectures which are discretized only at the end of the training [44]. Similar techniques have also been applied to learning data-augmentation policies [20] and meta-learning [27, 25]. These differentiable algorithms are often much faster and seamlessly scale to millions of hyperparameters [45]. However, the generalization capability of HPO/NAS with such large search spaces and the benefits of the train-validation split on this generalization remain largely mysterious.

Addressing the above challenge is particularly important in modern overparameterized learning regimes where the training loss is often not indicative of the model's performance as large networks with many parameters can easily overfit to training data and achieve zero loss. To be concrete, let $n_{\mathcal{T}}$ and $n_{\mathcal{V}}$ denote the training and validation sample sizes and p and h the number parameters and hyperparmeters of the model. In deep learning, NAS and HPO problems typically operate in a regime where

$$\boxed{p := \#\text{ params.} \geq n_{\mathcal{T}} \geq n_{\mathcal{V}} \geq h := \#\text{ hyperparams.}} \tag{1.1}$$

Figure 1 depicts such a regime (e.g. when $p \gg \text{poly}(n_{\mathcal{T}})$)) where the neural network model is in fact expressive enough to perfectly fit the dataset *for all possible combinations of hyperparameters*. Nevertheless, training with a train-validation split tends to select the right hyperparameters where the corresponding

network achieves stellar test accuracy. This leads us to the main challenge of this paper[1]:

How does train-validation split for NAS/HPO over large continuous search spaces discover near-optimal hyperparameters that generalize well despite the overparameterized nature of the problem?

To this aim, in this paper, we explore the statistical aspects of NAS with train-validation split and provide theoretical guarantees to explain its generalization capability in the practical data/parameter regime of (1.1). Specifically, our contributions and the basic outline of the paper are as follows:

• **Generalization with Train-Validation Split (Section 3):** We provide general-purpose uniform convergence arguments to show that refined properties of the validation loss (such as risk and hyper-gradients) are indicative of the test-time properties. This is shown when the lower-level of the bilevel train-validation problem is optimized by an algorithm which is (approximately) Lipschitz with respect to the hyperparameters. Our result applies as soon as the validation sample size scales proportionally with the *effective dimension of the hyperparameter space and only logarithmically in this Lipschitz constant*. We then utilize this result to obtain an end-to-end generalization bound for bilevel optimization with train-validation split under generic conditions. We also show that the aforementioned Lipschitzness condition holds in a variety of settings, such as: When the lower-level problem is strongly convex (the ridge regularization strength is allowed to be one of the hyperparameters) as well as a broad class of kernel and neural network learning problems (not necessarily strongly-convex) discussed next.

• **Generalization Guarantees for NAS (Sections 4 & 5):** We also develop guarantees for NAS problems. Specifically, we first develop results for a *neural activation search* problem that aims to determine the best activation function (among a continuously parameterized family) for overparameterized shallow neural network training. We study this problem in connection to a *feature-map/kernel learning* problem involving the selection of the best feature-map among a continuously parameterized family of feature-maps. Furthermore, when the lower-level problem is optimized via gradient descent, we show that the *bilevel problem is guaranteed to select the activation that has the best generalization capability*. This holds despite the fact that with any choice of the activation, the network can perfectly fit the training data. We then extend our results to deeper networks by similarly linking the problem of finding the optimal architecture to the search for the optimal kernel function. Using this connection, we show that train-validation split achieves the best excess risk bound among all architectures while requiring few validation samples and provide insights on the role of depth and width.

• **Algorithmic Guarantees via Connection to Low-rank Learning (Section 6):** The results stated so far focus on generalization and are not fully algorithmic in the sense that they assume access to an approximately optimal solution of the upper-level (validation) problem (see (TVO)). As mentioned earlier, this is not the case for the lower-level problem: we specifically consider algorithms such as gradient descent. This naturally raises the question: Can one provably find such an approximate solution with a few validation samples and a computationally tractable algorithm? Towards addressing this question, we connect the shallow neural activation search problem to a novel low-rank matrix learning problem with an overparameterized dimension p. We then provide a two stage algorithm on a train-validation split of the data to find near-optimal hyperparameters via a spectral estimator that also achieves a near-optimal generalization risk. Perhaps unexpectedly, this holds as long as the matrix dimensions obey $h \times p \lesssim (n_{\mathcal{T}} + n_{\mathcal{V}})^2$ which allows for the regime (1.1). In essence, this demonstrates that it is possible to tractably solve the upper problem in the regime of (1.1) even when the problem can easily be overfitting for all choices of hyperparameters. This is similar in spirit to practical NAS problems where the network can fit the data even with poor architectures.

[1]While we do provide guarantees for generic HPO problems (cf. Sec. 3), the emphasis of this work is NAS and the search for the optimal architecture rather than broader class of hyperparameters.

2 Preliminaries and Problem Formulation

We begin by introducing some notation used throughout the paper. We use $\boldsymbol{X}^\dagger$ to denote the Moore–Penrose inverse of a matrix X. $\gtrsim, \lesssim$ denote inequalities that hold up to an absolute constant. We define the norm $\|\cdot\|_{\mathcal{X}}$ over an input space $\mathcal{X}$ as $\|f\|_{\mathcal{X}} := \sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})|$. $\widetilde{\mathcal{O}}(\cdot)$ implies equality up to constant/logarithmic factors. $c, C > 0$ are used to denote absolute constants. Finally, we use $\mathcal{N}_\varepsilon(\boldsymbol{\Delta})$ to denote an ε-Euclidean ball cover of a set $\boldsymbol{\Delta}$.

Throughout, we use $(\boldsymbol{x}, y) \sim \mathcal{D}$ with $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, to denote the data distribution of the feature/label pair. We also use $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_\mathcal{T}}$ to denote the training dataset and $\mathcal{V} = \{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)\}_{i=1}^{n_\mathcal{V}}$ the validation dataset and assume $\mathcal{T}$ and $\mathcal{V}$ are drawn i.i.d. from $\mathcal{D}$. Given a loss function ℓ and a hypothesis $f : \mathcal{X} \to \mathcal{Y}$, we define the population risk and the empirical validation risk as follows

$$\mathcal{L}(f) = \mathbb{E}_\mathcal{D}[\ell(y, f(\boldsymbol{x}))],\ \widehat{\mathcal{L}}_\mathcal{V}(f) = \frac{1}{n_\mathcal{V}} \sum_{i=1}^{n_\mathcal{V}} \ell(\widetilde{y}_i, f(\widetilde{\boldsymbol{x}}_i)). \tag{2.1}$$

For binary classification with $y \in \{-1, 1\}$ also define the test classification error as $\mathcal{L}^{0-1}(f) = \mathbb{P}(yf(\boldsymbol{x}) \leq 0)$. We focus on a *bilevel empirical risk minimization* (ERM) problem over train/validation datasets involving a hyperparameter $\boldsymbol{\alpha} \in \mathbb{R}^h$ and a hypothesis f. Here, the model f (which depends on the hyperparameter $\boldsymbol{\alpha}$) is typically trained over the training data $\mathcal{T}$ with the hyperparameters fixed (lower problem). Then, the best hyperparameter is selected based on the validation data (upper-level problem).

While the training of the lower problem is typically via optimizing an (possibly regularized) empirical risk of the form $\widehat{\mathcal{L}}_\mathcal{T}(f) = \frac{1}{n_\mathcal{T}} \sum_{i=1}^{n_\mathcal{T}} \ell(y_i, f(\boldsymbol{x}_i))$, we do not explicitly require a global optima of this empirical loss and assume that we have access to an algorithm $\mathcal{A}$ that returns a model based on the training data $\mathcal{T}$ with hyperparameters fixed at $\boldsymbol{\alpha}$

$$f_{\boldsymbol{\alpha}}^\mathcal{T} = \mathcal{A}(\boldsymbol{\alpha}, \mathcal{T}).$$

We provide some example scenarios with the corresponding algorithm below.

Scenario 1: Strongly Convex Problems. The lower-level problem ERM is strongly convex with respect to the parameters of the model and $\mathcal{A}$ returns its unique solution. A specific example is learning the optimal kernel given a predefined set of kernels per §4.1.

Scenario 2: Gradient Descent & NAS. In NAS, f is typically a neural network and $\boldsymbol{\alpha}$ encodes the network architecture. Given this architecture, starting from randomly initialized weights, $\mathcal{A}$ trains the weights of f on dataset $\mathcal{T}$ by running fixed number of gradient descent iterations. See §4.2 and §5 for more details.

As mentioned earlier, modern NAS problems typically obey (1.1) where the lower-level problem involves fitting an overparameterized network with many parameters whereas the number of architectural parameters h is typically less than 1000 and obeys $h = \dim(\boldsymbol{\alpha}) \leq n_\mathcal{V}$. Intuitively, this is the regime in which all lower-level problems have solutions perfectly fitting the data. However, the under-parameterized upper problem can potentially guide the algorithm towards the right model. Our goal is to provide theoretical insights for this regime. To select the optimal model, given hyperparameter space $\boldsymbol{\Delta}$ and tolerance $\delta > 0$, the following Train-Validation Optimization (TVO) returns a δ-approximate solution to the validation risk $\widehat{\mathcal{L}}_\mathcal{V}$ (upper problem)

$$\widehat{\boldsymbol{\alpha}} \in \{\boldsymbol{\alpha} \in \boldsymbol{\Delta} \mid \widehat{\mathcal{L}}_\mathcal{V}(f_{\boldsymbol{\alpha}}^\mathcal{T}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \widehat{\mathcal{L}}_\mathcal{V}(f_{\boldsymbol{\alpha}}^\mathcal{T}) + \delta\}. \tag{TVO}$$

3 Generalization with Train-Validation Split

In this section we state our generic generalization bounds for bilevel optimization problems with train-validation split. Next, in Sections 4 and 5, we utilize these generic bounds to establish guarantees for

neural architecture/activation search –which will necessitate additional technical innovations. We start by introducing the problem setting in Section 2. We then introduce our first result in Section 3.1 which controls the generalization gap between the test and validation risk as well as the corresponding gradients. Then, in Section 3.2, we relate training and validation risks which, when combined with our first result, yields an end-to-end generalization bound for the train-validation split.

3.1 Low validation risk implies good generalization

Our first result connects the test (generalization) error to that of the validation error. A key aspect of our result is that we establish uniform convergence guarantees that hold over continuous hyperparameter spaces which is particularly insightful for differentiable HPO/NAS algorithms such as DARTS [44]. Besides validation loss, we will also establish the uniform convergence of the hyper-gradient $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}})$ of the upper problem under similar assumptions. Such concentration of hyper-gradient is insightful for gradient-based bilevel optimization algorithms to solve (TVO). Specifically, we will answer how many validation samples are required so that upper-level problems (hyper-)gradient concentrates around its expectation. Our results rely on the following definition and assumptions.

Definition 1 (Effective dimension) *For a set $\boldsymbol{\Delta} \in \mathbb{R}^h$ of hyperparameters we define its effective dimension h_{eff} as the smallest value of $h_{\mathit{eff}} > 0$ such that $|\mathcal{N}_\varepsilon(\boldsymbol{\Delta})| \leq (\bar{C}/\varepsilon)^{h_{\mathit{eff}}}$ for all $\varepsilon > 0$ and a constant $\bar{C} > 0$.*

The effective dimension captures the degrees of freedom of a set $\boldsymbol{\Delta}$. In particular, if $\boldsymbol{\Delta} \in \mathbb{R}^h$ has Euclidean radius R, then $h_{\text{eff}} = h$ with $\bar{C} = 3R$ so that it reduces to the number of hyperparameters. However, h_{eff} is more nuanced and can also help incorporate problem structure/prior knowledge (e.g. sparse neural architectures have less degrees of freedom).[2]

Assumption 1 *$\mathcal{A}(\cdot)$ is an L-Lipschitz function of $\boldsymbol{\alpha}$ in $\|\cdot\|_{\mathcal{X}}$ norm, that is, for all pairs $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \boldsymbol{\Delta}$, we have $\|f_{\boldsymbol{\alpha}_1}^{\mathcal{T}} - f_{\boldsymbol{\alpha}_2}^{\mathcal{T}}\|_{\mathcal{X}} \leq L\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2}$.*

Assumption 2 *For all hypotheses $f_{\boldsymbol{\alpha}}^{\mathcal{T}}$, the loss $\ell(y, \cdot)$ is Γ-Lipschitz over the feasible set $\{f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$. Additionally, $\ell(y, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x})) - \mathbb{E}[\ell(y, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x}))]$ has bounded subexponential ($\|\cdot\|_{\psi_1}$) norm with respect to the randomness in $(\boldsymbol{x}, y) \sim \mathcal{D}$.*

Assumption 1 (and a less stringent version stated in Assumption 6) is key to our NAS generalization analysis and we show it holds in a variety of scenarios. Assumption 2 requires the loss or gradient on a sample $(\boldsymbol{x}, y)$ to have a sub-exponential tail. While the above two assumptions allow us to show that the validation error is indicative of the test error, the two additional assumptions (which parallel those above) allow us to show that the hypergradient is concentrated around gradient of the true loss with respect to the hyperparameters. As mentioned earlier such concentration of the hyper-gradient is insightful for gradient-based bilevel optimization algorithms.

Assumption 1′ *For some $R \geq 1$ and all $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \boldsymbol{\Delta}$ and $\boldsymbol{x} \in \mathcal{X}$, hyper-gradient obeys $\|\nabla_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}_1}^{\mathcal{T}}(\boldsymbol{x})\|_{\ell_2} \leq R$ and $\|\nabla_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}_1}^{\mathcal{T}}(\boldsymbol{x}) - \nabla_{\boldsymbol{\alpha}} f_{\boldsymbol{\alpha}_2}^{\mathcal{T}}(\boldsymbol{x})\|_{\ell_2} \leq RL\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2}$.*

Assumption 2′ *$\ell'(y, \cdot)$ is Γ-Lipschitz and the hyper-gradient noise $\nabla\ell(y, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x})) - \mathbb{E}[\nabla\ell(y, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x}))]$ over the random example $(\boldsymbol{x}, y) \sim \mathcal{D}$ has bounded subexponential norm as well.*

Our first result establishes a generalization guarantee for (TVO) under these assumptions.

[2] In the empirical process theory literature this is sometimes also referred to as the uniform entropy number e.g. see [52, Definition 2.5]

Theorem 1 *Suppose Assumptions 1&2 hold. Let $\widehat{\boldsymbol{\alpha}}$ be an approximate minimizer of the empirical validation risk per (TVO) and set $\bar{h}_{\mathit{eff}} := h_{\mathit{eff}} \log(\bar{C} L \Gamma n_{\mathcal{V}} / h_{\mathit{eff}})$. Also assume $n_{\mathcal{V}} \geq \bar{h}_{\mathit{eff}} + \tau$ for some $\tau > 0$. Then, with probability at least $1 - 2e^{-\tau}$,*

$$\sup_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} |\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) - \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}})| \leq \sqrt{\frac{C(\bar{h}_{\mathit{eff}} + \tau)}{n_{\mathcal{V}}}}, \tag{3.1}$$

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \leq \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) + 2\sqrt{\frac{C(\bar{h}_{\mathit{eff}} + \tau)}{n_{\mathcal{V}}}} + \delta. \tag{3.2}$$

Suppose also Assumptions 1′& 2′ hold and $n_{\mathcal{V}} \geq h + \bar{h}_{\mathit{eff}} + \tau$ for some $\tau > 0$. Then, with probability at least $1 - 2e^{-\tau}$, the hyper-gradient of the validation risk converges uniformly. That is,

$$\sup_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \|\nabla \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) - \nabla \mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}})\|_{\ell_2} \leq \sqrt{\frac{C(h + \bar{h}_{\mathit{eff}} + \tau)}{n_{\mathcal{V}}}}. \tag{3.3}$$

This result shows that as soon as the size of the validation data exceeds the effective number of hyperparameters $n_{\mathcal{V}} \gtrsim h_{\text{eff}}$ (up to log factors) then (1) as evident per (3.1) the test error is close to the validation error (i.e. validation error is indicative of the test error) and (2) per (3.2) the optimization over validation is guaranteed to return a hypothesis on par with the best choice of hyperparameters in $\boldsymbol{\Delta}$. Theorem 1 has two key distinguishing features, over the prior art on cross-validation [37, 38], which makes it highly relevant for modern learning problems. The first distinguishing contribution of this result is that it applies to continuous hyperparameters and bounds the size of $\boldsymbol{\Delta}$ via the refined notion of effective dimension, establishing a logarithmic dependence on problem parameters. This is particularly important for the Lipschtizness parameter L which can be rather large in practice. The second distinguishing factor is that besides the loss function, per (3.3) we also establish the uniform convergence of hyper-gradients. The reason the latter is useful is that if the validation loss satisfies favorable properties (e.g. Polyak-Lojasiewicz condition), one can obtain generalization guarantees based on the stationary points of validation risk via (3.3) (see [26, 64]). We defer a detailed study of such gradient-based bilevel optimization guarantees to future work. Finally, we note that (3.3) requires at least h samples - which is the ambient dimension and greater than h_{eff}. This is unavoidable due to the vectorial nature of the (hyper)-gradient and is consistent with related results on uniform gradient concentration [49].

Theorem 1 is the simplest statement of our results and there are a variety of possible extensions that are more general and/or require less restrictive assumptions (see Theorem 6 in Appendix B.1 for further detail). First, the loss function or gradient can be viewed as special cases of functionals of the loss function and as long as such functionals are Lipschitz with subexponential behavior, they will concentrate uniformly. Second, while Theorem 1 aims to highlight our ability to handle continuous hyperparameters via the Lipschitzness of the algorithm $\mathcal{A}$, Assumption 1 can be replaced with a much weaker (see Assumption 6 in the Appendix). In general, $\mathcal{A}$ can be discontinuous as long as it is approximately locally-Lipschitz over the set $\boldsymbol{\Delta}$. This would allow for discrete $\boldsymbol{\Delta}$ (requiring $n_{\mathcal{V}} \propto \log |\boldsymbol{\Delta}|$ samples). Additionally, when analyzing neural nets, we indeed prove approximate Lipschitzness (rather than exact Lipschitzness).

Finally, we note that the results above do not directly imply good generalization as they do not guarantee that the validation error ($\min_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}})$) or the generalization error ($\min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}})$) of the model trained with the best hyperparameters is small. This is to be expected as when there are very few training data one can not hope for the model $f_{\boldsymbol{\alpha}}^{\mathcal{T}}$ to have good generalization even with optimal hyperparameters. However, whether the training phase is successful or not, the validation phase returns approximately the best hyperparameters even with a bad model! In the next section we do in fact show that with enough training data the

validation/generalization of the model trained with the best hyperparameter is indeed small allowing us to establish an end-to-end generalization bound.

3.2 End-to-end generalization with Train-Validation Split

We begin by briefly discussing the role of the training data which is necessary for establishing an end-to-end bound. To accomplish this, we need to characterize how the population loss of the algorithm $\mathcal{A}$ scales with the training data $n_{\mathcal{T}}$. To this aim, let us consider the limiting case $n_{\mathcal{T}} \to +\infty$ and define the corresponding model for a given set of hyperparameters $\boldsymbol{\alpha}$ as

$$f_{\boldsymbol{\alpha}}^{\mathcal{D}} := \mathcal{A}(\boldsymbol{\alpha}, \mathcal{D}) := \lim_{n_{\mathcal{T}} \to \infty} \mathcal{A}(\boldsymbol{\alpha}, \mathcal{T}).$$

Classical learning theory results typically bound the difference between the population loss/risk of a model that is trained with finite training data ($\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}})$) and the loss achieved by the idealized infinite data model ($\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{D}})$) in terms of an appropriate complexity measure of the class and the size of the training data. In particular, for a specific choice of the hyperparameter $\boldsymbol{\alpha}$, based on classical learning theory [13]) a typical behavior is to have

$$\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) \leq \mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{D}}) + \frac{\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}} + C_0\sqrt{t}}{\sqrt{n_{\mathcal{T}}}}, \tag{3.4}$$

with probability at least $1 - e^{-t}$. Here, $\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}}$ is a dataset-dependent complexity measure for the hypothesis set of the lower-level problem and C_0 is a positive scalar. We are now ready to state our end-to-end bound which ensures a bound of the form (3.4) holds simultaneously for all choices of hyperparameters $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$.

Proposition 1 (Train-validation bound) *Consider the setting of Theorem 1 and for any fixed $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ assume (3.4) holds. Also assume $f_{\boldsymbol{\alpha}}^{\mathcal{D}}$ (in $\|\cdot\|_{\mathcal{X}}$ norm) and $\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}}$ have bounded Lipschitz constants with respect to $\boldsymbol{\alpha}$ over $\boldsymbol{\Delta}$. Then with probability at least $1 - 3e^{-t}$ over the train $\mathcal{T}$ and validation $\mathcal{V}$ datasets*

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \leq \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \left(\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{D}}) + \frac{\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}}}{\sqrt{n_{\mathcal{T}}}} \right) + \sqrt{\frac{\widetilde{\mathcal{O}}(h_{\mathit{eff}} + t)}{n_{\mathcal{V}}}} + \delta.$$

In a nutshell, the above bound shows that the generalization error of a model trained with train-validation split is on par with the best train-only generalization achievable by picking the best hyperparameter $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$. The only loss incurred is an extra $\sqrt{h_{\text{eff}}/n_{\mathcal{V}}}$ term which is vanishingly small as soon as the validation data is sufficiently larger than the effective dimension of the hyperparameters. We note that the Lipschitzness condition on $f_{\boldsymbol{\alpha}}^{\mathcal{D}}$ and $\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}}$ can be relaxed. For instance, Proposition 2, stated in Appendix B.3, provides a strict generalization where the Lipschitz property is only required to hold over a subset of the search space $\boldsymbol{\Delta}$.

We note that classical literature on this topic [36, 38] typically use model selection to select the complexity from a nested set of growing hypothesis spaces by using explicit regularizers of a form similar in spirit to $\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}}$. Instead, Proposition 1 aims to implicitly control the model capacity via the complexity measure $\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}}$ of the outcome of the lower-level algorithm $\mathcal{A}$. This implicit capacity control is what we will utilize in Theorem 3 (via norm-based generalization) which is of interest in practical NAS settings[3]. This is because capacity of modern deep nets are rarely controlled explicitly and in fact, larger capacity often benefits generalization ability. In the next section, we demonstrate how one can utilize this end-to-end guarantee within specific problems.

[3]Indeed, to obtain meaningful bounds in the regime (1.1), $\mathcal{C}_{\boldsymbol{\alpha}}^{\mathcal{T}}$ should not be dimension-dependent (as # of params $p \gtrsim n_{\mathcal{T}}$).

4 Feature Maps and Shallow Networks

In this section and Section 5, we provide our main results on neural architecture/activation search which will utilize the generalization bounds provided above. Towards understanding the NAS problem, we first introduce the feature map selection problem [39]. This problem is similar in spirit to the multiple kernel learning [29] problem which aims to select the best kernel for a learning task. This problem can be viewed as a simplified linear NAS problem where the hyperparameters control a linear combination of features and the parameters of the network are shared across all hyperparameters. Building on our findings on feature maps/kernels, Section 4.2 will establish our main results on activation search for shallow networks.

4.1 Feature map selection for kernel learning

Below, the hyperparameter vector $\boldsymbol{\alpha} \in \mathbb{R}^{h+1}$ controls both the choice of the feature map and the ridge regularization coefficient.

Definition 2 (Optimal Feature Map Regression) *Suppose we are given h feature maps $\phi_i : \mathcal{X} \to \mathbb{R}^p$. Define the superposition $\phi_{\boldsymbol{\alpha}}(\cdot) = \sum_{i=1}^{h} \boldsymbol{\alpha}_i \phi_i(\cdot)$. Given training data $\mathcal{T}$, the algorithm $\mathcal{A}$ solves the ridge regression with feature matrix $\boldsymbol{\Phi}_{\boldsymbol{\alpha}} := \boldsymbol{\Phi}_{\boldsymbol{\alpha}}^{\mathcal{T}}$ via*

$$\boldsymbol{\theta}_{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\Phi}_{\boldsymbol{\alpha}}\boldsymbol{\theta}\|_{\ell_2}^2 + \boldsymbol{\alpha}_{h+1}\|\boldsymbol{\theta}\|_{\ell_2}^2 \tag{4.1}$$

$$\text{where} \quad \boldsymbol{\Phi}_{\boldsymbol{\alpha}} = [\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_1)\ \phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_2)\ \ldots\ \phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_{n_{\mathcal{T}}})]^T. \tag{4.2}$$

Here $\boldsymbol{\alpha}_{h+1} \in [\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}^+ \cup \{0\}$ controls the regularization strength. We then solve for optimal choice $\widehat{\boldsymbol{\alpha}}$ via (TVO) with hypothesis $f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{v}) = \boldsymbol{v}^T\boldsymbol{\theta}_{\boldsymbol{\alpha}}$.

The main motivation behind studying problems of the form (4.1), is obtaining the best linear superposition of the feature maps minimizing the validation risk. This is in contrast to building h individual models and then applying ensemble learning, which would correspond to a linear superposition of the h kernels induced by these feature maps. Instead this problem formulation models weight-sharing which has been a key ingredient of state-of-the-art NAS algorithms [63, 41] as same parameter $\boldsymbol{\theta}$ has compatible dimension with all feature maps. In essence, for NAS, this parameter will correspond to the (super)network's weights and the feature maps will be induced by different architecture choices so that the formulation above can be viewed as the simplest of NAS problems with linear networks. Nevertheless, as we will see in the forthcoming sections this analysis serves as a stepping stone for more complex NAS problems. To apply Theorem 1 to the optimal feature map regression problem we need to verify its assumptions/characterize $\bar{h}_{\text{eff}}$.

Lemma 1 *Suppose the feature maps and labels are bounded i.e. $\sup_{\boldsymbol{x}\in\mathcal{X}, 1\le i\le h} \|\phi_i(\boldsymbol{x})\|_{\ell_2}^2 \le B$ and $|y| \le 1$. Also assume the loss ℓ is bounded and 1-Lipschitz w.r.t. the model output. Set $\lambda_0 = \lambda_{\min} + \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \sigma_{\min}^2(\boldsymbol{\Phi}_{\boldsymbol{\alpha}}) > 0$. Additionally let $\boldsymbol{\Delta}$ be a convex set with ℓ_1 radius $R \ge 1$. Then, Theorem 1 holds with $\bar{h}_{\text{eff}} = (h+1)\log(20R^3Bn_{\mathcal{T}}^2\lambda_0^{-2}(Bn_{\mathcal{T}}+1))$.*

An important component of the proof of this lemma is that we show that when $\lambda_0 > 0$, $f_{\boldsymbol{\alpha}}$ is a Lipschitz function of $\boldsymbol{\alpha}$ and Theorem 1 applies. Thus per (TVO) in this setting one can provably and jointly find the optimal feature map and the optimal regularization strength as soon as the size of the validation exceeds the number of hyperparameters.

We note that there are two different mechanisms by which we establish Lipschitzness w.r.t. $\boldsymbol{\alpha}$ in the above theorem. When $\lambda_{\min} > 0$, the lower problem is strongly-convex with respect to the model parameters. As we show in the next lemma, this is more broadly true for any training procedure which is based on minimizing a loss which is strongly convex with respect to the model parameters.

Lemma 2 *Let $\boldsymbol{\Delta}$ be a convex set. Suppose $f_{\boldsymbol{\alpha}}$ is parameterized by $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ where $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ is obtained by minimizing a loss function $\bar{\mathcal{L}}_{\mathcal{T}}(\boldsymbol{\alpha}, \boldsymbol{\theta}) : \boldsymbol{\Delta} \times \mathbb{R}^p \to \mathbb{R}$. Suppose $\bar{\mathcal{L}}_{\mathcal{T}}(\boldsymbol{\alpha}, \boldsymbol{\theta})$ is μ strongly-convex in $\boldsymbol{\theta}$ and $\bar{L}$ smooth in $\boldsymbol{\alpha}$. Then $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ is $\sqrt{\bar{L}/\mu}$-Lipschitz in $\boldsymbol{\alpha}$.*

Importantly, Lemma 1 can also operate in the ridgeless regime ($\lambda_{\min} = 0$) even when the training loss is not strongly convex. This holds as long as the feature maps are not poorly-conditioned in the sense that

$$\inf_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \sigma_{\min} \left(\boldsymbol{\Phi}_{\boldsymbol{\alpha}} \boldsymbol{\Phi}_{\boldsymbol{\alpha}}^T \right) = \lambda_0 > 0. \tag{4.3}$$

We note the exact value of λ_0 is not too important as the effective number of hyperparameters only depends logarithmically on this quantity. Such a ridgeless regression setting has attracted particular interest in recent years as deep nets can often generalize well in an overparameterized regime without any regularization despite perfectly interpolating the training data. In the remainder of the manuscript, we focus on ridgeless regression problems with an emphasis on neural nets (thus we drop the $h+1$'th index of $\boldsymbol{\alpha}$).

Our next result utilizes Proposition 1 to provide an end-to-end generalization bound for feature map selection involving both training and validation sample sizes. Below we assume that (4.3) holds with high probability over $\mathcal{T}$.

Theorem 2 (End-to-end generalization for feature map selection) *Consider the setup in Definition 2 with $\boldsymbol{\alpha}_{h+1} = 0$. Set $R = \sup_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \|\boldsymbol{\alpha}\|_{\ell_1}$ and assume $\sup_{\boldsymbol{x} \in \mathcal{X}, 1 \le i \le h} \|\phi_i(\boldsymbol{x})\|_{\ell_2}^2 \le B$ and ℓ in (2.1) is Γ-Lipschitz and bounded by a constant. Suppose (4.3) holds with probability at least $1 - p_0$. Also $p \ge n_{\mathcal{T}} \ge n_{\mathcal{V}} \gtrsim h \log(M)$ with $M = 30R^4 B^2 \lambda_0^{-2} \Gamma (n_{\mathcal{T}}^2 + n_{\mathcal{V}}^2) \|\boldsymbol{y}\|_{\ell_2}$. Furthermore, let $\boldsymbol{y} = [y_1 \; y_2 \; \ldots \; y_{n_{\mathcal{T}}}]$. Then with probability at least $1 - 4e^{-t} - p_0$, the population risk (over $\mathcal{D}$) obeys*

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}) \le \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} 2\Gamma \sqrt{\frac{B \boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}}{n_{\mathcal{T}}}} + C \sqrt{\frac{h \log(M) + \tau}{n_{\mathcal{V}}}} + \delta.$$

In this result, the excess risk term $\sqrt{\frac{\boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}}{n_{\mathcal{T}}}}$ becomes smaller as the kernel induced by $\boldsymbol{\alpha}$ becomes better aligned with the labeling function e.g. when $\boldsymbol{y}$ lies on the principal eigenspace of $\boldsymbol{K}_{\boldsymbol{\alpha}}$. This theorem shows that for the optimal feature map regression problem, bilevel optimization via a train-validation split returns a generalization guarantee on par with that of the best feature map (minimizing the excess risk) as soon as the size of the validation data exceeds the number of hyperparameters.

4.2 Activation search for shallow networks

In this section we focus on an *activation search* problem where the goal is to find the best activation among a parameterized family of activations for training a shallow neural networks based on a train-validation split. To this aim we consider a one-hidden layer network of the form $\boldsymbol{x} \mapsto f_{\text{nn}}(\boldsymbol{x}) = \boldsymbol{v}^T \sigma(\boldsymbol{W}\boldsymbol{x})$ and focus on a binary classification task with $y \in \{-1, +1\}$ labels. Here, $\sigma : \mathbb{R} \to \mathbb{R}$ denotes the activation, $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ input-to-hidden weights, and $\boldsymbol{v} \in \mathbb{R}^d$ hidden-to-output weights. We focus on the case where the activation belongs to a family of activations of the form $\sigma_{\boldsymbol{\alpha}} = \sum_{i=1}^h \boldsymbol{\alpha}_i \sigma_i$ with $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ denoting the hyperparameters. Here, $\{\sigma_i\}_{i=1}^h$ are a list of candidate activation functions (e.g., ReLU, sigmoid, Swish). The neural net with hyperparameter $\boldsymbol{\alpha}$ is thus given by $f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}) = \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$. For simplicity of exposition in this section we will only use the input layer for training thus the training weights are $\boldsymbol{W}$ with dimension $p = \dim(\boldsymbol{W}) = k \times d$ and fix $\boldsymbol{v}$ to have $\pm\sqrt{c_0/k}$ entries (roughly half of each) with a proper choice of $c_0 > 0$. In Section 5 we further discuss how our results can be extended to NAS beyond activation search and to deeper networks where all the layers are trained.

Bilevel optimization for shallow activation search: We now explain the specific gradient-based algorithm we consider for the lower-level optimization problem. For a fixed hyperparameter $\boldsymbol{\alpha}$, the lower-level optimization aims to minimize a quadratic loss over the training data of the form

$$\widehat{\mathcal{L}}_{\mathcal{T}}(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{n_{\mathcal{T}}}(y_i - f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}_i, \boldsymbol{W}))^2.$$

To this aim, for a fixed hyperparameter $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, starting from a random initialization of the form $\boldsymbol{W}_0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ we run gradient descent updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\widehat{\mathcal{L}}_{\mathcal{T}}(\boldsymbol{W}_\tau)$ for T iterations. Thus, the lower algorithm $\mathcal{A}$ returns the model

$$f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x}) = \boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_T\boldsymbol{x}).$$

We then solve for the δ-approximate optimal activation $\widehat{\boldsymbol{\alpha}}$ via (TVO) by setting ℓ in (2.1) to be the hinge loss.

To state our end-to-end generalization guarantee, we need a few definitions. First, we introduce neural feature maps induced by the Neural Tangent Kernel (NTK) [34].

Definition 3 (Neural feature maps & NTK) *Let $f_{nn,\boldsymbol{\alpha}}(\cdot,\boldsymbol{\theta})$ be a neural net parameterized by weights $\boldsymbol{\theta} \in \mathbb{R}^p$ and architecture $\boldsymbol{\alpha}$. Define $\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \frac{\partial f_{nn,\boldsymbol{\alpha}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}_0}$ to be the* neural feature map *at the random initialization $\boldsymbol{\theta}_0 \sim \mathcal{D}_{init}$. Define the neural feature matrix $\boldsymbol{\Phi}_{\boldsymbol{\alpha}} = [\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_1)\ \ldots\ \phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_{n_{\mathcal{T}}})]^T \in \mathbb{R}^{n_{\mathcal{T}}\times p}$ as in (4.2) i.e.*

$$\boldsymbol{\Phi}_{\boldsymbol{\alpha}} = \left[\frac{\partial f_{nn,\boldsymbol{\alpha}}(\boldsymbol{x}_1)}{\partial \boldsymbol{\theta}_0}\ \ldots\ \frac{\partial f_{nn,\boldsymbol{\alpha}}(\boldsymbol{x}_n)}{\partial \boldsymbol{\theta}_0}\right]^T. \tag{4.4}$$

We define the gram matrix as $\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} = \boldsymbol{\Phi}_{\boldsymbol{\alpha}}\boldsymbol{\Phi}_{\boldsymbol{\alpha}}^T \in \mathbb{R}^{n_{\mathcal{T}}\times n_{\mathcal{T}}}$ with (i,j)th entry equal to $\langle\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_i), \phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_j)\rangle$ and NTK matrix is as $\boldsymbol{K}_{\boldsymbol{\alpha}} = \mathbb{E}_{\boldsymbol{\theta}_0}[\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}]$.

Neural feature maps are in general nonlinear function of $\boldsymbol{\alpha}$ (cf. Sec. 5). However, in case of shallow networks, it is nicely additive and obeys $\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}_i) = \sum_{i=1}^h \boldsymbol{\alpha}_i\phi_1(\boldsymbol{x}_i)$ regardless of random initialization $\boldsymbol{\theta}_0$ establishing a link to Def. 2. The next assumption ensures the expressivity of the NTK to interpolate the data and enables us to analyze regularization-free training.

Assumption 3 (Expressive Neural Kernels) *There exists $\lambda_0 > 0$ such that for any $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, the NTK matrix $\boldsymbol{K}_{\boldsymbol{\alpha}} \succeq \lambda_0\boldsymbol{I}_{n_{\mathcal{T}}}$.*

This assumption is similar to (4.3) but we take expectation over random $\boldsymbol{\theta}_0$. Assumptions in a similar spirit to this are commonly used for the optimization/generalization analysis of neural nets, especially in the interpolating regime [4, 19, 16, 50]. For fixed $\boldsymbol{\alpha}$, $\boldsymbol{K}_{\boldsymbol{\alpha}} \succ 0$ as long as no two training inputs are perfectly correlated and $\phi_{\boldsymbol{\alpha}}$ is analytic and not a polynomial [22]. The key aspect of our assumption is that we require the NTK matrices to be lower bounded for all $\boldsymbol{\alpha}$. Later in Theorem 4 of §5 we shall show how to circumvent this assumption with a small ridge regularization.

With these definitions in place we are now ready to state our end-to-end generalization guarantee for Shallow activation search where the lower-level problem is optimized via gradient descent. The reader is referred to Theorem 14 for the precise statement. Note that λ_0 and $\boldsymbol{K}_{\boldsymbol{\alpha}}$ scales linearly with initialization variance c_0. To state a result invariant to initialization, we will state our result in terms of the normalized eigen lower bound $\bar{\lambda}_0 = \lambda_0/c_0$ and kernel matrix $\bar{\boldsymbol{K}}_{\boldsymbol{\alpha}} = \boldsymbol{K}_{\boldsymbol{\alpha}}/c_0$.

Theorem 3 (Neural activation search) *Suppose input features have unit Euclidean norm i.e.* $\|\boldsymbol{x}\|_{\ell_2} = 1$ *and labels take values in* $\{-1, 1\}$. *Pick* $\boldsymbol{\Delta}$ *to be a subset of the unit* ℓ_1 *ball. Suppose Assumption 3 holds for* $\boldsymbol{\theta}_0 \leftrightarrow \boldsymbol{W}_0$ *and the candidate activations have first two derivatives* $(|\sigma_i'|, |\sigma_i''|)$ *upper bounded by* $B > 0$. *Furthermore, fix* $\boldsymbol{v}$ *with half* $\sqrt{c_0/k}$ *and half* $-\sqrt{c_0/k}$ *entries for a proper* c_0 *(see supplementary). Define the normalized lower bound* $\bar{\lambda}_0 = \lambda_0/c_0$ *and kernel matrix* $\bar{\boldsymbol{K}}_{\boldsymbol{\alpha}} = \boldsymbol{K}_{\boldsymbol{\alpha}}/c_0$. *Also assume the network width obeys*

$$k \gtrsim \mathit{poly}(n_{\mathcal{T}}, \bar{\lambda}_0^{-1}, \varepsilon^{-1}).$$

for a tolerance level $1 > \varepsilon > 0$ *and the size of the validation data obeys* $n_{\mathcal{V}} \gtrsim \widetilde{\mathcal{O}}(h)$. *Following the aforementioned bilevel optimization scheme with a proper* $\eta > 0$ *choice and any choice of number of iterations obeying* $T \gtrsim \widetilde{\mathcal{O}}(\frac{n_{\mathcal{T}}}{\bar{\lambda}_0} \log(\varepsilon^{-1}))$, *the classification error (0-1 loss) on the data distribution* $\mathcal{D}$ *obeys*

$$\mathcal{L}^{0-1}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \leq \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} 2B \sqrt{\frac{\boldsymbol{y}^T \bar{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}}{n_{\mathcal{T}}}} + C \sqrt{\frac{\widetilde{\mathcal{O}}(h) + t}{n_{\mathcal{V}}}} + \varepsilon + \delta,$$

with probability at least $1 - 4(e^{-t} + n_{\mathcal{T}}^{-3} + e^{-10h})$ *(over the randomness in* $\boldsymbol{W}_0, \mathcal{T}, \mathcal{V}$*). Here,* $\boldsymbol{y} = [y_1\ y_2\ \ldots\ y_{n_{\mathcal{T}}}]$. *On the same event, for all* $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, *the training classification error obeys* $\hat{\mathcal{L}}_{\mathcal{T}}^{0-1}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) \leq \varepsilon$.

For a fixed $\boldsymbol{\alpha}$, the norm-based excess risk term $\sqrt{\frac{\boldsymbol{y}^T \bar{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}}{n_{\mathcal{T}}}}$ quantifies the alignment between the kernel and the labeling function (which is small when $\boldsymbol{y}$ lies on the principal eigenspace of $\boldsymbol{K}_{\boldsymbol{\alpha}}$). This generalization bound is akin to expressions that arise in norm-based NTK generalization arguments such as [4]. Critically, however, going beyond a fixed $\boldsymbol{\alpha}$, our theorem establishes this for all activations uniformly to conclude that the minimizer of the validation error also achieves minimal excess risk. The final statement of the theorem shows that the training error is arbitrarily small (essentially zero as $T \to \infty$) over all activations uniformly. Together, these results formally establish the pictorial illustration in Figures 1(a) & (b).

The proof strategy has two novelties with respect to standard NTK arguments. First, it requires a subtle uniform convergence argument on top of the NTK analysis to show that certain favorable properties that are essential to the NTK proof hold *uniformly* for all activations (i.e. choices of the hyperparameters) simultaneously with the same random initialization $\boldsymbol{W}_0$. Second, since neural nets may not obey Assumption 1, to be able to apply our generalization bounds we need to construct a uniform Lipschitz approximation via its corresponding linearized feature map ($f_{\text{lin},\boldsymbol{\alpha}}(\boldsymbol{x}) = \boldsymbol{x}^T \phi_{\boldsymbol{\alpha}}(\boldsymbol{x})$) and bound the neural net's risk over train-validation procedure in terms of this proxy. This uniform approximation is in contrast to pointwise approximation results of [5]. Sec. 5 provides further insights on our strategy. Finally, we mention a few possible extensions. One can (i) use logistic loss in the algorithm $\mathcal{A}$ rather than least-squares loss –which is often more suitable for classification tasks– [81] or (ii) use regularization techniques such as early stopping or ridge penalization [42] or (iii) refine the excess risk term using further technical arguments such as local Rademacher complexity [10].

5 Extension to NAS for Deep Architectures

In this section, we provide further discussion extending our results in Sec. 4.2 to multi-layer networks and general NAS beyond simple activation search. Our intention is to provide the core building blocks for an NTK-based NAS argument over a continuous architecture space $\boldsymbol{\Delta}$. Recall the neural feature map introduced in Definition 3 where $f_{\text{nn},\boldsymbol{\alpha}}(\cdot, \boldsymbol{\theta})$ can be any architecture induced by hyperparameters $\boldsymbol{\alpha}$. For instance, in DARTS, the architecture is a directed acyclic graph where each node $\boldsymbol{x}^{(i)}$ is a latent representation of the raw input $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ dictates the operations $\sigma^{(i,j)}$ on the edges that transform $\boldsymbol{x}^{(i)}$ to obtain $\boldsymbol{x}^{(j)}$.

Recall the matrices $\boldsymbol{K}_{\boldsymbol{\alpha}}, \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} \in \mathbb{R}^{n_{\mathcal{T}} \times n_{\mathcal{T}}}$ from Definition 3 which are constructed from the neural feature map $\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) := \frac{\partial f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}_0}$. For infinite-width networks (with proper initialization), NTK perfectly governs the training and generalization dynamics and the architectural hyperparameters $\boldsymbol{\alpha}$ controls the NTK kernel matrix $\boldsymbol{K}_{\boldsymbol{\alpha}} \in \mathbb{R}^{n_{\mathcal{T}} \times n_{\mathcal{T}}}$ (associated to the training set $\mathcal{T}$). A critical challenge for the general architectures is that the relation between the NTK kernel $\boldsymbol{K}_{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}$ can be highly nonlinear. Here, we introduce a generic result for NTK-based generalization where we assume that $\boldsymbol{K}_{\boldsymbol{\alpha}}$ is possibly nonlinear but Lipschitz function of $\boldsymbol{\alpha}$. Recall that Theorem 1 achieves logarithmic dependency on the Lipschitz constant thus unless the Lipschitz constant is extremely large, good generalization bounds are achievable. Our arguments will utilize this fact.

Assumption 4 (Lipschitz Kernel) *$\boldsymbol{K}_{\boldsymbol{\alpha}}, \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} \in \mathbb{R}^{n \times n}$ are L-Lipschitz functions of $\boldsymbol{\alpha}$ in spectral norm.*

To be able to establish generalization bounds for learning with generic neural feature maps in connection to NTK (Thm 4 below) we need to ensure that wide networks converge to their infinite-width counterparts uniformly over $\boldsymbol{\Delta}$. Let $k_\star$ be a width parameter associated with the network. For instance, for a fully-connected network, $k_\star$ can be set to the minimum number of hidden units across all layers. Similar to random features, it is known that, the network at random initialization converges to its infinite width counterpart exponentially fast. We now formalize this assumption.

Assumption 5 (Neural Feature Concentration) *Recall Def. 3. There exists a width parameter $k_\star > 0$ and scalar $\nu > 0$ such that, for any fixed $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, at initialization $\boldsymbol{\theta}_0 \sim \mathcal{D}_{init}$, we have*

$$\mathbb{P}\Big\{\|\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} - \boldsymbol{K}_{\boldsymbol{\alpha}}\| \geq \sqrt{\nu t / k_\star}\Big\} \geq \mathrm{e}^{-t}. \tag{5.1}$$

For deep ResNets, fully-connected deep nets and DARTS architecture space (with zero, skip, conv operations), this assumption holds with proper choice of $\nu > 0$ (cf. Theorem E.1 of [22] and Lemma 22 of [78] which sets $\nu \propto n^2$).

Assumptions 4 and 5, allows us to establish uniform convergence to the NTK over a continuous architecture space $\boldsymbol{\Delta}$. Specifically, given tolerance $\varepsilon > 0$, for $k_\star \gtrsim \widetilde{\mathcal{O}}(\varepsilon^{-2} \nu h_{\text{eff}} \log(L))$, with high probability (over initialization), $\|\boldsymbol{K}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}\| \leq \varepsilon$ holds for all $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ uniformly (cf. Lemma 13).

Our generalization result for generic architectures is provided below and establishes a bound similar to Theorem 3 by training a linearized model with neural feature maps. However, unlike Theorem 3, here we employ a small ridge regularization to promote Lipschitzness which helps us circumvent Assumption 3.

Theorem 4 *Suppose Assumptions 4 and 5 hold and for all $\boldsymbol{\alpha}$ and some $B > 0$ neural feature maps obey $\|\frac{\partial f_{nn,\boldsymbol{\alpha}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}_0}\|_{\ell_2}^2 \leq B$ almost surely. We solve feature map regression (Def. 2) with neural feature maps and fixed ridge penalty λ. Fix some eigen-cutoff $\lambda_0 > 0$ and tolerance $\varepsilon > 0$. Set $0 < \lambda \leq \frac{\varepsilon \lambda_0^2}{4\sqrt{B n_{\mathcal{T}}}}$ and $\bar{h}_{eff} = \widetilde{\mathcal{O}}(h_{eff} \log(\lambda^{-2} L n_{\mathcal{T}}^3))$. Finally define the λ_0-positive set*

$$\boldsymbol{\Delta}_0 = \{\boldsymbol{\alpha} \in \boldsymbol{\Delta} \;\big|\; \boldsymbol{K}_{\boldsymbol{\alpha}} \geq \lambda_0\}.$$

Also assume $k_\star \gtrsim \widetilde{\mathcal{O}}(\varepsilon^{-4} \lambda_0^{-4} \nu^2 h_{eff} \log(L))$ and $n_{\mathcal{V}} \gtrsim \bar{h}_{eff}$. Finally, set ℓ in (2.1) to be the hinge loss. Then, with probability at least $1 - 5e^{-t}$, for some constant $C > 0$, the binary classification error obeys

$$\mathcal{L}^{0-1}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \leq \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}_0} 2\sqrt{\frac{B \boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}}{n_{\mathcal{T}}}} + C\sqrt{\frac{\bar{h}_{eff} + t}{n_{\mathcal{V}}}} + \varepsilon + \delta.$$

Here the set $\mathbf{\Delta}_0$ is the set of architectures that are favorable in the sense that the NTK associated with their infinite-width network can rather quickly fit the data as its kernel matrix has a large minimum eigenvalue. This also ensures that the neural nets trained on $\mathcal{T}$ are Lipschitz with respect to $\boldsymbol{\alpha}$ over $\mathbf{\Delta}_0$. In essence, the result states that the bilevel procedure (TVO) is guaranteed to return a model at least as good as the best model over $\mathbf{\Delta}_0$. Importantly, one can enlarge the set $\mathbf{\Delta}_0$ by reducing the penalty λ at the cost of a larger $\bar{h}_{\text{eff}}$ term which grows logarithmically in $1/\lambda$. The above theorem also relates to optimal kernel selection using validation data which connects NAS to the multiple kernel learning (MKL) [29] problem. However in the MKL literature, the kernel $\boldsymbol{K}_{\boldsymbol{\alpha}}$ is typically a linear function of $\boldsymbol{\alpha}$ whereas in NAS the dependence is more involved and nonconvex, especially for realistic search spaces.

Deep activation search. As a concrete example, we briefly pause to demonstrate that the above assumption holds for a deep multilayer activation search problem. In this case we will show that L is at most exponential in depth D. To be precise, following Section 4.2, fix a pre-defined set of activations $\{\sigma_j\}_{j=1}^h$. For a feedforward depth D network, set $\boldsymbol{\alpha} \in \mathbb{R}^{Dh}$ to be the concatenation of D subvectors $\{\boldsymbol{\alpha}^{(i)}\}_{i=1}^D \subset \mathbb{R}^h$. The layer ℓ activation is then given via

$$\sigma_{\boldsymbol{\alpha}^{(i)}}(\cdot) = \sum_{j=1}^{h} \boldsymbol{\alpha}_j^{(i)} \sigma_j(\cdot).$$

Now, given input $\boldsymbol{x} = \boldsymbol{x}^{(0)}$ and weight matrices $\boldsymbol{\theta} = \{\boldsymbol{W}^{(i)}\}_{i=1}^{D+1}$, define the corresponding feedforward network $f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R} = \boldsymbol{W}^{(D+1)} \boldsymbol{x}^{(D)}$ where the hidden features $\boldsymbol{x}^{(i)}$ are defined as

$$\boldsymbol{x}^{(i)} = \sigma_{\boldsymbol{\alpha}^{(i)}}(\boldsymbol{W}^{(i)} \boldsymbol{x}^{(i-1)}) \quad \text{for} \quad 1 \le i \le D.$$

The lemma below shows that in this case the Lipschitzness L with respect to the hyperparameters is at most exponential in D. The result is stated for a fairly flexible random Gaussian initialization. The precise statement is deferred to Lemma 15.

Lemma 3 *Suppose for all $1 \le i \le h$, $|\phi_i(0)|, |\phi_i'(x)|, |\phi_i''(x)|$ are bounded. Input features are normalized to ensure $\|\boldsymbol{x}\|_{\ell_2} \lesssim \sqrt{d}$. Let layer i have k_i neurons and $\boldsymbol{W}^{(i)} \in \mathbb{R}^{k_i \times k_{i-1}}$. Suppose the aspect ratios k_i/k_{i-1} are bounded for layers $i \ge 2$. For the first layer, denote $k_0 = d$ and $k_1 = k$. Each layer ℓ is initialized with i.i.d. $\mathcal{N}(0, c_\ell)$ entries satisfying*

$$c_\ell \le \begin{cases} \bar{c} & \text{if} \quad i = 1 \\ \bar{c}/k_{i-1} & \text{if} \quad \ell \ge 2 \end{cases} \qquad \text{for some constant} \quad \bar{c} > 0.$$

Then, Assumption 4 holds (with high probability for $\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}$) with $\log(L) \lesssim D + \log(k + d + n_{\mathcal{T}})$.

We remark that this initialization scheme corresponds to Xavier/He initializations ($1/\text{fan\_in}$ variance) as well as our setting in Theorem 3. We suspect that the exponential dependence on D can be refined by enforcing normalization schemes during initialization to ensure that hidden features don't grow exponentially with depth. Recall that sample complexity in Theorem 1 depends logarithmically on L, which grows at most linearly in D up to log factors. Furthermore, as stated earlier Assumption 5 is known to hold in this setting as well (cf. discussion above). Thus for a depth D network, using the above Lipschitzness bound, Theorem 4 allows for good generalization with a validation sample complexity of $n_{\mathcal{V}} \propto \widetilde{\mathcal{O}}(h_{\text{eff}} \log(L)) = \widetilde{\mathcal{O}}(D \times h_{\text{eff}})$. Finally, note that $\log(L)$ also exhibits a logarithmic dependence on k. Thus as long as network width is not exponentially large (which holds for all practical networks), our excess risk bounds remain small leading to meaningful generalization guarantees. We do remark that results can also be extended to infinite-width networks (which are of interest due to NTK). Here, the key idea is constructing a Lipschitz approximation

to the infinite-width problem via a finite-width problem. In similar spirit to Lemma 5, such approximation can be made arbitrarily accurate with a polynomial choice of k [78, 5]. As discussed earlier, Theorem 7 in the appendix provides a clear path to fully formalize this and the proof of Theorem 3 already employs such a Lipschitz approximation scheme to circumvent imperfect Lipschitzness of nonlinear neural net training.

6 Algorithmic Guarantees via Connection to Low-rank Matrix Learning

The results stated so far focus on generalization and are not fully algorithmic in nature in the sense that they assume access to an approximately optimal hyperparameter of the upper-level problem per (TVO) based on the validation data. In this section we wish to investigate whether it is possible to provably find such an approximate solution with a few validation samples and a computationally tractable algorithm. To this aim, in this section, we establish algorithmic connections between our activation/feature-map search problems of Section 4 to a rank-1 matrix learning problem. In Def. 2 –instead of studying $\boldsymbol{\Phi}_{\boldsymbol{\alpha}}$ given $\boldsymbol{\alpha}$– let us consider the matrix of feature maps

$$\boldsymbol{X} = [\phi_1(\boldsymbol{x})\ \phi_2(\boldsymbol{x})\ \ldots\ \phi_h(\boldsymbol{x})]^T \in \mathbb{R}^{h\times p}$$

for a given input $\boldsymbol{x}$. Then, the population squared-loss risk of a $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ pair predicting $\boldsymbol{\theta}^T\phi_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is given by

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\theta}) := \mathbb{E}[(y - \boldsymbol{\alpha}^T\boldsymbol{X}\boldsymbol{\theta})^2] = \mathbb{E}[(y - \langle \boldsymbol{X}, \boldsymbol{\alpha}\boldsymbol{\theta}^T\rangle)^2].$$

Thus, the search for the optimal model pair $(\boldsymbol{\alpha}_\star, \boldsymbol{\theta}_\star)$ is simply a rank-1 matrix learning task with $\boldsymbol{M}_\star = \boldsymbol{\alpha}_\star\boldsymbol{\theta}_\star{}^T$. Can we learn the right matrix with a tractable algorithm in the regime (1.1)?

This is a rather subtle question as in the regime (1.1) there is not enough samples to reconstruct $\boldsymbol{M}_\star$ as anything algorithm regardless of computational tractability requires $n_{\mathcal{T}} + n_{\mathcal{V}} \gtrsim p + h$! But this of course does not rule out the possibility of finding an approximately optimal hyperparameter close to $\boldsymbol{\alpha}_\star$. To answer this –rather tricky question– we study a discriminative data model commonly used for modeling low-rank learning. Consider a realizable setup $y = \boldsymbol{\alpha}_\star{}^T\boldsymbol{X}\boldsymbol{\theta}_\star$ where we ignore noise for ease of exposition, see supplementary for details. We also assume that the feature matrix $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Suppose we have $\mathcal{T} = (y_i, \boldsymbol{X}_i)_{i=1}^{n_{\mathcal{T}}=n}, \mathcal{V} = (\widetilde{y}_i, \tilde{\boldsymbol{X}}_i)_{i=1}^{n_{\mathcal{V}}=n}$ datasets with equal sample split $n = n_{\mathcal{T}} = n_{\mathcal{V}}$. If we combine these datasets into $\mathcal{T}$ and solve ERM, when $2n \le p$, for *any choice of* $\boldsymbol{\alpha}$, weights $\boldsymbol{\theta} \in \mathbb{R}^p$ can perfectly fit the labels. Instead, we propose the following two-stage algorithm to achieve a near-optimal learning guarantee. Set $\hat{\boldsymbol{M}} = \sum_{i=1}^n \widetilde{y}_i\tilde{\boldsymbol{X}}_i$.

$$\text{1. } \textbf{Spectral estimator on } \mathcal{V}: \quad \text{Set } \widehat{\boldsymbol{\alpha}} = \text{top\_eigen\_vec}(\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^T). \tag{6.1}$$

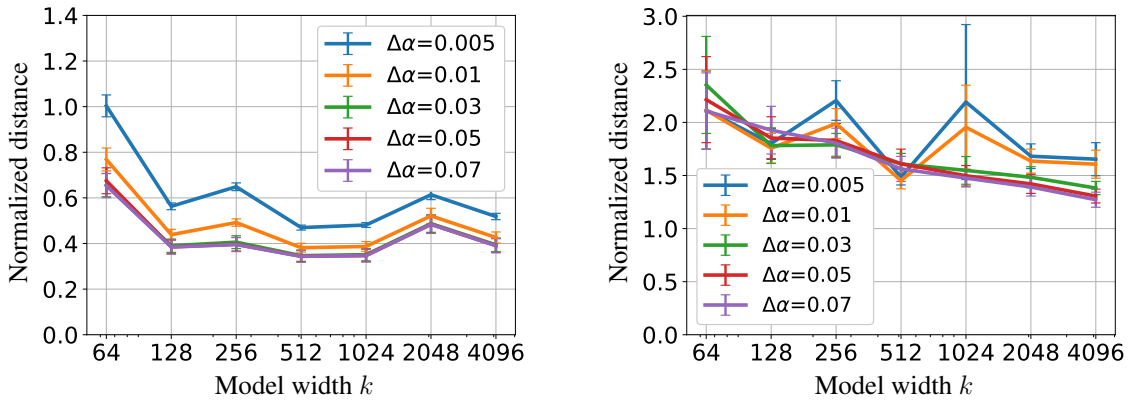
$$\text{2. } \textbf{Solve ERM on } \mathcal{T}: \quad \text{Set } \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y_i - \widehat{\boldsymbol{\alpha}}^T\boldsymbol{X}_i\boldsymbol{\theta})^2. \tag{6.2}$$

We have the following guarantee for this procedure.

Theorem 5 (Low-rank learning with $p > n$) *Let $(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i)_{i=1}^n$ be i.i.d. matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Let $y_i = \boldsymbol{\alpha}_\star{}^T\boldsymbol{X}_i\boldsymbol{\theta}_\star$ for unit norm $\boldsymbol{\alpha} \in \mathbb{R}^h, \boldsymbol{\theta} \in \mathbb{R}^p$. Consider an asymptotic setting where p, n, h grow to infinity with fixed ratios given by $\bar{p} = p/n > 1$, $\bar{h} = h/n < 1$ and consider the asymptotic performance of $(\widehat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$.*

Let $1 \ge \rho_{\boldsymbol{\alpha}_\star,\widehat{\boldsymbol{\alpha}}} \ge 0$ be the absolute correlation between $\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}$ i.e. $\rho_{\boldsymbol{\alpha}_\star,\widehat{\boldsymbol{\alpha}}} = |\boldsymbol{\alpha}^T\widehat{\boldsymbol{\alpha}}|$. Suppose $\bar{p}\bar{h} \le 1/6$. We have that

$$\lim_{n\to\infty} \rho^2_{\boldsymbol{\alpha}_\star,\widehat{\boldsymbol{\alpha}}} \ge 1 - 64\bar{p}\bar{h} \tag{6.3}$$



(a) Input layer stability for a one-hidden layer network

(b) Stability of weights of a deeper four layer network

Figure 2: We visualize the Lipschitzness of the algorithm when $\mathcal{A}(\cdot)$ is stochastic gradient descent. We train networks with activation parameters α and $\alpha + \Delta\alpha$ and display the normalized distances $\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_{\alpha+\Delta\alpha}\|_{\ell_2}/\Delta\alpha$ for different perturbation strengths $\Delta\alpha$.

Additionally, if $\bar{p}\bar{h} \le c$ for sufficiently small constant $c > 0$,

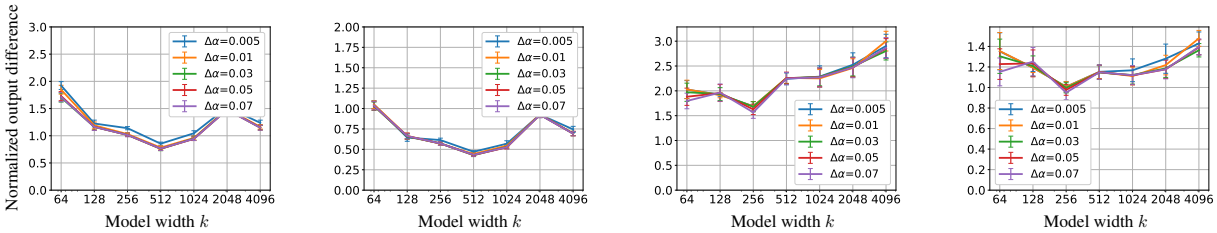
$$\lim_{n\to\infty} \mathcal{L}(\widehat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) \le \underbrace{1 - \frac{1}{\bar{p}}}_{\textit{risk}(\boldsymbol{\alpha}_\star)} + \underbrace{\frac{200\bar{h}}{1 - 1/\bar{p}}}_{\textit{estimating } \boldsymbol{\alpha}_\star} . \tag{6.4}$$

A few remarks are in order. First, the result applies in the regime $p \gg n$ as long as –the rather surprising condition– $ph \lesssim n^2$ holds (see (6.3)). Numerical experiments in Section 7 verify that (specifically Figure 5) this condition is indeed necessary. Here, $\text{risk}(\boldsymbol{\alpha}_\star) = 1 - n/p$ is the *exact asymptotic risk* one would achieve by solving ERM with the knowledge of optimal $\boldsymbol{\alpha}_\star$. Our result shows that one can approximately recover this optimal $\boldsymbol{\alpha}_\star$ up to an error that scales with ph/n^2. Our second result achieves a near-optimal risk via $\widehat{\boldsymbol{\alpha}}$ without knowing $\boldsymbol{\alpha}_\star$. Since $1 - 1/\bar{p}$ is essentially constant, the risk due to $\boldsymbol{\alpha}_\star$-search is proportional to $\bar{h} = h/n$. This rate is consistent with Theorem 1 which would achieve a risk of $1 - n/p + \widetilde{\mathcal{O}}(\sqrt{h/n})$. Remarkably, we obtain a slightly more refined rate ($h/n \le \sqrt{h/n}$) using a spectral estimator with a completely different mathematical machinery based on high-dimensional learning. To the best of our knowledge, our spectral estimation result (6.3) in the $p > n$ regime is first of its kind (with a surprising $ph \lesssim n^2$ condition) and might be of independent interest. Finally, while this result already provides valuable algorithmic insights, it would be desirable to extend this result to general feature distributions to establish algorithmic guarantees for the original activation/feature map search problems.

7 Numerical Experiments

To verify our theory, we provide three sets of experiments. First, to test Theorem 3, we verify the (approximate) Lipschitzness of trained neural nets to perturbations in the activation function. Second, to test Theorem 1, we will study the test-validation gap for DARTS search space. Finally, we verify our claims on

a. Lipschitzness of Trained Networks. First, we wish to verify Assumption 1 for neural nets by demonstrating their Lipschitzness under proper conditions. In these experiments, we consider a single hyperparameter $\alpha \in \mathbb{R}$ to control the activation via a combination of ReLU and Sigmoid i.e. $\sigma_\alpha(x) = (1 - \alpha)\text{ReLU}(x) +$



(a) Maximum output variability for shallow network (b) Average output variability for shallow network (c) Maximum output variability for deep network (d) Average output variability for deep network

Figure 3: We visualize the Lipschitzness of the algorithm when $\mathcal{A}(\cdot)$ is stochastic gradient descent. In Figure 3(a) and 3(b), we train the input layer of 2-layer shallow networks with activation parameters α and $\alpha + \Delta\alpha$ which have the same setup as Figure 2(a). Then we display the maximum and average output variability defined in (7.1) and (7.2) respectively for different perturbation strengths $\Delta\alpha$ and model width k. In Figure 3(c) and 3(d), we train 4-layer deep fully connected networks with activation parameters α and $\alpha + \Delta\alpha$ and also display the maximum and average output variability for different perturbation strengths $\Delta\alpha$.

$\alpha \cdot \text{Sigmoid}(x)$. Training the network weights $\boldsymbol{\theta}$ with this activation from the same random initialization leads to the weights $\boldsymbol{\theta}_\alpha$. We are interested in testing the stability of these weights to slight α perturbations by studying the normalized distance $\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_{\alpha+\Delta\alpha}\|_{\ell_2}/\Delta\alpha$. This in turn ensures the Lipschitzness of the model output via a standard bound (see supp. for numerical experiments). Fig. 2 presents our results on both shallow and deeper networks on a binary MNIST task which uses the first two classes with squared loss. This setup is in a similar spirit to our theory. In Fig. 2(a) we train input layer of a shallow network $f_\alpha(x) = \boldsymbol{v}^T\sigma_\alpha(\boldsymbol{W}\boldsymbol{X})$ where $\boldsymbol{W} \in \mathbb{R}^{k\times 784}$. In Fig. 2(b), a deeper fully connected network with 4 layers is trained. Here, the number of neurons from input to output are k, $k/2$, $k/4$ and 1 and the same activation $\sigma_\alpha(\boldsymbol{X})$ is used for all layers. Finally, we initialize the network with He initialization and train the model for 60 epochs with batch size 128 with SGD optimizer and learning rate 0.003. For each curve and width level, we average 20 experiments where we first pick 20 random $\alpha \in [0, 1]$ and their perturbation $\alpha + \Delta\alpha$. We then compute the average of normalized distances $\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_{\alpha+\Delta\alpha}\|_{\ell_2}/\Delta\alpha$.

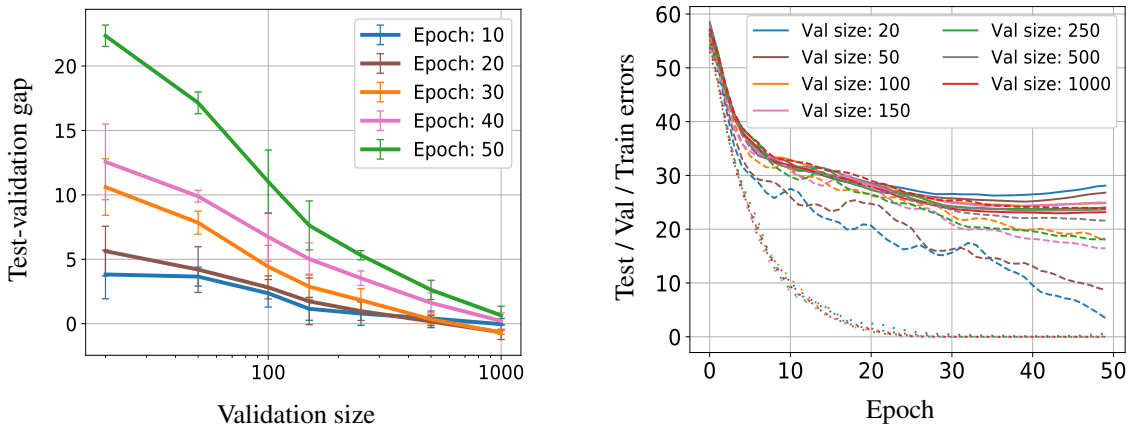
We now provide further experiments to better verify Assumption 1. Let $\mathcal{T}_{\text{test}}$ be the test data (of MNIST) which provides a proxy for the input domain $\mathcal{X}$ [4]. Our goal is to assess the Lipschitzness of the network prediction over $\mathcal{T}_{\text{test}}$ which exactly corresponds to the setup of Assumption 1. Specifically, we will evaluate two quantities as a function of the activation perturbation $\Delta\alpha$

$$\text{Maximum output variability: } \max(f, \Delta\alpha, \mathcal{T}_{\text{test}}) = \frac{\sup_{\boldsymbol{x}\in\mathcal{T}_{\text{test}}} |f(\boldsymbol{\theta}_\alpha, \boldsymbol{x}) - f(\boldsymbol{\theta}_{\alpha+\Delta\alpha}, \boldsymbol{x})|}{\Delta\alpha}, \tag{7.1}$$

$$\text{Average output variability: } \text{avg}(f, \Delta\alpha, \mathcal{T}_{\text{test}}) = \frac{1}{n_{\mathcal{T}}\Delta\alpha} \sum_{\boldsymbol{x}\in\mathcal{T}_{\text{test}}} |f(\boldsymbol{\theta}_\alpha, \boldsymbol{x}) - f(\boldsymbol{\theta}_{\alpha+\Delta\alpha}, \boldsymbol{x})|. \tag{7.2}$$

Here, the maximum output variability is the most relevant quantity as it directly corresponds to the Lipschitzness of the function over the input domain. We keep the same setup in Figure 3(a) and 3(b) as Figure 2(a), however, in Figure 3(a) and 3(b), instead of computing the distance between trained models, we plot maximum output variability (7.1) and average output variability (7.2) respectively for different perturbation level Δ_α and model width. Figure 3(c) and 3(d) demonstrate the output variability on 4-layer neural networks which has the same setup as 2(b).

[4] We note that these values are calculated over the test data however we found the behavior over the training data to be similar.



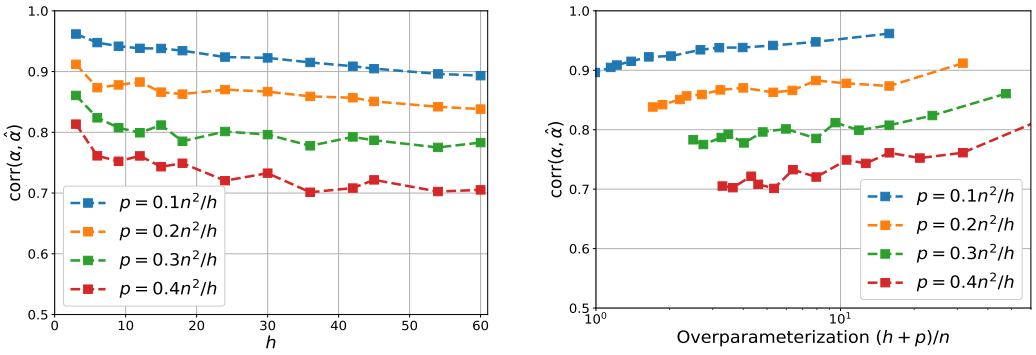
(a) The test-validation gap for the continuously parameterized architecture during the search phase of DARTS.

(b) The train (dotted curves), validation (dashed curves) and test (solid curves) errors during the search phase of DARTS training.

Figure 4: The test-validation gap and errors for the continuously parameterized architecture during the search phase of DARTS. The result is average of 5 different runs. Evaluations are for different validation sample sizes and epoch checkpoints. Figure 4(b) shows that all models overfit to training samples quickly. The DARTS runs with less validation examples also trend to overfit the validation dataset which leads to the increase in test error (most visibly for $n_{\mathcal{V}} \in \{20, 50\}$).

All figures support our theory and show that, the normalized distance is indeed stable to the perturbation level $\Delta\alpha$ across different widths and only mildly changes. Note that $\Delta\alpha \in \{0.01, 0.005\}$ result in a slightly larger normalized distance compared to larger perturbations. Such behavior for small $\Delta\alpha$ is not surprising and is likely due to the imperfect Lipschitzness of the network (especially with ReLU activation). Fortunately, our theory allows for this as it only requires an approximate Lipschitz property (recall the discussion below Theorem 1).

b. Test-Validation Gap for DARTS. In this experiment, we study a realistic architecture search space via DARTS algorithm [44] over CIFAR-10 dataset using 10k training samples. We only consider the search phase of DARTS and train for 50 epochs using SGD. This phase outputs a *continuously parameterized architecture*, which can be computed on DARTS' supernet. Each operation on the edges of the final architecture is a linear superposition of eight predefined operations (e.g. conv3x3, zero, skip). The curves are obtained by averaging three independent runs. In Fig. 4, we assess the gap between the test and validation errors while varying validation sizes from 20 to 1000. Our experiments reveal two key findings via Figure 4(a). First, the train-validation gap indeed decreases rapidly as soon as the validation size is only mildly large, e.g. around $n_{\mathcal{V}} = 250$ –much smaller than the typical validation size used in practice. This is consistent with Theorem 1 as the architecture has 224 hyperparameters. On the other hand, there is indeed a potential of overfitting to validation for $n_{\mathcal{V}} \leq 100$. We also observe that the gap noticeably increases with more epochs. The small gaps at initial epochs may be due to insufficient training i.e. network does not yet achieve zero training loss. For later epochs, since early-stopping (i.e. using earlier epoch checkpoints) has a ridge regularization effect, we suspect that widening gap may be due to the growing Lipschitz constant with respect to the architecture choice. Such behavior would be consistent with Thm 4 (smaller ridge penalty leads to more excess validation risk). Figure 4(b) displays the train/validation/test errors by epoch for different validation sample sizes. This figure is also consistent with our core setup and expectations (1.1). The training



(a) Correlation as a function of the left singular dimension h (corresponding to # of hyperparameters)

(b) Correlation as a function of overparameterization level (left-dim+right-dim / sample size)

Figure 5: Overparameterized rank-1 learning setup of Sec. 6. The (absolute value of) the correlation coefficient between $\boldsymbol{\alpha}_\star$ and the estimate $\widehat{\boldsymbol{\alpha}}$ as a function of h. Each curve (with distinct color) corresponds to a fixed ph/n^2 choice. Here we kept the notation consistent with Sec. 6 and set $n = n_{\mathcal{T}} = n_{\mathcal{V}}$.

loss/error quickly goes down to zero. Validation contains much fewer samples but it is difficult to overfit (despite continuously parameterized architecture). However, as discussed above, below a certain threshold ($n_{\mathcal{V}} \leq 100$), differentiable search indeed overfits to the validation leading to deteriorating test risk.

c. Overparameterized Rank-1 Learning. We now aim to verify the theoretical claims of Sec. 6 on rank-1 learning. Specifically, we will verify our claim (6.3) and empirically demonstrate that recovery of the ground-truth hyperparameter $\boldsymbol{\alpha}_\star$ requires $hp \lesssim n^2$ where we set $n = n_{\mathcal{T}} = n_{\mathcal{V}}$. This is in contrast to the arguably more intuitive $h \lesssim n$ requirement. Figure 5 summarizes our numerical results. Our experiment is constructed as follows. We generate an $h \times p$ rank-1 matrix $\boldsymbol{M} = \boldsymbol{\alpha}_\star \boldsymbol{\theta}_\star{}^T$ with left and right singular vectors $\boldsymbol{\alpha}_\star, \boldsymbol{\theta}_\star$ generated as i.i.d. Gaussians normalized to unit norm. We collect n noiseless labels via $y_i = \boldsymbol{\alpha}_\star{}^T \boldsymbol{X}_i \boldsymbol{\theta}_\star$ where $\boldsymbol{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ as in Theorem 5 and apply the spectral estimator (6.1) to estimate the $\boldsymbol{\alpha}_\star$ vector. In our experiments, we vary h between 0 to 60 and we set $p = \gamma n^2/h$. γ is also varied from 0.1 to 0.4. Figure 5(a) displays the (absolute value of) the correlation coefficient between $\boldsymbol{\alpha}_\star$ and the estimate $\widehat{\boldsymbol{\alpha}}$ as a function of h. Each curve (with distinct color) corresponds to a fixed ph/n^2 choice. Observe that these curves remain constant even if h is varying more than a factor of 20. This indicates that correlation indeed depends on hp rather than solely h. When we increase γ, p increases and correlation noticeably decreases as we move from a higher curve to a lower curve again indicating the dependence on p. Finally, Figure 5(b) displays the exact same information but the x-axis is the total number of parameters normalized by the total data (used for spectral initialization). This shows that, just as predicted by Theorem 5, hyperparameter $\boldsymbol{\alpha}_\star$ can be learned even when p is much larger than n as long as $hp \lesssim n^2$.

8 Related Works

Our work establishes generalization guarantees for architecture search and is closely connected to the literature on deep learning theory, statistical learning, and hyperparameter optimization / NAS. These connections are discussed below.

Statistical learning: The statistical learning theory provide rich tools for analyzing test performance of algorithms [13, 66]. Our discussion on learning with bilevel optimization and train-validation split connects to the model selection literature [36, 37, 69] as well as the more recent works on architecture search [40, 39].

The model selection literature is mostly concerned with controlling the model complexity (e.g. via nested hypothesis), which is not directly applicable to high-capacity deep nets. The latter two works are closer to us and also establish connections between feature maps and NAS. However, there are key distinctions. First, we operate on continuous hyperparameter spaces whereas these consider discrete hyperparameters which are easier to analyze. Second, their approaches do not directly apply to neural nets as they have to control the space of all networks with zero training loss which can be large. In contrast, we analyze tractable lower-level algorithms such as gradient-descent and study the properties of a single model returned by the algorithm. [30] discuss methods for determining train-validation split ratios. Favorable learning theoretic properties of (cross-)validation are studied by [38, 73]. These works either apply to specific scenarios such as tuning lasso penalty or do not consider hyperparameters. We also note that algorithmic stability of [14] utilizes stability of an algorithm to changes in the training set. In contrast, we consider the stability with respect to hyperparamters. [7] discusses the importance of train-validation split in meta-learning problems, which also accept a bilevel formulation. Finally, [71] explores tuning the learning rate for improved generalization. They focus on a simple quadratic objective using hyper-gradient methods and characterize when train-validation split provably helps.

Generalization in deep learning: The statistical study of neural networks can be traced back to 1990's [3, 12, 9]. With the success of deep learning, the generalization properties of deep networks received a renewed interest in recent years [24, 6, 55, 28]. [11, 56] establish spectrally normalized risk bounds for deep networks and [54] provides refined bounds by exploiting inter-layer Jacobian. [6] proposes tighter bounds using compression techniques. These provide a solid foundation on the generalization ability of deep nets. More recently, [34] has introduced the neural tangent kernel which enables the analysis of deep nets trained with gradient-based algorithms: With proper initialization, wide networks behave similarly to kernel regression. NTK has received significant attention for analyzing the optimization and learning dynamics of wide networks [23, 2, 18, 76, 57, 80, 70, 60]. Closer to us, [16, 4, 48, 59, 1, 4] provide generalization bounds for gradient descent training. A line of research implements neural kernels for convolutional networks and ResNets [75, 65, 43, 33]. Related to us [5] mention the possibility of using NTK for NAS and recent work by [62] shows that such an approach can indeed produce good results and speed up NAS. In connection to these, §4 and 5 establish the first provable guarantees for NAS and also provide a rigorous justification of the NTK-based NAS by establishing data-dependent bounds under verifiable assumptions.

Neural Architecture Search and Bilevel Optimization: HPO and NAS find widespread applications in machine learning. These are often formulated as a bilevel optimization problem, which seeks the optimal hyperparameter at the upper-level optimization minimizing a validation loss. There are variety of NAS approaches employing reinforcement learning, evolutionary search, and Bayesian optimization [79, 8, 77]. Recently differentiable optimization methods have emerged as a powerful tool for NAS (and HPO) problem [44, 72, 15, 74, 41, 63, 21] which use continuous relaxations of the architecture. Specifically, DARTS proposed by [44] uses a continuous relaxation on the upper-level optimization and weight sharing in the lower level. The algorithm optimizes the lower and upper level simultaneously by gradient descent using train-validation split. [72, 74, 63, 21] provide further improvements on differentiable NAS. The success of differentiable HPO/NAS methods has further raised interest in large continuous hyperparameter spaces and accelerating bilevel optimization [27, 41, 45]. In this paper we initiate some theoretical understanding of this exciting and expanding empirical literature.

High-dimensional learning: In §6, we use ideas from high-dimensional learning to establish algorithmic results. Closest to us are the works on spectral estimators. The recent literature utilizes spectral methods for low-rank learning problems such as phase-retrieval and clustering [68]. Spectral algorithm is used to initialize the algorithm within a basin of attraction for a subsequent method such as convex optimiza-

tion or gradient descent. This is in similar flavor to our Theorem 5 which employs a two-stage algorithm. [46, 47, 53] provide asymptotic/sharp analysis for spectral methods for phase retrieval. However, unlike our problem, these works all focus symmetric matrices and operate in the low-dimensional regime where sample size is more than the parameter size. While not directly related, we remark that sparse phase retrieval and sparse PCA problems [35, 82] do lead to a high-dimensional regime (sample size less than parameter size) due to the sparsity prior on the parameter.

9 Conclusion and Future Directions

In this paper, we explored theoretical aspects of the NAS problem. We first provided statistical guarantees when solving bilevel optimization with train-validation split. We showed that even if the lower-level problem overfits –which is common in deep learning– the upper-level problem can guarantee generalization with a few validation data. We applied these results to establish guarantees for the optimal activation search problem and extended our theory to generic neural architectures. These formally established the high-level intuition in Figure 1. We also showed interesting connections between the activation search and a novel low-rank matrix learning problem and provided sharp algorithmic guarantees for the latter.

There are multiple exciting directions for future research. First, one can develop variants of Theorems 3 and 4 by studying other lower-level algorithms (e.g. different loss functions, incorporate regularization) and, most importantly, developing a better understanding of the architecture search spaces and architectural operations. Second, our results are established for the NTK (i.e. lazy training) regime and it would be desirable to obtain similar results for other learning regimes such as the mean-field regime [51]. Finally, it would be interesting to study both computational and statistical aspects of the gradient-based solutions to the upper-level problem (TVO). To this aim, our Theorem 1 established that gradient of the validation loss (i.e. hyper-gradient) uniformly converges under mild assumptions. However, besides this, we require a deeper understanding of the population landscape (i.e. as $n_{\mathcal{V}} \to \infty$) of the the upper-level validation problem (even for the shallow NAS problem of Section 4) which might necessitate new advances in bilevel optimization theory.

Acknowledgements

S.O. and M.L. are supported by NSF-CNS under award #1932254 and by an NSF-CAREER under award #2046816. M.S. is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award # FA9550-18-1-0078, DARPA Learning with Less Labels (LwLL) and FastNICS programs, and NSF-CIF awards #1813877 and #2008443.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

[3] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

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

[5] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

[6] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[7] Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason D Lee, Sham Kakade, Huan Wang, and Caiming Xiong. How important is the train-validation split in meta-learning? *arXiv preprint arXiv:2010.05843*, 2020.

[8] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167*, 2016.

[9] Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[10] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[11] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6241–6250, 2017.

[12] Peter L Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc-dimension bounds for piecewise polynomial networks. *Neural computation*, 10(8):2159–2173, 1998.

[13] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[14] Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. *Advances in Neural Information Processing Systems*, pages 196–202, 2001.

[15] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

[16] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019.

[17] Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. *AAAI*, 2021.

[18] Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[20] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[21] Xuanyi Dong, Mingxing Tan, Adams Wei Yu, Daiyi Peng, Bogdan Gabrys, and Quoc V Le. Autohas: Efficient hyperparameter and architecture search. *arXiv preprint arXiv:2006.03656*, 2020.

[22] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[23] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[24] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[25] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[26] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8759–8770, 2018.

[27] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

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

[29] Mehmet Gonen and Ethem Alpaydin. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[30] Isabelle Guyon et al. A scaling law for the validation-set training-set size ratio. *AT&T Bell Laboratories*, 1(11), 1997.

[31] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[32] Reinhard Heckel and Mahdi Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds a smooth approximation. In *International Conference on Machine Learning*, pages 4149–4158. PMLR, 2020.

[33] Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feedforward networks?—a neural tangent kernel perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

[34] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[35] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Sparse phase retrieval: Convex algorithms and limitations. In *2013 IEEE International Symposium on Information Theory*, pages 1022–1026. IEEE, 2013.

[36] Michael Kearns. A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split. *Advances in Neural Information Processing Systems*, pages 183–189, 1996.

[37] Michael Kearns, Yishay Mansour, Andrew Y Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1):7–50, 1997.

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

[39] Mikhail Khodak, Liam Li, Maria-Florina Balcan, and Ameet Talwalkar. On weight-sharing and bilevel optimization in architecture search. *preprint available at https://openreview.net/forum?id=HJgRCyHFDr*, 2019.

[40] Mikhail Khodak, Liam Li, Nicholas Roberts, Maria-Florina Balcan, and Ameet Talwalkar. A simple setting for understanding neural architecture search with weight-sharing. *ICML AutoML Workshop*, 2020.

[41] Liam Li, Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Geometry-aware gradient algorithms for neural architecture search. *arXiv preprint arXiv:2004.07802*, 2020.

[42] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020.

[43] Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

[44] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[45] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.

[46] Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020.

[47] Wangyu Luo, Wael Alghamdi, and Yue M Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9):2347–2356, 2019.

[48] Chao Ma, Lei Wu, et al. A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics. *arXiv preprint arXiv:1904.04326*, 2019.

[49] Song Mei, Yu Bai, Andrea Montanari, et al. The landscape of empirical risk for nonconvex losses. *Annals of Statistics*, 46(6A):2747–2774, 2018.

[50] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[51] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[52] Shahar Mendelson. A few notes on statistical learning theory. In *Advanced lectures on machine learning*, pages 1–40. Springer, 2003.

[53] Marco Mondelli, Christos Thrampoulidis, and Ramji Venkataramanan. Optimal combination of linear and spectral estimators for generalized linear models. *arXiv preprint arXiv:2008.03326*, 2020.

[54] Prabhat Nagarajan, Garrett Warnell, and Peter Stone. Deterministic implementations for reproducibility in deep reinforcement learning. *arXiv preprint arXiv:1809.05676*, 2018.

[55] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.

[56] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[57] Atsushi Nitanda and Taiji Suzuki. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

[58] Samet Oymak. Learning compact neural networks with regularization. In *International Conference on Machine Learning*, pages 3966–3975. PMLR, 2018.

[59] Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

[60] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

[61] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[62] Daniel S Park, Jaehoon Lee, Daiyi Peng, Yuan Cao, and Jascha Sohl-Dickstein. Towards nngp-guided neural architecture search. *arXiv preprint arXiv:2011.06006*, 2020.

[63] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pages 4095–4104. PMLR, 2018.

[64] Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *arXiv preprint arXiv:2002.08538*, 2020.

[65] Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Jonathan Ragan-Kelley, Ludwig Schmidt, and Benjamin Recht. Neural kernels without tangents. In *International Conference on Machine Learning*, pages 8614–8623. PMLR, 2020.

[66] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[67] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[68] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[69] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

[70] Haoxiang Wang, Ruoyu Sun, and Bo Li. Global convergence and induced kernels of gradient-based meta-learning with neural nets. *arXiv preprint arXiv:2006.14606*, 2020.

[71] Xiang Wang, Shuai Yuan, Chenwei Wu, and Rong Ge. Guarantees for tuning the step size using a learning-to-learn approach. *arXiv preprint arXiv:2006.16495*, 2020.

[72] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. Snas: stochastic neural architecture search. *arXiv preprint arXiv:1812.09926*, 2018.

[73] Ning Xu, Timothy CG Fisher, and Jian Hong. Rademacher upper bounds for cross-validation errors with an application to the lasso. *arXiv preprint arXiv:2007.15598*, 2020.

[74] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong. Pc-darts: Partial channel connections for memory-efficient architecture search. *arXiv preprint arXiv:1907.05737*, 2019.

[75] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[76] Huishuai Zhang, Da Yu, Wei Chen, and Tie-Yan Liu. Training over-parameterized deep resnet is almost as easy as training a two-layer network. *arXiv preprint arXiv:1903.07120*, 2019.

[77] Zhao Zhong, Zichen Yang, Boyang Deng, Junjie Yan, Wei Wu, Jing Shao, and Cheng-Lin Liu. Blockqnn: Efficient block-wise neural network architecture generation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[78] Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Theory-inspired path-regularized differential network architecture search. *arXiv preprint arXiv:2006.16537*, 2020.

[79] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[80] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

[81] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

[82] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Organization of the Appendix

1. We gather some useful statistical learning and concentration results in Section A.
2. Proofs of our generic train-validation bounds and feature maps are provided in Section B.
3. Proofs for general architectures appear in Sections D and E.
4. Proofs for algorithmic guarantees (i.e. overparameterized low-rank learning) are the subject of Section F and G.
5. Finally Theorem 3 (activation search for shallow networks) is proven in Sections H and J.

A Useful Statistical Learning and Concentration Results

The lemma below is a standard generalization result for linear models. Let $(\varepsilon_i)_{i=1}^{n_{\mathcal{T}}}$ be Rademacher random variables. Define Rademacher complexity of a function class $\mathcal{F}$ as

$$\mathcal{R}_{n_{\mathcal{T}}}(\mathcal{F}) = \frac{1}{n_{\mathcal{T}}} \mathbb{E}_{\mathcal{T},\varepsilon_i} \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n_{\mathcal{T}}} \varepsilon_i f(\boldsymbol{x}_i) \right]. \tag{A.1}$$

Lemma 4 *Suppose the loss function ℓ is bounded in $[0, C]$ and Γ-Lipschitz in the second argument. Also define*

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \widehat{\mathcal{L}}_{\mathcal{T}}(f) \quad \textit{where} \quad \widehat{\mathcal{L}}_{\mathcal{T}}(f) = \frac{1}{n_{\mathcal{T}}} \sum_{i=1}^{n_{\mathcal{T}}} \ell(y_i, f(\boldsymbol{x}_i)).$$

Then with probability at least $1 - \mathrm{e}^{-t}$, for all $f \in \mathcal{F}$ we have

$$\sup_{f\in\mathcal{F}} \mathcal{L}_{\mathcal{D}}(f) \le \widehat{\mathcal{L}}_{\mathcal{T}}(f) + 2\Gamma\mathcal{R}_{n_{\mathcal{T}}}(\mathcal{F}) + C\sqrt{\frac{t}{n_{\mathcal{T}}}}$$

Corollary 1 (Linear models) *Let $\mathcal{F}_R^{lin} = \{f \mid f(\boldsymbol{x}) = \boldsymbol{\theta}^T\phi(\boldsymbol{x}),\ \|\boldsymbol{\theta}\|_{\ell_2} \le R\}$. Suppose $\|\phi(\boldsymbol{x})\|_{\ell_2}^2 \le B$ for all $\boldsymbol{x} \in \mathcal{X}$. Then, with probability at least $1 - e^{-t}$, for all $f \in \mathcal{F}_R^{lin}$, we have*

$$\mathcal{L}_{\mathcal{D}}(f) \le \widehat{\mathcal{L}}_{\mathcal{T}}(f) + \frac{2\Gamma R\sqrt{B} + C\sqrt{t}}{\sqrt{n_{\mathcal{T}}}}.$$

Define $\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1)\ \ldots\ \phi(\boldsymbol{x}_{n_{\mathcal{T}}})]^T$. Set $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$. Suppose $n > p = dim(\boldsymbol{\theta})$ and $\sigma_{\min}(\boldsymbol{\Phi}) > 0$. Consider the min Euclidean norm estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{\Phi}^\dagger\boldsymbol{y}$ and $\hat{f}(\boldsymbol{x}) = \hat{\boldsymbol{\theta}}^T\phi(\boldsymbol{x})$. Noting that $\|\hat{\boldsymbol{\theta}}\|_{\ell_2} = \|\boldsymbol{\Phi}^\dagger\boldsymbol{y}\|_{\ell_2} = \sqrt{\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y}}$, we arrive at

$$\mathcal{L}_{\mathcal{D}}(\hat{f}) \le \frac{2\Gamma\sqrt{B\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y}} + C\sqrt{t}}{\sqrt{n_{\mathcal{T}}}}. \tag{A.2}$$

Proof The Rademacher complexity of $\mathcal{F}_R^{\text{lin}}$ is bounded as

$$\begin{aligned}\mathcal{R}_{n_\mathcal{T}}(\mathcal{F}_R^{\text{lin}}) &= \frac{1}{n_\mathcal{T}}\,\mathbb{E}_{\mathcal{T},\varepsilon_i}\left[\sup_{\|\boldsymbol{\theta}\|_{\ell_2}\le R}\sum_{i=1}^{n_\mathcal{T}}\varepsilon_i\boldsymbol{\theta}^T\phi(\boldsymbol{x}_i)\right]\\ &= \frac{R}{n_\mathcal{T}}\,\mathbb{E}\left[\left\|\sum_{i=1}^{n_\mathcal{T}}\varepsilon_i\phi(\boldsymbol{x}_i)\right\|_{\ell_2}\right] \le \frac{R\sqrt{B}}{\sqrt{n_\mathcal{T}}}.\end{aligned}$$

To finish the proof observe that $\hat{f}\subset\mathcal{F}_R^{\text{lin}}$ with $R=\sqrt{\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y}}$. ■

Lemma 5 (Moment concentration with Gaussian tail) *Let X be nonnegative satisfy the tail bound $\mathbb{P}(X\ge E+t)\le e^{-t^2/2}$ for some $E\ge 0$. Then*

$$\mathbb{E}[|X|^k]\le 2^{k+1}\max(E,k+2)^k,$$

which implies

$$\mathbb{E}[|X|^k]^{1/k}\le 2\max(E,k+2).$$

Proof If $E\le k+2$, we will simply use the bound $\mathbb{P}(X\ge k+2+t)\le e^{-t^2/2}$. The tail condition implies

$$\mathbb{P}(X\ge 2Et)\le\begin{cases}1\ if\ t\le 1\\ e^{-Et^2/2}\ \text{ if }\ t\ge 1\end{cases},$$

which implies that

$$\mathbb{P}(X^k\ge 2^k tE^k)\le\begin{cases}1\ if\ t\le 1\\ e^{-Et^{2/k}/2}\ \text{ if }\ t\ge 1\end{cases},$$

Let f,Q be the pdf and tail of $(X/2E)^k$. Then

$$\frac{\mathbb{E}[X^k]}{2^kE^k} = \int_0^\infty f(t)tdt = -\int_0^\infty tdQ(t) = -[Q(t)t]_0^\infty + \int_0^\infty Q(t)dt \tag{A.3}$$

$$= \int_0^\infty Q(t)dt \le 1+\int_1^\infty e^{-Et^{2/k}/2}dt = 1+\sum_{i=0}^\infty\int_{e^{ik/2}}^{e^{(i+1)k/2}} e^{-Et^{2/k}/2}dt \tag{A.4}$$

$$\le 1+\sum_{i=0}^\infty e^{(i+1)k/2}e^{-Ee^i/2} \tag{A.5}$$

$$\le 1+\sum_{i=0}^\infty e^{-i-1}\le 2. \tag{A.6}$$

For the final line to hold, we simply need $e^{Ee^i/2}\ge e^{i+1}e^{(i+1)k/2}$. Taking logs of both sides, this reduces to $Ee^i/2\ge(i+1)(k+2)/2$ which is the same as

$$Ee^i\ge(i+1)(k+2).$$

Clearly, this holds for all i when $E\ge k+2$. ■

B Generalizations and Proofs for Learning with Train-Validation Split (Theorems 1, 2 & Proposition 1)

B.1 Uniform Convergence of Loss Functionals

We will show uniform concentration of an m-dimensional functional $\mathbb{F}$ of the loss function i.e. we shall first study $\mathbb{F}\ell(y, f(\boldsymbol{x})) \in \mathbb{R}^m$. To obtain the results on loss function or its gradient, we can set $\mathbb{F}$ to be identity and ∇ operator respectively. More generally, $\mathbb{F}$ can also represent the Hessian or projection of the gradient to proper subspaces of interest. Associated with the functional $\mathbb{F}$ we define the $\mathbb{F}$ distance metric to be

$$\text{dist}_{\mathbb{F}}(f_1, f_2) = \sup_{y\in\mathcal{Y}, \boldsymbol{x}\in\mathcal{X}} \|\mathbb{F}\ell(y, f_1(\boldsymbol{x})) - \mathbb{F}\ell(y, f_2(\boldsymbol{x}))\|_{\ell_2}.$$

Note that, for simplicity of exposition in our notation we dropped the dependence on $\mathcal{X}, \mathcal{Y}, \ell$. This distance will serve as a proxy for the $\|\cdot\|_{\mathcal{X}}$-norm in the generalized analysis. We also loosen Assumption 1 to relax global Lipschitzness condition so that $\mathcal{A}$ can instead be **approximately locally-Lipschitz**. Thus, the following provides a generalization of Assumption 1.

Assumption 6 *Suppose there exists a partitionining $\mathcal{P} = (\boldsymbol{\Delta}_i)_{i=1}^P$ of $\boldsymbol{\Delta}$ such that $\log(P) \leq h_{\text{eff}} \log(\bar{C})$ and the algorithm $\mathcal{A}$ satisfies the following. Over each $\boldsymbol{\Delta}_i \subset \mathcal{P}$, $\mathcal{A}(\cdot)$ is an (L, ε_0)-Lipschitz approximable function of $\boldsymbol{\alpha}$ in dist$_{\mathbb{F}}(\cdot,\cdot)$ distance. That is, there exists a function $g_{\boldsymbol{\alpha}}$ such that, dist$_{\mathbb{F}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}, g_{\boldsymbol{\alpha}}) \leq \varepsilon_0$ over $\boldsymbol{\Delta}_i$ and $g_{\boldsymbol{\alpha}}$ is L-Lipschitz function of $\boldsymbol{\alpha}$ in dist$_{\mathbb{F}}(\cdot,\cdot)$ that is dist$_{\mathbb{F}}(g_{\boldsymbol{\alpha}_1}, g_{\boldsymbol{\alpha}_2}) \leq L\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2}$.*

Assumption 7 *$\mathbb{F}\ell(y, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x})) - \mathbb{E}[\mathbb{F}\ell(y, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x}))]$ has subexponential ($\|\cdot\|_{\psi_1}$) norm bounded by some $S \geq 1$ with respect to the randomness in $(\boldsymbol{x}, y) \sim \mathcal{D}$.*

The following lemma is obtained as a corollary of Lemma D.7 of [58] by specializing it to the unit Euclidean ball.

Lemma 6 *Consider the empirical and population functionals $\mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) = \frac{1}{n_{\mathcal{T}}} \sum_{i=1}^{n_{\mathcal{T}}} \mathbb{F}\ell(y_i, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x}_i))$ and $\mathbb{F}\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) = \mathbb{E}[\mathbb{F}\ell(y, f_{\boldsymbol{\alpha}}^{\mathcal{T}}(\boldsymbol{x}))]$ respectively. Suppose $n_{\mathcal{V}} \geq m$. Then*

$$\mathbb{P}\left\{ \left\| \mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) - \mathbb{F}\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) \right\|_{\ell_2} \gtrsim S\frac{\sqrt{m+t^2}}{\sqrt{n_{\mathcal{V}}}} \right\} \leq 2e^{-\min(t^2, t\sqrt{n_{\mathcal{V}}})}.$$

With this lemma in place we are now ready to state our uniform concentration result.

Theorem 6 (Uniform Concentration over Validation) *Suppose Assumptions 6 and 7 hold. Fix $\tau > 0$ and set*

$$\bar{h}_{\text{eff}} := h_{\text{eff}} \log\left(\bar{C}L\sqrt{\frac{n_{\mathcal{V}}}{h_{\text{eff}}}}\right).$$

Then, as long as $n_{\mathcal{V}} \geq \bar{h}_{\text{eff}} + m + \tau$, with probability at least $1 - 2\mathrm{e}^{-\tau}$, $\mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}$ uniformly converges as follows

$$\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \left\| \mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) - \mathbb{F}\mathcal{L}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) \right\|_{\ell_2} \leq S\sqrt{\frac{C(\bar{h}_{\text{eff}} + m + \tau)}{n_{\mathcal{V}}}} + 4\varepsilon_0, \tag{B.1}$$

with $C > 0$ a fixed numerical constant.

Proof Let $\mathcal{C}_\varepsilon^i$ be an ε/L-cover of $\mathbf{\Delta}_i$ in ℓ_2 norm and $\mathcal{C}_\varepsilon = \bigcup_{i=1}^P \mathcal{C}_\varepsilon^i$. Recall from Definition 1 that the covering bound for $\mathbf{\Delta}_i \subset \mathbf{\Delta}$ obeys $|\mathcal{C}_\varepsilon^i| \leq h_{\text{eff}} \log(\bar{C}L/\varepsilon)$. Using the bound on P (via Assumption 6), we arrive at

$$\log|\mathcal{C}_\varepsilon| \leq \log(P) + h_{\text{eff}} \log(\bar{C}L/\varepsilon) \leq 2h_{\text{eff}} \log(\bar{C}L/\varepsilon).$$

Let $F = \{f_{\boldsymbol{\alpha}} \,|\, \boldsymbol{\alpha} \in \mathbf{\Delta}\}$. Let $G = \{g_{\boldsymbol{\alpha}} \,|\, \boldsymbol{\alpha} \in \mathbf{\Delta}\}$ be the set of (locally) Lipschitz functions $g_{\boldsymbol{\alpha}}$ within ε_0 neighborhood of F. Following Assumption 6, let G_ε be the $\text{dist}_{\mathbb{F}}(\cdot,\cdot)$ ε-cover of G induced by $\mathcal{C}_\varepsilon$. Additionally, let $F_\varepsilon \subset F$ be a set of hypothesis with same cardinality as G_ε that are within ε_0 distance to their counterparts in G_ε. For a fixed $f \in F_\varepsilon$, using sub-exponentiality of the loss functional (Assumption 7) and subexponential concentration provided in Lemma 6, for a proper choice of constant $C > 0$, with probability $1 - 2\mathrm{e}^{-\min(t\sqrt{n}, t^2)}$, we have that

$$\left\| \mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f) - \mathbb{F}\mathcal{L}(f) \right\|_{\ell_2} \leq \frac{S\sqrt{C}\sqrt{m+t^2}}{2\sqrt{n_{\mathcal{V}}}}.$$

Recall that we choose $n_{\mathcal{V}} \geq h_{\text{eff}} \log(\bar{C}L/\varepsilon) + m + \tau$. Also set the short-hand notation $\tilde{h}_{\text{eff}} = h_{\text{eff}} \log(\bar{C}L/\varepsilon)$. Thus setting $t = \sqrt{2\tilde{h}_{\text{eff}} + \tau} \leq \sqrt{n_{\mathcal{V}}}$ ensures

$$\min(t\sqrt{n}, t^2) \geq t^2 \geq \log|\mathcal{C}_\varepsilon| + \tau.$$

Thus, union bounding over F_ε, we find that with probability $1 - 2\mathrm{e}^{-\tau}$, for all $f \in F_\varepsilon$

$$\|\mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f) - \mathbb{F}\mathcal{L}(f)\|_{\ell_2} \leq S\sqrt{C}\frac{\sqrt{m + \tilde{h}_{\text{eff}} + \tau}}{\sqrt{n_{\mathcal{V}}}}. \tag{B.2}$$

Now, fix any $f \in F$. Pick $g \in G$ and $g' \in G_\varepsilon$ such that $\text{dist}_{\mathbb{F}}(g, g') \leq \varepsilon$ and $\text{dist}_{\mathbb{F}}(g, f) \leq \varepsilon_0$. Additionally pick $f' \in F_\varepsilon$ which is ε_0-neighbor of $g' \in G_\varepsilon$. This implies that, for all feasible $(\boldsymbol{x}, y)$

$$\begin{aligned}
&\|\mathbb{F}\ell(y, f(\boldsymbol{x})) - \mathbb{F}\ell(y, f'(\boldsymbol{x}))\|_{\ell_2} \\
&\quad \leq \|\mathbb{F}\ell(y, f(\boldsymbol{x})) - \mathbb{F}\ell(y, g(\boldsymbol{x}))\|_{\ell_2} + \|\mathbb{F}\ell(y, g(\boldsymbol{x})) - \mathbb{F}\ell(y, f'(\boldsymbol{x}))\|_{\ell_2} + \|\mathbb{F}\ell(y, g'(\boldsymbol{x})) - \mathbb{F}\ell(y, f'(\boldsymbol{x}))\|_{\ell_2} \\
&\quad \leq \varepsilon + 2\varepsilon_0,
\end{aligned}$$

via Assumption 6. This further implies the same bound for population and empirical functionals

$$\|\mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f) - \mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f')\|_{\ell_2}, \|\mathbb{F}\mathcal{L}(f) - \mathbb{F}\mathcal{L}(f')\|_{\ell_2} \leq \varepsilon + 2\varepsilon_0.$$

Combining this with (B.2) leads to the following uniform convergence bound for all $f \in F$

$$\sup_{f \in F} \|\mathbb{F}\mathcal{L}(f) - \mathbb{F}\widehat{\mathcal{L}}_{\mathcal{V}}(f')\|_{\ell_2} \leq S\frac{\sqrt{C(\tilde{h}_{\text{eff}} + m + \tau)}}{\sqrt{n_{\mathcal{V}}}} + 2(\varepsilon + 2\varepsilon_0).$$

To proceed, select $\varepsilon = \sqrt{\frac{h_{\text{eff}}}{n_{\mathcal{V}}}}$ and set $\bar{h}_{\text{eff}} = h_{\text{eff}} \log(\bar{C}L\sqrt{n_{\mathcal{V}}/h_{\text{eff}}})$. Thus,

$$\sup_{f \in F} |\mathcal{L}(f_{\boldsymbol{\alpha}}) - \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}})| \leq S\sqrt{\frac{C(\bar{h}_{\text{eff}} + m + \tau)}{n_{\mathcal{V}}}} + 4\varepsilon_0, \tag{B.3}$$

concluding the proof of the bound. ■

B.2 Proof of Theorem 1

Let us state slight generalization of Assumptions 1 and 1′ which will be utilized in our proof.

Assumption 8 *There exists a function $g_{\boldsymbol{\alpha}}$ such that, for all pairs $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \boldsymbol{\Delta}$, $\|g_{\boldsymbol{\alpha}_1} - f^{\mathcal{T}}_{\boldsymbol{\alpha}_1}\|_{\mathcal{X}} \leq \varepsilon_0$ and $\|g_{\boldsymbol{\alpha}_1} - g_{\boldsymbol{\alpha}_2}\|_{\mathcal{X}} \leq L\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2}$.*

Assumption 8′ *For some $R \geq 1$ and all $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \boldsymbol{\Delta}$ and $\boldsymbol{x} \in \mathcal{X}$, hyper-gradient obeys $\|\nabla_{\boldsymbol{\alpha}} f^{\mathcal{T}}_{\boldsymbol{\alpha}_1}(\boldsymbol{x})\|_{\ell_2} \leq R$, $\|\nabla_{\boldsymbol{\alpha}} f^{\mathcal{T}}_{\boldsymbol{\alpha}_1}(\boldsymbol{x}) - \nabla_{\boldsymbol{\alpha}} f^{\mathcal{T}}_{\boldsymbol{\alpha}_2}(\boldsymbol{x})\|_{\ell_2} \leq RL\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2}$ and $\|\nabla_{\boldsymbol{\alpha}} g_{\boldsymbol{\alpha}_1}(\boldsymbol{x}) - \nabla_{\boldsymbol{\alpha}} f^{\mathcal{T}}_{\boldsymbol{\alpha}_1}(\boldsymbol{x})\|_{\ell_2} \leq R\varepsilon_0$.*

The proof will be a corollary of the generalized result Theorem 6. Let us first state a lemma to show that Assumptions 2 and 8 imply Assumptions 6 and 7.

Lemma 7 *Let $C > 0$ be a proper choice of constant.*

- *Assumptions 2 and 8 imply Assumptions 6 and 7 hold for the loss function (setting $\mathbb{F} = $ Identity), with $m = 1$, $L \to L\Gamma$, $\varepsilon_0 \to \Gamma\varepsilon_0$, and $S = C$.*
- *Assumptions 2′ and 8′ imply Assumptions 6 and 7 hold for the gradient (setting $\mathbb{F} = \nabla$), with $m = h$, $L \to 2RL\Gamma$, $\varepsilon_0 \to 2R\Gamma\varepsilon_0$, and $S = C$.*

Proof For the first statement, we conclude that Assumption 6 holds with $L \to L\Gamma$ via

$$\sup_{y\in\mathcal{Y},\boldsymbol{x}\in\mathcal{X}} |\ell(y, f_{\boldsymbol{\alpha}_1}(\boldsymbol{x})) - \ell(y, f_{\boldsymbol{\alpha}_2}(\boldsymbol{x}))| \leq \Gamma \sup_{\boldsymbol{x}\in\mathcal{X}} |f_{\boldsymbol{\alpha}_1}(\boldsymbol{x}) - f_{\boldsymbol{\alpha}_2}(\boldsymbol{x})| \leq L\Gamma\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2}.$$

Following the same argument, we plug in $\varepsilon_0 \to \Gamma\varepsilon_0$. To prove the result for the gradient mapping note that if Assumptions 2′ & 8′ hold then, using the fact that $\nabla_{\boldsymbol{\alpha}}\ell(y, f_{\boldsymbol{\alpha}}(\boldsymbol{x})) = \ell'(y, f_{\boldsymbol{\alpha}}(\boldsymbol{x}))\nabla f_{\boldsymbol{\alpha}}(\boldsymbol{x})$ we have

$$\begin{aligned}\|\nabla\ell(y, f_{\boldsymbol{\alpha}_1}(\boldsymbol{x})) - \nabla\ell(y, f_{\boldsymbol{\alpha}_2}(\boldsymbol{x}))\|_{\ell_2} &\leq |\ell'(y, f_{\boldsymbol{\alpha}_1}(\boldsymbol{x})) - \ell'(y, f_{\boldsymbol{\alpha}_2}(\boldsymbol{x}))|\|\nabla f_{\boldsymbol{\alpha}_1}(\boldsymbol{x})\|_{\ell_2} + \Gamma\|\nabla f_{\boldsymbol{\alpha}_1}(\boldsymbol{x}) - \nabla f_{\boldsymbol{\alpha}_2}(\boldsymbol{x})\|_{\ell_2} \\ &\leq \Gamma|f_{\boldsymbol{\alpha}_1}(\boldsymbol{x}) - f_{\boldsymbol{\alpha}_2}(\boldsymbol{x})|R + \Gamma RL\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2} \\ &\leq 2\Gamma LR\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_{\ell_2}.\end{aligned}$$

Following the same argument, we plug in $\varepsilon_0 \to 2R\Gamma\varepsilon_0$. Subexponentiality with $S = C$ follows from the boundedness condition in Assumption 2. ∎

Theorem 1 directly follows by setting $\varepsilon_0 = 0$ and plugging in the looser bound $n_{\mathcal{V}}/h_{\text{eff}} \geq \sqrt{n_{\mathcal{V}}/h_{\text{eff}}}$ in the definition of $\bar{h}_{\text{eff}}$ term. Note that, the theorem below is a corollary of the more general result in Theorem 6.

Theorem 7 (Learning with Validation – Lipschitz Approximation) *Suppose Assumptions 2 & 8 hold. Let $\widehat{\boldsymbol{\alpha}}$ be the minimizer of the empirical risk (TVO) over validation. Fix $\tau > 0$ and set $\bar{h}_{\mathit{eff}} := h_{\mathit{eff}} \log(\bar{C}L\Gamma\sqrt{n_{\mathcal{V}}/h_{\mathit{eff}}})$. There exists a constant $C > 0$ such that, whenever $n_{\mathcal{V}} \geq \bar{h}_{\mathit{eff}} + \tau$, with probability at least $1 - 2\mathrm{e}^{-\tau}$, $f_{\widehat{\boldsymbol{\alpha}}}$ achieves the risk bound*

$$\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} |\mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) - \widehat{\mathcal{L}}_{\mathcal{V}}(f^{\mathcal{T}}_{\boldsymbol{\alpha}})| \leq \sqrt{\frac{C(\bar{h}_{\mathit{eff}} + \tau)}{n_{\mathcal{V}}}} + 4\Gamma\varepsilon_0, \tag{B.4}$$

$$\mathcal{L}(f^{\mathcal{T}}_{\widehat{\boldsymbol{\alpha}}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) + 2\sqrt{\frac{C(\bar{h}_{\mathit{eff}} + \tau)}{n_{\mathcal{V}}}} + 8\Gamma\varepsilon_0 + \delta. \tag{B.5}$$

Furthermore, suppose Assumptions 2′ & 8′ hold as well. Set $\bar{h}^{\nabla}_{\mathit{eff}} := h + h_{\mathit{eff}} \log(2\bar{C}RL\Gamma\sqrt{n_{\mathcal{V}}/h_{\mathit{eff}}})$. *Whenever* $n_{\mathcal{V}} \geq \bar{h}^{\nabla}_{\mathit{eff}} + \tau$, *with probability at least* $1 - 2\mathrm{e}^{-\tau}$,

$$\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \|\nabla\widehat{\mathcal{L}}_{\mathcal{V}}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) - \nabla\mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}})\|_{\ell_2} \leq \sqrt{\frac{C(\bar{h}^{\nabla}_{\mathit{eff}} + \tau)}{n_{\mathcal{V}}}} + 8R\Gamma\varepsilon_0. \tag{B.6}$$

Proof Applying Theorem 6 and plugging in the first statement of Lemma 7, with probability $1 - 2e^{-\tau}$, we obtain the statement (B.4). Here we used the fact that $\bar{h}_{\text{eff}} + \tau + m \leq 2(\bar{h}_{\text{eff}} + \tau)$ and factor 2 can be subsumed in the constant C.

We obtain the advertised bound (B.5) via (a) observing that the bound is valid for $\widehat{\boldsymbol{\alpha}}$ and the optimal population hypothesis $\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \mathcal{L}(f_{\boldsymbol{\alpha}})$ and (b) using the fact that

$$\widehat{\mathcal{L}}_{\mathcal{V}}(f_{\widehat{\boldsymbol{\alpha}}}) \leq \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}) + \delta \leq \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}^\star}) + \delta.$$

Following (B.4), we first have $\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}) \leq \min_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_{\mathcal{V}}(f_{\boldsymbol{\alpha}}) + \sqrt{\frac{C(\bar{h}_{\text{eff}}+\tau)}{n_{\mathcal{V}}}} + 4\varepsilon_0\Gamma + \delta$. Using a triangle inequality with $\boldsymbol{\alpha}^\star$, we find

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}) \leq \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \mathcal{L}(f_{\boldsymbol{\alpha}}) + 2\sqrt{\frac{C(\bar{h}_{\text{eff}} + \tau)}{n_{\mathcal{V}}}} + 8\varepsilon_0\Gamma + \delta \tag{B.7}$$

Applying Theorem 6 and plugging in the second statement of Lemma 7, with probability $1 - 2e^{-\tau}$, we obtain the statement (B.6). Here, we used the fact that $m = h$ and $\bar{h}_{\text{eff}} = h_{\text{eff}} \log(2\bar{C}RL\Gamma\sqrt{n_{\mathcal{V}}/h_{\text{eff}}})$. We then set $\bar{h}^{\nabla}_{\text{eff}} = \bar{h}_{\text{eff}} + h$ to conclude. ■

B.3 Proof of Proposition 1

For the purpose of our neural network analysis assuming $\mathcal{C}_{\boldsymbol{\alpha}}$ to be Lipschitz might be too restrictive. As a result, we will state a slight generalization which allows for Lipschitzness of $\mathcal{C}_{\boldsymbol{\alpha}}$ and $f^{\mathcal{D}}_{\boldsymbol{\alpha}}$ over a smaller domain $\boldsymbol{\Delta}_\star$ which provides more flexibility.

Assumption 9 *Over a domain* $\boldsymbol{\Delta}_\star \subset \boldsymbol{\Delta}$: *(1)* $f^{\mathcal{D}}_{\boldsymbol{\alpha}}$ *is* L*-Lipschitz function of* $\boldsymbol{\alpha}$ *in* $\|\cdot\|_{\mathcal{X}}$ *norm and (2) the excess risk term* $\mathcal{C}_{\boldsymbol{\alpha}}$ *is* $\kappa L\sqrt{n_{\mathcal{T}}}$*-Lipschitz for some* $\kappa > 0$.

Note that, the first condition is equivalent to Assumption 1 holding over $\boldsymbol{\Delta}_\star$ in population (i.e. $n_{\mathcal{T}} \to \infty$). We intentionally parameterized the Lipschitz constant by the same notation to simplify exposition. The following result is a slight generalization of Proposition 1 where we allow for non-Lipschitzness of $\boldsymbol{\Delta}$ by instead assuming it holds over a smaller subset $\boldsymbol{\Delta}_\star$.

Proposition 2 (Train-Validation Bounds) *Consider the setting of Theorem 1 and for any fixed* $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ *assume (3.4) holds. Additionally suppose Assumption 9 holds. Then with probability at least* $1 - e^{-\tau}$,

$$\min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}_\star} \left(\mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}}) + \frac{\mathcal{C}^{\mathcal{T}}_{\boldsymbol{\alpha}}}{\sqrt{n_{\mathcal{T}}}}\right) + 2C_0\sqrt{\frac{h_{\mathit{eff}}\log(2(\Gamma+\kappa)n_{\mathcal{T}}\bar{C}L/h_{\mathit{eff}}) + \tau}{n_{\mathcal{T}}}}. \tag{B.8}$$

Using Theorem 1, this in turn implies that with probability at least $1-3e^{-\tau}$

$$\mathcal{L}(f^{\mathcal{T}}_{\widehat{\boldsymbol{\alpha}}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}_\star}\left(\mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}}) + \frac{\mathcal{C}^{\mathcal{T}}_{\boldsymbol{\alpha}}}{\sqrt{n_{\mathcal{T}}}}\right) + 2C_0\sqrt{\frac{h_{\mathit{eff}}\log(2(\Gamma+\kappa)n_{\mathcal{T}}\bar{C}L/h_{\mathit{eff}})+\tau}{n_{\mathcal{T}}}} + \sqrt{\frac{Ch_{\mathit{eff}}\log(2\bar{C}L\Gamma n_{\mathcal{V}}/h_{\mathit{eff}})}{n_{\mathcal{V}}}} + \delta. \tag{B.9}$$

Observe that (B.9) *implies Proposition 1 as the sample size setting of the paper is* $n_{\mathcal{T}} \geq n_{\mathcal{V}}$ *via* (1.1).

Proof The proof of (B.8) uses a covering argument. Create an ε/L covering $\boldsymbol{\Delta}_\varepsilon$ of the set $\boldsymbol{\Delta}_\star$ of size $\log|\boldsymbol{\Delta}_\varepsilon| \leq h_{\text{eff}}\log(\bar{C}L/\varepsilon)$. For each $\boldsymbol{\alpha}\in\boldsymbol{\Delta}_\star$, setting $t = h_{\text{eff}}\log(\bar{C}L/\varepsilon)+\tau$ we have that for all $\boldsymbol{\alpha}_\varepsilon \in \boldsymbol{\Delta}_\varepsilon$, the bound (3.4) holds with probability $1-e^{-\tau}$. To conclude set $\varepsilon = \frac{C_0\sqrt{h_{\text{eff}}/n_{\mathcal{T}}}}{2\Gamma+2\kappa} \geq \frac{C_0 h_{\text{eff}}/n_{\mathcal{T}}}{2\Gamma+2\kappa}$ and $\bar{h}_{\text{eff}} = h_{\text{eff}}\log(2(\Gamma+\kappa)n_{\mathcal{T}}\bar{C}L/h_{\text{eff}})+\tau$. Also observe, via Assumption 9, that

$$|\mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}_\varepsilon}) - \mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}})| + \frac{|\mathcal{C}^{\mathcal{T}}_{\boldsymbol{\alpha}_\varepsilon} - \mathcal{C}^{\mathcal{T}}_{\boldsymbol{\alpha}}|}{\sqrt{n_{\mathcal{T}}}} \leq \Gamma\varepsilon + \kappa\varepsilon \leq 0.5C_0\sqrt{h_{\text{eff}}/n_{\mathcal{T}}}.$$

To proceed, via Assumption 1 we also have $|\mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}_\varepsilon}) - \mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}})| \leq \Gamma\varepsilon \leq 0.5C_0\sqrt{h_{\text{eff}}/n_{\mathcal{T}}}$. Together, using triangle inequality, these imply for all $\boldsymbol{\alpha}\in\boldsymbol{\Delta}_\star$

$$\begin{aligned}
&\mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) \leq \mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}_\varepsilon}) + 0.5C_0\sqrt{h_{\text{eff}}/n_{\mathcal{T}}} + C_0\sqrt{\frac{\bar{h}_{\text{eff}}}{n_{\mathcal{T}}}} \leq \mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}}) + \frac{\mathcal{C}^{\mathcal{T}}_{\boldsymbol{\alpha}}}{\sqrt{n_{\mathcal{T}}}} + 2C_0\sqrt{\bar{h}_{\text{eff}}/n_{\mathcal{T}}}, \\
\implies &\mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) \leq \mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}}) + \frac{\mathcal{C}^{\mathcal{T}}_{\boldsymbol{\alpha}}}{\sqrt{n_{\mathcal{T}}}} + 2C_0\sqrt{\frac{\bar{h}_{\text{eff}}}{n_{\mathcal{T}}}}.
\end{aligned} \tag{B.10}$$

Finally to conclude with (B.8) notice the fact that $\min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}_\star}\mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}})$.

To conclude with the final statement (B.9), we apply Theorem 1 which bounds $\widehat{\mathcal{L}}_{\mathcal{V}}(f^{\mathcal{T}}_{\widehat{\boldsymbol{\alpha}}})$ in terms of $\min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}})$. This results in an overall success probability of $1-3e^{-\tau}$. ■

B.4 Proof of Theorem 2

Proof Below $C>0$ is an absolute constant. First under the provided conditions (which include (4.3) with probability $1-p_0$), applying Lemma 12, we find that Assumption 1 ($\mathcal{A}$ is Lipschitz) holds with Lipschitz constant $L = 6R^3B^2\sqrt{n_{\mathcal{T}}^3 h\lambda_0^{-2}}\|\boldsymbol{y}\|_{\ell_2}$. Assumption 2 holds automatically and Assumption 1 holds with $h_{\text{eff}} = h$ and $\bar{C} = 3$. Thus, applying Theorem 7 (with $\varepsilon_0 = 0$), with probability at least $1-2e^{-\tau}$

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}) \leq \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\mathcal{L}(f_{\boldsymbol{\alpha}}) + C\frac{\sqrt{h\log(6RL\Gamma\sqrt{n_{\mathcal{V}}/h})+\tau}}{\sqrt{n_{\mathcal{V}}}} + \delta.$$

The remaining task is bounding the $\inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\mathcal{L}(f_{\boldsymbol{\alpha}}) - \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\mathcal{L}(f^{\mathcal{D}}_{\boldsymbol{\alpha}})$ term. We do this via Lemma 8 which yields that with probability $1-2e^{-t}$, for all $\boldsymbol{\alpha}\in\boldsymbol{\Delta}$ (B.11) holds. Together, these imply that with probability

$1 - 4e^{-\tau} - p_0$

$$\begin{aligned}\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}) - \delta &\leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} C\frac{\sqrt{h\log(6R\times 6R^3B^2\sqrt{n_{\mathcal{T}}^3 h\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}\Gamma\sqrt{n_{\mathcal{V}}/h})+\tau}}}{\sqrt{n_{\mathcal{V}}}} \\ &\quad + \frac{2\Gamma\sqrt{B\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}} + C\sqrt{h\log(20R^4B^2\lambda_0^{-2}\Gamma n_{\mathcal{T}}^2\|\boldsymbol{y}\|_{\ell_2})+\tau}}{\sqrt{n_{\mathcal{T}}}} \\ &\leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \frac{C\sqrt{h\log(M)+\tau}}{\sqrt{n_{\mathcal{V}}}} + \frac{2\Gamma\sqrt{B\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}} + C\sqrt{h\log(M)+\tau}}{\sqrt{n_{\mathcal{T}}}} \\ &\leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \frac{2\Gamma\sqrt{B\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}}}{\sqrt{n_{\mathcal{T}}}} + \frac{C\sqrt{h\log(M)+\tau}}{\sqrt{\min(n_{\mathcal{T}},n_{\mathcal{V}})}}\end{aligned}$$

where $M = 30R^4B^2\lambda_0^{-2}\Gamma(n_{\mathcal{T}}^2 + n_{\mathcal{V}}^2)\|\boldsymbol{y}\|_{\ell_2}$. ■

B.5 Uniform Concentration of Excess Risk for Feature Maps

Lemma 8 *Consider the setup of Definition 2. Let* $\sup_{\boldsymbol{x}\in\mathcal{X},1\leq i\leq h}\|\phi_i(\boldsymbol{x})\|_{\ell_2}^2 \leq B$. *Declare* $\boldsymbol{K}_{\boldsymbol{\alpha}} = \boldsymbol{\Phi}_{\boldsymbol{\alpha}}\boldsymbol{\Phi}_{\boldsymbol{\alpha}}^T$. *Suppose* (4.3) *holds with probability* $1-p_0$ *over the training data. Assume* ℓ *is a* Γ *Lipschitz loss bounded by* $C \geq 1$. *With probability at least* $1 - p_0 - \mathrm{e}^{-t}$, *we have that for all* $\boldsymbol{\alpha}\in\boldsymbol{\Delta}$

$$\mathcal{L}(\boldsymbol{\theta}_{\boldsymbol{\alpha}}) \leq \frac{2\Gamma\sqrt{B\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}} + 2C\sqrt{h\log(20R^4B^2\lambda_0^{-2}\Gamma n_{\mathcal{T}}^2\|\boldsymbol{y}\|_{\ell_2})+\tau}}{\sqrt{n_{\mathcal{T}}}} \tag{B.11}$$

Proof The proof strategy follows that of Proposition 1. Using (A.2) of Corollary 1, we know that, for any choice of $\boldsymbol{\alpha}\in\boldsymbol{\Delta}$ with probability $1-\mathrm{e}^{-t}$ we have that

$$\mathcal{L}_{\mathcal{D}}(\hat{f}) \leq \frac{2\Gamma\sqrt{B\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}} + C\sqrt{t}}{\sqrt{n_{\mathcal{T}}}}. \tag{B.12}$$

where $\boldsymbol{K}_{\boldsymbol{\alpha}} = \boldsymbol{\Phi}_{\boldsymbol{\alpha}}\boldsymbol{\Phi}_{\boldsymbol{\alpha}}^T$. To proceed, we will apply a covering argument to the population loss of the empirical solutions. This will require Lipschitzness of the population loss. Observe that

$$\sqrt{\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}} = \|\boldsymbol{\theta}_{\boldsymbol{\alpha}}\|_{\ell_2} = \|\boldsymbol{\Phi}_{\boldsymbol{\alpha}}^{\dagger}\boldsymbol{y}\|_{\ell_2}.$$

Lemma 12 shows the Lipschitzness of $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ with $L = 5R^2\sqrt{B^3 n_{\mathcal{T}}^3 h\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}}$ which implies that

$$\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2} \leq L\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}.$$

To proceed, let $\mathcal{C}_\varepsilon$ be an ε cover of $\boldsymbol{\Delta}$ of size $\log|\mathcal{C}_\varepsilon| \leq h\log(3R/\varepsilon)$. Setting $t = h\log(3R/\varepsilon)+\tau$, with probability $1-\mathrm{e}^{-\tau}$, we find that all $\boldsymbol{\alpha}\in\mathcal{C}_\varepsilon$ obeys (B.12) with this choice of t. Observe that, with probability at least $1-p_0$, via the lower bound (4.3), we have $\|\boldsymbol{\theta}_{\boldsymbol{\alpha}}\|_{\ell_2}, \|\boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2} \leq \|\boldsymbol{y}\|_{\ell_2}/\sqrt{\lambda_0}$. To proceed, for any $\boldsymbol{\alpha}$

pick $\boldsymbol{\alpha}' \in \mathcal{C}_\varepsilon$ with $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2} \le \varepsilon$ and observe that

$$\begin{aligned}
\mathcal{L}(f_{\boldsymbol{\alpha}}) - \mathcal{L}(f_{\boldsymbol{\alpha}'}) &= \mathbb{E}[\ell(y, \boldsymbol{\theta}_{\boldsymbol{\alpha}}^T \phi_{\boldsymbol{\alpha}}(\boldsymbol{x})] - \mathbb{E}[\ell(y, \boldsymbol{\theta}_{\boldsymbol{\alpha}'}^T \phi_{\boldsymbol{\alpha}'}(\boldsymbol{x})] \\
&\le \Gamma\, \mathbb{E}[|\boldsymbol{\theta}_{\boldsymbol{\alpha}}^T \phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}^T \phi_{\boldsymbol{\alpha}'}(\boldsymbol{x})|] \\
&\le \Gamma[\mathbb{E}[|(\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'})^T \phi_{\boldsymbol{\alpha}}(\boldsymbol{x})|] + \mathbb{E}[|\boldsymbol{\theta}_{\boldsymbol{\alpha}'}^T (\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \phi_{\boldsymbol{\alpha}'}(\boldsymbol{x}))|]] \\
&\le \Gamma[\mathbb{E}[\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2} \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{x})\|_{\ell_2}] + \mathbb{E}[\|\boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2} \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \phi_{\boldsymbol{\alpha}'}(\boldsymbol{x})\|_{\ell_2}]] \\
&\le \Gamma(RL\sqrt{B}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2} + \sqrt{hB}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2} \|\boldsymbol{y}\|_{\ell_2}/\sqrt{\lambda_0}) \\
&\le \Gamma\sqrt{B}\|\boldsymbol{y}\|_{\ell_2}(5R^3 B^{3/2}\sqrt{n_{\mathcal{T}}^3 h}\lambda_0^{-2} + \sqrt{\frac{h}{\lambda_0}})\varepsilon \\
&\le 6R^3 B^2 \Gamma \sqrt{n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}\varepsilon,
\end{aligned}$$

where we used the fact that $\lambda_0 \le \sigma_{\min}^2(\boldsymbol{\Phi}_{\boldsymbol{\alpha}}) \le n_{\mathcal{T}} R^2 B$. Thus we obtain that for all $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$

$$\mathcal{L}_{\mathcal{D}}(\hat{f}) \le \frac{2\Gamma\sqrt{B\boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}} + C\sqrt{h\log(3R/\varepsilon) + \tau}}{\sqrt{n_{\mathcal{T}}}} + 6R^3 B^2 \Gamma \sqrt{n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}\varepsilon.$$

Here setting $\varepsilon^{-1} = 20R^4 B^2 \lambda_0^{-2} \Gamma n_{\mathcal{T}}^2 \|\boldsymbol{y}\|_{\ell_2}/3R$, we obtain the desired bound (B.11). ■

C Results on Algorithmic Lipschitzness

C.1 Proof of Lemma 1

Proof First applying Lemma 12, we find that Assumption 1 holds with (C.8). Plugging in the boundedness of labels (i.e. $\|\boldsymbol{y}\|_{\ell_2} \le \sqrt{n_{\mathcal{T}}}$), $\Gamma = 1$, $n_{\mathcal{T}} \ge h$, $R \ge 1$, we end up with the refined bound

$$L \le 6R^3 B n_{\mathcal{T}} \lambda_0^{-2}(B n_{\mathcal{T}}\sqrt{h} + 1).$$

Finally, using the fact that $\boldsymbol{\alpha}$ is $h+1$ dimensional, we obtain $\bar{h}_{\text{eff}} = (h+1)\log(20R^3 B n_{\mathcal{T}}^2 \lambda_0^{-2}(B n_{\mathcal{T}} + 1))$. ■

C.2 Proof of Lemma 2

The following lemma is a rephrasing of Lemma 2.

Lemma 9 (Lipschitz Solutions under Strong-Convexity / Smoothness) *Let $\boldsymbol{\Delta}$ be a convex set. Let $\mathcal{F}(\boldsymbol{\alpha}, \boldsymbol{\theta}) : \boldsymbol{\Delta} \times \mathbb{R}^p \to \mathbb{R}$ be a μ strongly-convex function of $\boldsymbol{\theta}$ and L-smooth function of $\boldsymbol{\alpha}$ over the feasible domain $\mathbb{R}^p \times \boldsymbol{\Delta}$. Let*

$$\boldsymbol{\theta}_{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\alpha}, \boldsymbol{\theta}).$$

Then, $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ is $\sqrt{L/\mu}$-Lipschitz function of $\boldsymbol{\alpha}$.

Proof Pick a pair $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \boldsymbol{\Delta}$. From strong-convexity of $\boldsymbol{\theta}$ and optimality of $\boldsymbol{\theta}_{\boldsymbol{\alpha}'}$, we have that

$$\mathcal{F}(\boldsymbol{\alpha}', \boldsymbol{\theta}_{\boldsymbol{\alpha}}) - \mathcal{F}(\boldsymbol{\alpha}', \boldsymbol{\theta}_{\boldsymbol{\alpha}'}) \ge \frac{\mu}{2}\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2}^2.$$

On the other hand, smoothness of $\boldsymbol{\alpha}$ implies

$$|\mathcal{F}(\boldsymbol{\alpha}', \boldsymbol{\theta}_{\boldsymbol{\alpha}}) - \mathcal{F}(\boldsymbol{\alpha}, \boldsymbol{\theta}_{\boldsymbol{\alpha}})| \leq \frac{L}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}^2. \tag{C.1}$$

Putting these together,

$$\mathcal{F}(\boldsymbol{\alpha}, \boldsymbol{\theta}_{\boldsymbol{\alpha}}) + \frac{L}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}^2 \geq \mathcal{F}(\boldsymbol{\alpha}', \boldsymbol{\theta}_{\boldsymbol{\alpha}}) \geq \mathcal{F}(\boldsymbol{\alpha}', \boldsymbol{\theta}_{\boldsymbol{\alpha}'}) + \frac{\mu}{2}\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2}^2.$$

Since the inequality is symmetric with respect to $\boldsymbol{\alpha}, \boldsymbol{\alpha}'$, we also have

$$\mathcal{F}(\boldsymbol{\alpha}', \boldsymbol{\theta}_{\boldsymbol{\alpha}'}) + \frac{L}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}^2 \geq \mathcal{F}(\boldsymbol{\alpha}, \boldsymbol{\theta}_{\boldsymbol{\alpha}}) + \frac{\mu}{2}\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2}^2. \tag{C.2}$$

Summing up both sides yield the desired result

$$L\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}^2 \geq \mu\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2}^2.$$

■

C.3 Stability of Linear Regression

We begin by stating the following useful lemma.

Lemma 10 *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two positive semidefinite matrices with minimum eigenvalue of $\boldsymbol{A}$ bounded below by $\gamma > 0$. Set $\boldsymbol{P} = \boldsymbol{B} - \boldsymbol{A}$ and suppose $\|\boldsymbol{P}\| \leq \delta$. Then*

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}\| \leq \frac{\delta}{\gamma(\gamma - \delta)}. \tag{C.3}$$

Proof Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two positive semidefinite matrices with minimum eigenvalue of $\boldsymbol{A}$ bounded below by $\gamma > 0$. Set $\boldsymbol{P} = \boldsymbol{B} - \boldsymbol{A}$ and suppose $\|\boldsymbol{P}\| \leq \delta$. Let $\boldsymbol{A}$ have eigen decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$. Let $\bar{\boldsymbol{P}} = \boldsymbol{U}^T\boldsymbol{P}\boldsymbol{U}$ and $\tilde{\boldsymbol{P}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^T\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$. Observe that $\|\tilde{\boldsymbol{P}}\| \leq \gamma^{-1}\delta$. We have that

$$\begin{aligned}
\|\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}\| &= \|\boldsymbol{\Sigma}^{-1} - \boldsymbol{U}^T\boldsymbol{B}^{-1}\boldsymbol{U}\| \\
&= \|\boldsymbol{\Sigma}^{-1} - \boldsymbol{U}^T(\boldsymbol{A} + \boldsymbol{P})^{-1}\boldsymbol{U}\| \\
&= \|\boldsymbol{\Sigma}^{-1} - \boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T + \boldsymbol{U}\bar{\boldsymbol{P}}\boldsymbol{U}^T)^{-1}\boldsymbol{U}\| \\
&= \|\boldsymbol{\Sigma}^{-1} - (\boldsymbol{\Sigma} + \bar{\boldsymbol{P}})^{-1}\| \\
&\leq \|\boldsymbol{\Sigma}^{-1} - (\boldsymbol{\Sigma} + \bar{\boldsymbol{P}})^{-1}\| \\
&\leq \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{P}}\boldsymbol{\Sigma}^{-1/2})^{-1})\boldsymbol{\Sigma}^{-1/2}\| \\
&\leq \sigma_{\min}^{-1}(\boldsymbol{\Sigma})\|\boldsymbol{I} - (\boldsymbol{I} + \tilde{\boldsymbol{P}})^{-1}\| \\
&\leq \frac{1}{\gamma}\frac{\gamma^{-1}}{1 - \delta\gamma^{-1}} = \frac{\delta}{\gamma(\gamma - \delta)}.
\end{aligned}$$

For the last statement, we observe the fact that $(\boldsymbol{I} + \tilde{\boldsymbol{P}})^{-1} = \sum_{i \geq 0}(-1)^i\tilde{\boldsymbol{P}}^i$ which yields $\|\boldsymbol{I} - (\boldsymbol{I} + \tilde{\boldsymbol{P}})^{-1}\| \leq \frac{\gamma^{-1}\delta}{1 - \gamma^{-1}\delta}$. ■

The lemma below shows the stability of ridge(less) regression to feature or regularizer changes.

Lemma 11 (Feature Robustness of Linear Models) *Fix $\boldsymbol{X}, \bar{\boldsymbol{X}} \in \mathbb{R}^{n\times p}$ with $\|\boldsymbol{X}\|, \|\bar{\boldsymbol{X}}\| \leq B$. Fix $\lambda, \bar{\lambda} \geq 0$ and assume $\lambda_0 = \lambda + \sigma_{\min}^2(\boldsymbol{X}) > 0$. Suppose $2B\|\boldsymbol{X} - \bar{\boldsymbol{X}}\| + |\lambda - \bar{\lambda}| < \lambda_0/2$. Consider the ridge regression (or min Euclidean solution)*

$$\boldsymbol{\theta} = \mathcal{A}_{\textit{ridge}}(\boldsymbol{y}, \boldsymbol{X}, \lambda) = \begin{cases} (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} & \textit{if} \quad \lambda > 0 \textit{ or } n \geq p \\ \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y} & \textit{if} \quad \lambda = 0 \textit{ and } n < p \end{cases}.$$

Also define $\bar{\boldsymbol{\theta}} = \mathcal{A}_{\textit{ridge}}(\boldsymbol{y}, \bar{\boldsymbol{X}}, \bar{\lambda})$ which solves the problem with features $\bar{\boldsymbol{X}}$. In either cases (whether $\lambda > 0$ or not), we have that

$$\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_{\ell_2} \leq \frac{5B^2\|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2}\|\boldsymbol{X} - \bar{\boldsymbol{X}}\| + \frac{2B\|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2}|\lambda - \bar{\lambda}|. \tag{C.4}$$

Proof Suppose $\lambda > 0$ or $n > p$. Recall that $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I} \succeq \lambda_0\boldsymbol{I}$. Observe that $\boldsymbol{X}^T\boldsymbol{X} - \bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} = \boldsymbol{X}^T(\boldsymbol{X} - \bar{\boldsymbol{X}}) + (\boldsymbol{X} - \bar{\boldsymbol{X}})^T\bar{\boldsymbol{X}}$ which implies $\|\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I} - \bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} + \bar{\lambda}\boldsymbol{I}\| \leq \bar{B}$ where $\bar{B} = 2\varepsilon B + |\lambda - \bar{\lambda}|$. Recall that $\bar{B} \leq \lambda_0/2$. Using this and the result (C.3) above, we can bound

$$\begin{aligned} \|(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} + \bar{\lambda}\boldsymbol{I})^{-1} - (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\| &\leq \frac{2B\varepsilon + |\lambda - \bar{\lambda}|}{\lambda_0(\lambda_0 - (2B\varepsilon + |\lambda - \bar{\lambda}|))} && \text{(C.5)} \\ &\leq \frac{\bar{B}}{\lambda_0(\lambda_0 - 2B\varepsilon)} && \text{(C.6)} \\ &\leq \frac{2\bar{B}}{\lambda_0^2} && \text{(C.7)} \end{aligned}$$

We then find

$$\begin{aligned} \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\ell_2} &\leq \|(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} + \lambda\boldsymbol{I})^{-1}\bar{\boldsymbol{X}}^T - (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\|\|\boldsymbol{y}\|_{\ell_2} \\ &\leq (\|(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{P}^T\| + \|(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} + \lambda\boldsymbol{I})^{-1} - (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\|\|\boldsymbol{X}\|)\|\boldsymbol{y}\|_{\ell_2} \\ &\leq (\varepsilon\lambda_0^{-1} + \frac{2B\bar{B}}{\lambda_0^2})\|\boldsymbol{y}\|_{\ell_2} \\ &\leq \frac{5B^2\|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2}\varepsilon + \frac{2B\|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2}|\lambda - \bar{\lambda}|. \end{aligned}$$

Similarly when $\lambda = 0$ and $n < p$, using $\boldsymbol{X}\boldsymbol{X}^T \succeq \lambda_0\boldsymbol{I}$

$$\begin{aligned} \|(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T)^{-1} - (\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T)^{-1}\| &\leq \frac{2B\varepsilon}{\sigma_{\min}^2(\boldsymbol{X})(\sigma_{\min}^2(\boldsymbol{X}) - \|\boldsymbol{X}\boldsymbol{X}^T - \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\|)} \\ &\leq \frac{4B\varepsilon}{\lambda_0^2} \end{aligned}$$

Thus, in an essentially identical fashion, we find

$$\begin{aligned} \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\ell_2} &\leq \|\bar{\boldsymbol{X}}^T(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T)^{-1} - \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\|\|\boldsymbol{y}\|_{\ell_2} \\ &\leq (\|(\bar{\boldsymbol{X}} - \boldsymbol{X})^T(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T)^{-1}\| + \|\boldsymbol{X}^T\|\|(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T)^{-1} - (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\|)\|\boldsymbol{y}\|_{\ell_2} \\ &\leq \frac{5B^2\|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2}\varepsilon. \end{aligned}$$

This finishes the proof of $\frac{5B^2\|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2}$ Lipschitzness.

Finally, let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ be the singular value decomposition with ith singular value σ_i. Suppose $\lambda = 0$, $\sigma_i^2 \geq \lambda_0$ and $\bar{\lambda} \neq 0$. We consider the gap

$$\begin{aligned}\|(\boldsymbol{X}^T\boldsymbol{X} + \bar{\lambda}\boldsymbol{I})^{-1}\boldsymbol{X}^T - \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\| &= \|\boldsymbol{V}(\boldsymbol{\Lambda}^2 + \bar{\lambda}\boldsymbol{I})^{-1}\boldsymbol{\Lambda}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\| \\ &\leq \sup_{1\leq i\leq n} |\frac{\lambda_i}{\lambda_i^2 + \bar{\lambda}} - \frac{\lambda_i}{\lambda_i^2}| \leq \frac{\bar{\lambda}}{\lambda_0^2}.\end{aligned}$$

Thus, using this bound and the triangle inequality, for the scenario $\lambda = 0$ and $\bar{\lambda} \neq 0$, we again find the desired bound. ■

C.4 Lipschitzness of Feature Maps

Lemma 12 (Lipschitzness of Feature Map Solutions) *Let* $\sup_{\boldsymbol{x}\in\mathcal{X}, 1\leq i\leq h} \|\phi_i(\boldsymbol{x})\|_{\ell_2}^2 \leq B$. *Let* $\lambda \geq 0$ *be the strength of regularization. Suppose* $n \leq p$ *and* $\lambda + \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \sigma_{\min}^2(\boldsymbol{\Phi}_{\boldsymbol{\alpha}}) \geq \lambda_0 > 0$. *Set* $R = \sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \|\boldsymbol{\alpha}\|_{\ell_1}$. *Suppose* $\boldsymbol{\Delta}$ *is convex and the Algorithm* $\mathcal{A}$ *solves the ridge regression given* $\boldsymbol{\Phi}_{\boldsymbol{\alpha}}$ *i.e.*

$$\boldsymbol{\theta}_{\boldsymbol{\alpha}} = \mathcal{A}_{\textit{ridge}}(\boldsymbol{y}, \boldsymbol{\Phi}_{\boldsymbol{\alpha}}, \lambda).$$

Then, $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ *is* $5R^2\sqrt{B^3 n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}$*-Lipschitz function of* $\boldsymbol{\alpha}$ *(in* ℓ_2 *norm) and Assumption 1 holds with* $L = 6R^3B^2\sqrt{n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}$.

Additionally, consider $\boldsymbol{\theta}_{\boldsymbol{\alpha}} = \mathcal{A}_{\textit{ridge}}(\boldsymbol{y}, \boldsymbol{\Phi}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_{h+1})$ *where* $\boldsymbol{\alpha}_{h+1} \in [\lambda, \lambda_{\max}]$ *where* $\lambda + \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \sigma_{\min}^2(\boldsymbol{\Phi}_{\boldsymbol{\alpha}}) \geq \lambda_0$. *Then Assumption 1 holds with*

$$L = 6R^3B^2\sqrt{n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2} + 2R^2B\sqrt{n_{\mathcal{T}}}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}. \tag{C.8}$$

Proof Observe that $\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{x})\|_{\ell_2} \leq R\sqrt{B}$ which also implies $\|\boldsymbol{\Phi}\| \leq \bar{B} = R\sqrt{Bn_{\mathcal{T}}}$. Fix $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \boldsymbol{\Delta}$ satisfying $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2} = \varepsilon < \frac{\lambda_0}{4RBn_{\mathcal{T}}}$ where $\varepsilon > 0$ is to be determined. This implies that

$$\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \phi_{\boldsymbol{\alpha}'}(\boldsymbol{x})\|_{\ell_2} = \|\sum_{i=1}^{h}(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i')\phi_i(\boldsymbol{x})\|_{\ell_2} \leq \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_1}\sqrt{B} \tag{C.9}$$

$$\implies \|\boldsymbol{\Phi}_{\boldsymbol{\alpha}} - \boldsymbol{\Phi}_{\boldsymbol{\alpha}'}\| \leq \sqrt{n_{\mathcal{T}}}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_1}\sqrt{B} \leq \sqrt{n_{\mathcal{T}}h}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}\sqrt{B} \leq \varepsilon\sqrt{n_{\mathcal{T}}hB}. \tag{C.10}$$

Using the fact that problem is λ_0-strongly convex, applying Lemma 11 with λ_0, and observing that the initial choice of ε implies $\|\boldsymbol{\Phi}_{\boldsymbol{\alpha}} - \boldsymbol{\Phi}_{\boldsymbol{\alpha}'}\| \leq \varepsilon\sqrt{n_{\mathcal{T}}hB} < \frac{\lambda_0}{4R\sqrt{Bn_{\mathcal{T}}}}$, we find that

$$\frac{\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2}}{\|\boldsymbol{\Phi}_{\boldsymbol{\alpha}} - \boldsymbol{\Phi}_{\boldsymbol{\alpha}'}\|} \leq 5\bar{B}^2\|\boldsymbol{y}\|_{\ell_2}/\lambda_0^2 = 5R^2Bn_{\mathcal{T}}\|\boldsymbol{y}\|_{\ell_2}/\lambda_0^2.$$

This also implies the Lipschitzness $\frac{\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2}}{\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}} \leq \bar{L} := 5R^2\sqrt{B^3 n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}$. Observe that this argument showed the desired Lipschitzness around $\boldsymbol{\alpha}$ in a ball of radius $\varepsilon > 0$. However, this same local Lipschitzness holds for any choice of $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$. Since $\boldsymbol{\Delta}$ is convex, local Lipschitzness implies global (as we can draw a straightline between any two points and repeatedly apply local Lipschitzness).

Let $\boldsymbol{\theta}_{\boldsymbol{\alpha}}$ be $\bar{L}$ Lipschitz function of $\boldsymbol{\alpha}$ (which is provided above). To conclude with Assumption 1, we need Lipschitzness over the input space. Observe that $\|\boldsymbol{\theta}_{\boldsymbol{\alpha}}\|_{\ell_2} \leq \lambda_0^{-1/2}\|\boldsymbol{y}\|_{\ell_2}$. Consequently, using the fact $R^2B(n_{\mathcal{T}}-1) \geq \lambda$ (which implies $R^2Bn_{\mathcal{T}} \geq \lambda_0$), we find

$$\begin{aligned}
|f_{\boldsymbol{\alpha}}(\boldsymbol{x}) - f_{\boldsymbol{\alpha}'}(\boldsymbol{x})| &\leq |\boldsymbol{\theta}_{\boldsymbol{\alpha}}^T\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}^T\phi_{\boldsymbol{\alpha}'}(\boldsymbol{x})| \\
&\leq R\sqrt{B}\bar{L}\|\boldsymbol{\alpha}-\boldsymbol{\alpha}'\|_{\ell_2} + \|\boldsymbol{\theta}_{\boldsymbol{\alpha}}\|_{\ell_2}\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \phi_{\boldsymbol{\alpha}'}(\boldsymbol{x})\|_{\ell_2} \\
&\leq (R\sqrt{B}\bar{L} + \|\boldsymbol{y}\|_{\ell_2}\lambda_0^{-1/2}\sqrt{hB})\|\boldsymbol{\alpha}-\boldsymbol{\alpha}'\|_{\ell_2} \\
&\leq \sqrt{B}(R\bar{L} + \|\boldsymbol{y}\|_{\ell_2}\lambda_0^{-1/2}\sqrt{h})\|\boldsymbol{\alpha}-\boldsymbol{\alpha}'\|_{\ell_2} \\
&\leq 6R^3B^2\sqrt{n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}\|\boldsymbol{\alpha}-\boldsymbol{\alpha}'\|_{\ell_2},
\end{aligned}$$

concluding the first result.

To show the second point, using the exact same argument and applying (C.4), we find

$$\frac{\|\boldsymbol{\theta}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}'}\|_{\ell_2}}{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}'\|_{\ell_2}} \leq 5R^2\sqrt{B^3n_{\mathcal{T}}^3 h}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2} + 2R\sqrt{Bn_{\mathcal{T}}}\lambda_0^{-2}\|\boldsymbol{y}\|_{\ell_2}.$$

Repeating the input space argument, we find the advertised bound. ∎

D Proofs for Neural Feature Maps (Theorem 4)

The following theorem is a restatement of Theorem 4. The main difference is that conditions (D.1) on $\bar{h}_{\text{eff}}, k_\star$ are more precise versions compared to Theorem 4.

Theorem 8 (Restatement of Theorem 4) *Suppose Assumptions 4 and 5 hold and we solve feature map regression (Def. 2) with neural feature maps (4.4) upper bounded by $\sqrt{B}$ in ℓ_2-norm. Define the set*

$$\boldsymbol{\Delta}_0 = \{\boldsymbol{\alpha} \in \boldsymbol{\Delta} \;\big|\; \boldsymbol{K}_{\boldsymbol{\alpha}} \geq \lambda_0\}.$$

Let loss ℓ be Γ-Lipschitz and bounded by a constant. Suppose the following conditions on $k_\star, \lambda$ hold and define $\bar{h}_{\mathit{eff}}$ as

$$\begin{aligned}
&\bar{h}_{\mathit{eff}} = h_{\mathit{eff}}\log(4\bar{C}L(\lambda^{-2}B\Gamma + \sqrt{B})n_{\mathcal{T}}^3) \quad , \\
&k_\star \gtrsim \varepsilon^{-4}\lambda_0^{-4}\Gamma^4B^2\nu(h_{\mathit{eff}}\log(2L\bar{C}k_\star/(\nu h_{\mathit{eff}})) + t) \quad , \\
&\lambda \leq \frac{\varepsilon\lambda_0^2}{4\sqrt{Bn_{\mathcal{T}}}}.
\end{aligned} \tag{D.1}$$

Then, with probability at least $1-5e^{-t}$, for some constant $C>0$

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}_0} 2\Gamma\sqrt{\frac{B\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}}{n_{\mathcal{T}}}} + C\sqrt{\frac{\bar{h}_{\mathit{eff}}+\tau}{n_{\mathcal{T}}}} + \varepsilon + \delta. \tag{D.2}$$

Finally, this result also applies to the $0-1$ loss $\mathcal{L}^{0-1}$ by setting $\Gamma=1$. To see this, choose ℓ to be the Hinge loss and note that it dominates the $0-1$ loss.

Proof For a new input example $\boldsymbol{x}$, define the vector $\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x}) = k_{\boldsymbol{\alpha}}(\boldsymbol{x}_i, \boldsymbol{x})$ where $k_{\boldsymbol{\alpha}}$ is the empirical kernel function associated with neural feature map at initialization. By assumption $\kappa_{\boldsymbol{\alpha}}$ is L-Lipschitz in ℓ_2 norm.

First we settle a couple of bounds.

1. Set $\varepsilon_0 \leq \lambda_0/2$ and note that using Assumption 5 we have

$$k_\star \geq 4\varepsilon_0^{-2}\nu(h_{\text{eff}}\log(2L\bar{C}k_\star/(\nu h_{\text{eff}})) + t). \tag{D.3}$$

   With probability at least $1 - e^{-t}$, for all $\boldsymbol{\alpha}$, $\|\boldsymbol{K}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}\| \leq \varepsilon_0$.

2. Let $f^\lambda_{\boldsymbol{\alpha}}$ be the solution with ridge regularization $0 < \lambda \leq \lambda_0/2$ and $f^0_{\boldsymbol{\alpha}}$ be the ridgeless solution. For any input $\boldsymbol{x} \in \mathcal{X}$, using Lemma 10, we have that

$$\begin{aligned}
|f^\lambda_{\boldsymbol{\alpha}}(\boldsymbol{x}) - f^0_{\boldsymbol{\alpha}}(\boldsymbol{x})| &= |\boldsymbol{y}\boldsymbol{K}^{-1}_{\boldsymbol{\alpha}}\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{y}(\boldsymbol{K}_{\boldsymbol{\alpha}} + \lambda\boldsymbol{I})^{-1}\kappa_{\boldsymbol{\alpha}'}(\boldsymbol{x})| && \text{(D.4)}\\
&\leq \|\boldsymbol{y}\|_{\ell_2}\|\boldsymbol{K}^{-1}_{\boldsymbol{\alpha}} - (\boldsymbol{K}_{\boldsymbol{\alpha}} + \lambda\boldsymbol{I})^{-1}\|\|\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x})\|_{\ell_2} && \text{(D.5)}\\
&\leq \frac{2\sqrt{Bn}\lambda}{\lambda_0^2}. && \text{(D.6)}
\end{aligned}$$

3. Next, we apply Theorem 1. First note that (for sufficiently small $\boldsymbol{\alpha} - \boldsymbol{\alpha}'$)

$$\|(\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} + \lambda\boldsymbol{I})^{-1} - (\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'} + \lambda\boldsymbol{I})^{-1}\| \leq \frac{L}{\lambda^2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}.$$

   The Lipschitz constant of $\mathcal{A}$ can be found via

$$\begin{aligned}
|\boldsymbol{y}(\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} + \lambda\boldsymbol{I})^{-1}\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{y}(\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'} + \lambda\boldsymbol{I})^{-1}\kappa_{\boldsymbol{\alpha}'}(\boldsymbol{x})| &\leq \|\boldsymbol{y}\|_{\ell_2}\|(\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} + \lambda\boldsymbol{I})^{-1}\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x}) - (\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'} + \lambda\boldsymbol{I})^{-1}\kappa_{\boldsymbol{\alpha}'}(\boldsymbol{x})\|_{\ell_2}\\
&\leq \|\boldsymbol{y}\|_{\ell_2}\|(\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} + \lambda\boldsymbol{I})^{-1} - (\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'} + \lambda\boldsymbol{I})^{-1}\|\|\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x})\|_{\ell_2} + \|\widehat{\boldsymbol{K}}^{-1}_{\boldsymbol{\alpha}'}\|\|\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \kappa_{\boldsymbol{\alpha}'}(\boldsymbol{x})\|_{\ell_2}\\
&\leq (\frac{2L\|\boldsymbol{y}\|_{\ell_2}}{\lambda^2}\|\kappa_{\boldsymbol{\alpha}}(\boldsymbol{x})\|_{\ell_2} + \frac{L\|\boldsymbol{y}\|_{\ell_2}}{\lambda})\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}\\
&\leq \frac{2LB\|\boldsymbol{y}\|_{\ell_2}\sqrt{n_{\mathcal{T}}}}{\lambda^2}\varepsilon\\
&\leq \frac{2LBn_{\mathcal{T}}}{\lambda^2}\varepsilon
\end{aligned}$$

   Thus Theorem 1 yields with probability $1 - 2e^{-t}$

$$\mathcal{L}(f^{\mathcal{T}}_{\widehat{\boldsymbol{\alpha}}}) \leq \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \mathcal{L}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) + C\sqrt{\frac{h_{\text{eff}}\log(4\lambda^{-2}\bar{C}LB\Gamma n_{\mathcal{T}}n_{\mathcal{V}}) + \tau}{n_{\mathcal{V}}}} + \delta. \tag{D.7}$$

4. Finally, applying Theorem 1 on the ridgeless kernel estimators (trained with $\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}$) $f'_{\boldsymbol{\alpha}}$ over the set $\boldsymbol{\Delta}_0$, we obtain that for each $\boldsymbol{\alpha} \in \boldsymbol{\Delta}_0$

$$\mathcal{L}(f'_{\boldsymbol{\alpha}}) \leq \frac{2\Gamma\sqrt{B\boldsymbol{y}^T\widehat{\boldsymbol{K}}^{-1}_{\boldsymbol{\alpha}}\boldsymbol{y}} + C\sqrt{t}}{\sqrt{n_{\mathcal{T}}}}.$$

Observe that for $\boldsymbol{\alpha} \in \boldsymbol{\Delta}_0$

$$|\sqrt{\boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}} - \sqrt{\boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'}^{-1} \boldsymbol{y}}| \le \frac{|\boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y} - \boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'}^{-1} \boldsymbol{y}|}{\sqrt{\boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}}} \le \sqrt{Bn_{\mathcal{T}}} \frac{|\boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y} - \boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'}^{-1} \boldsymbol{y}|}{\|\boldsymbol{y}\|_{\ell_2}} \tag{D.8}$$

$$\le \sqrt{Bn_{\mathcal{T}}} \|\boldsymbol{y}\|_{\ell_2} \|\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'}^{-1}\| \tag{D.9}$$

$$\le \frac{2\sqrt{Bn_{\mathcal{T}}} L \|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2} \sqrt{\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2}} \tag{D.10}$$

Same argument also gives

$$|\sqrt{\boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}} - \sqrt{\boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}}| \le \sqrt{\boldsymbol{y}^T (\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} - \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}) \boldsymbol{y}} \le \frac{\|\boldsymbol{y}\|_{\ell_2} \sqrt{2\varepsilon_0}}{\lambda_0}. \tag{D.11}$$

Applying Lemma 1 with $\kappa = \frac{2\sqrt{B} n_{\mathcal{T}} \|\boldsymbol{y}\|_{\ell_2}}{\lambda_0^2}$ and using (D.11), we find with probability $1 - 2e^{-t}$ that

$$\min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \mathcal{L}(f'_{\boldsymbol{\alpha}}) \le \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}_0} \frac{2\Gamma \sqrt{B \boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}} + C \sqrt{h_{\text{eff}} \log(2\bar{C}\sqrt{B} n_{\mathcal{T}}^3 L) + \tau}}{\sqrt{n_{\mathcal{T}}}} \tag{D.12}$$

$$\le \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}_0} \frac{2\Gamma \sqrt{B \boldsymbol{y}^T \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}} + C \sqrt{h_{\text{eff}} \log(2\bar{C}\sqrt{B} n_{\mathcal{T}}^3 L) + \tau}}{\sqrt{n_{\mathcal{T}}}} + 3\Gamma \frac{\sqrt{B\varepsilon_0}}{\lambda_0}. \tag{D.13}$$

To stitch the results together, observe that with probability $1 - 5e^{-t}$ all events hold. Then, using (D.6) and right above, we obtain

$$\min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \mathcal{L}(f_{\boldsymbol{\alpha}}) \le \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \mathcal{L}(f'_{\boldsymbol{\alpha}}) + \frac{2\sqrt{Bn_{\mathcal{T}}}\lambda}{\lambda_0^2}$$

$$\le \min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}_0} \frac{2\Gamma \sqrt{B \boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}} + C \sqrt{h_{\text{eff}} \log(2\bar{C}\sqrt{B} n_{\mathcal{T}}^3 L) + \tau}}{\sqrt{n_{\mathcal{T}}}} + 3\Gamma \frac{\sqrt{B\varepsilon_0}}{\lambda_0} + \frac{2\sqrt{Bn_{\mathcal{T}}}\lambda}{\lambda_0^2}. \tag{D.14}$$

Now fix $\varepsilon > 0$ and set $\lambda \le \frac{\varepsilon \lambda_0^2}{4\sqrt{Bn_{\mathcal{T}}}}$ and $\varepsilon_0 = \frac{\varepsilon^2 \lambda_0^2}{36\Gamma^2 B}$. The sum $3\Gamma \frac{\sqrt{B\varepsilon_0}}{\lambda_0} + \frac{2\sqrt{Bn_{\mathcal{T}}}\lambda}{\lambda_0^2}$ on the right hand-side of (D.14) is upper bounded by ε as soon as the conditions (D.1) hold (recall (D.3) for $k_\star$ bound). Under these conditions, combining (D.14) with the validation bound (D.7), yields the following upper bound

$$\min_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}_0} \frac{2\Gamma \sqrt{B \boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y}} + C \sqrt{h_{\text{eff}} \log(2\bar{C}\sqrt{B} n_{\mathcal{T}}^3 L) + \tau}}{\sqrt{n_{\mathcal{T}}}} + C \sqrt{\frac{h_{\text{eff}} \log(4\lambda^{-2} \bar{C} L B \Gamma n_{\mathcal{T}} n_{\mathcal{V}}) + \tau}{n_{\mathcal{V}}}} + \delta + \varepsilon.$$

This is equivalent to (D.2) after plugging in the definition of $\bar{h}_{\text{eff}}$ and recalling $n_{\mathcal{T}} \ge n_{\mathcal{V}}$. ■

D.1 Uniform Convergence to Population Neural Kernel

Lemma 13 (Uniform Convergence to NTK) *Suppose Assumptions 4 and 5 hold. With probability at least* $1 - e^{-t}$ *over initialization, for all* $\boldsymbol{\alpha}$, $\|\boldsymbol{K}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}\| \le 2\sqrt{\frac{\nu h_{\text{eff}} \log(2L\bar{C}k_\star/(\nu h_{\text{eff}})) + t}{k_\star}}$.

Proof Fix $\boldsymbol{\alpha}, \boldsymbol{\alpha}'$ such that $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_{\ell_2} = \varepsilon/2L$ and $\mathcal{C}_\varepsilon$ be the cover set of size $\log|\mathcal{C}_\varepsilon| \leq h_{\text{eff}} \log(2L\bar{C}/\varepsilon)$. We find that $\|\boldsymbol{K}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}\| \leq \varepsilon/2$. Now, using the concentration bound (5.1) and setting $\tau = h_{\text{eff}} \log(2L\bar{C}/\varepsilon) + t$, we obtain that with probability $1 - e^{-t}$, all $\boldsymbol{\alpha} \in \mathcal{C}_\varepsilon$ obeys $\|\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} - \boldsymbol{K}_{\boldsymbol{\alpha}}\| \leq \sqrt{\frac{\nu h_{\text{eff}} \log(2L\bar{C}/\varepsilon) + \nu t}{k_\star}}$. For $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, picking a neighbor $\boldsymbol{\alpha}' \in \mathcal{C}_\varepsilon$, via triangle inequality, we find

$$\begin{aligned}
&\|\boldsymbol{K}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}\| \\
&\leq \|\boldsymbol{K}_{\boldsymbol{\alpha}} - \boldsymbol{K}_{\boldsymbol{\alpha}'}\| + \|\boldsymbol{K}_{\boldsymbol{\alpha}'} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'}\| + \|\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}'} - \widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}\| \\
&\leq \sqrt{\frac{\nu h_{\text{eff}} \log(2L\bar{C}/\varepsilon) + \nu t}{k_\star}} + \varepsilon.
\end{aligned}$$

Setting $\varepsilon = \sqrt{\nu h_{\text{eff}}/k_\star}$, we conclude with the result. ∎

E Proof of Lemma 3 and Structure of the Jacobian of a Deep Neural Net

E.1 Expression for Jacobian

In this section we wish to bound the Lipschitzness of the Jacobian matrix with respect to the architecture parameters $\boldsymbol{\alpha}$. To this aim we begin by considering the structure of the Jacobian. Specifically, assume we have n data points $\boldsymbol{x}_i \in \mathbb{R}^d$ for $i = 1, 2, \ldots, n$. We shall use

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$$

for the data matrix.

Set $\boldsymbol{v} := \boldsymbol{W}^{(D+1)}$. Assume a neural network mapping $f : \mathbb{R}^d \to \mathbb{R}$ with D hidden layers of the form

$$f(\boldsymbol{v}, \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(D)}, \boldsymbol{x}) = \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}^{(D)}} \left(\boldsymbol{W}^{(D)} \sigma_{\boldsymbol{\alpha}^{(D-1)}} \left(\boldsymbol{W}^{(D-1)} \ldots \sigma_{\boldsymbol{\alpha}^{(1)}} \left(\boldsymbol{W}^{(1)} \boldsymbol{x} \right) \right) \right)$$

with $\boldsymbol{W}^{(r)} \in \mathbb{R}^{d_r \times d_{r-1}}$ where $d_0 = d$ and $d_D = k$. We also define the hidden unit vector

$$\boldsymbol{h}^{(1)}(\boldsymbol{x}) = \sigma_{\boldsymbol{\alpha}^{(1)}} \left(\boldsymbol{W}^{(1)} \boldsymbol{x} \right)$$

for the first layer and inductively for the remaining layers via

$$\boldsymbol{h}^{(r)}(\boldsymbol{x}) = \sigma_{\boldsymbol{\alpha}^{(r)}} \left(\boldsymbol{W}^{(r)} \boldsymbol{h}^{(r-1)}(\boldsymbol{x}) \right)$$

for $r = 2, 3, \ldots, D$.

The Jacobian with respect to the weights of the last layer is equal to

$$\mathcal{J}^T(\boldsymbol{v}) = \sigma_{\boldsymbol{\alpha}^{(D)}} \left(\boldsymbol{W}^{(D)} \sigma_{\boldsymbol{\alpha}^{(D-1)}} \left(\boldsymbol{W}^{(D-1)} \ldots \sigma_{\boldsymbol{\alpha}^{(1)}} \left(\boldsymbol{W}^{(1)} \boldsymbol{X}^T \right) \right) \right)$$

which can also be rewritten as

$$\mathcal{J}(\boldsymbol{v}) = \begin{bmatrix} \sigma_{\boldsymbol{\alpha}^{(D)}} \left(\boldsymbol{h}^{(D)}(\boldsymbol{x}_1)\right)^T \\ \sigma_{\boldsymbol{\alpha}^{(D)}} \left(\boldsymbol{h}^{(D)}(\boldsymbol{x}_2)\right)^T \\ \vdots \\ \sigma_{\boldsymbol{\alpha}^{(D)}} \left(\boldsymbol{h}^{(D)}(\boldsymbol{x}_n)\right)^T \end{bmatrix}$$

With respect to $\boldsymbol{W}^{(D)}$ the columns take the form

$$\mathcal{J}^T\left(\boldsymbol{W}^{(D)}\right) = \operatorname{diag}(\boldsymbol{v})\, \sigma'_{\boldsymbol{\alpha}^{(D)}}\left(\boldsymbol{W}^{(D)}\boldsymbol{X}\right) * \boldsymbol{h}^{(D)}\left(\boldsymbol{X}^T\right)$$

To calculate the derivative with respect to $\boldsymbol{W}^{(r)}$ first note that

$$f\left(\boldsymbol{v}, \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(D)}, \boldsymbol{x}\right) = \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}^{(D)}}\left(\boldsymbol{W}^{(D)} \sigma_{\boldsymbol{\alpha}^{(D-1)}}\left(\boldsymbol{W}^{(D-1)} \ldots \sigma_{\boldsymbol{\alpha}^{(1)}}\left(\boldsymbol{W}^{(r)}\boldsymbol{h}^{(r-1)}(\boldsymbol{x})\right)\right)\right).$$

Furthermore, note that

$$f\left(\boldsymbol{v}, \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(D)}, \boldsymbol{x}\right) = \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}^{(D)}}\left(\boldsymbol{W}^{(D)} \sigma_{\boldsymbol{\alpha}^{(D-1)}}\left(\boldsymbol{W}^{(D-1)} \ldots \sigma_{\boldsymbol{\alpha}^{(r)}}\left(g\left(\boldsymbol{W}^{(r)}\right)\right)\right)\right).$$

where $g\left(\boldsymbol{W}^{(r)}\right) = \boldsymbol{W}^{(r)} h^{(r-1)}(\boldsymbol{x})$. Using the chain rule we have

$$\begin{aligned} \mathcal{D}_f\left(\boldsymbol{W}^{(r)}\right) &= \left(\boldsymbol{v}^T \prod_{s=r+1}^{D} \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(s)}}\left(\boldsymbol{h}^{(s)}(\boldsymbol{x})\right)\right) \boldsymbol{W}^{(s)}\right) \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(r)}}\left(\boldsymbol{h}^{(r)}(\boldsymbol{x})\right)\right) \mathcal{D}_g\left(\boldsymbol{W}^{(r)}\right) \\ &= \left(\boldsymbol{v}^T \prod_{s=r+1}^{D} \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(s)}}\left(\boldsymbol{h}^{(s)}(\boldsymbol{x})\right)\right) \boldsymbol{W}^{(s)}\right) \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(r)}}\left(\boldsymbol{h}^{(r)}(\boldsymbol{x})\right)\right) \left(\boldsymbol{I}_{d^{(r)}} \otimes \left(h^{(r-1)}(\boldsymbol{x})\right)^T\right) \end{aligned}$$

Therefore,

$$\mathcal{J}\left(\boldsymbol{W}^{(r)}\right) = \begin{bmatrix} \left(\boldsymbol{v}^T \prod_{s=r+1}^{D} \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(s)}}\left(\boldsymbol{h}^{(s)}(\boldsymbol{x}_1)\right)\right) \boldsymbol{W}^{(s)}\right) \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(r)}}\left(\boldsymbol{h}^{(r)}(\boldsymbol{x}_1)\right)\right) \left(\boldsymbol{I}_{d^{(r)}} \otimes \left(\boldsymbol{h}^{(r-1)}(\boldsymbol{x}_1)\right)^T\right) \\ \left(\boldsymbol{v}^T \prod_{s=r+1}^{D} \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(s)}}\left(\boldsymbol{h}^{(s)}(\boldsymbol{x}_2)\right)\right) \boldsymbol{W}^{(s)}\right) \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(r)}}\left(\boldsymbol{h}^{(r)}(\boldsymbol{x}_2)\right)\right) \left(\boldsymbol{I}_{d^{(r)}} \otimes \left(\boldsymbol{h}^{(r-1)}(\boldsymbol{x}_2)\right)^T\right) \\ \vdots \\ \left(\boldsymbol{v}^T \prod_{s=r+1}^{D} \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(s)}}\left(\boldsymbol{h}^{(s)}(\boldsymbol{x}_n)\right)\right) \boldsymbol{W}^{(s)}\right) \operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}^{(r)}}\left(\boldsymbol{h}^{(r)}(\boldsymbol{x}_n)\right)\right) \left(\boldsymbol{I}_{d^{(r)}} \otimes \left(\boldsymbol{h}^{(r-1)}(\boldsymbol{x}_n)\right)^T\right) \end{bmatrix}$$

For simplicity we shall use the short-hand

$$\mathcal{J}^{(\ell)} := \mathcal{J}\left(\boldsymbol{W}^{(\ell)}\right),$$

with $\mathcal{J}^{(D+1)} = \mathcal{J}(\boldsymbol{v})$. We shall also use $\mathcal{J}_i^{(\ell)}$ to denote the i-th row of $\mathcal{J}^{(\ell)}$. Finally, define the overall Jacobian to be

$$\mathcal{J} = \boldsymbol{\Phi}_{\boldsymbol{\alpha}} = [\mathcal{J}^{(D+1)}\ \mathcal{J}^{(D)}\ \ldots\ \mathcal{J}^{(1)}] \in \mathbb{R}^{n \times p}$$

Additionally, define the gram matrix as $\widehat{\boldsymbol{K}} = \mathcal{J}\mathcal{J}^T = \boldsymbol{\Phi}_{\boldsymbol{\alpha}}\boldsymbol{\Phi}_{\boldsymbol{\alpha}}^T$. For the discussion below set $S_\ell = \|\boldsymbol{W}^{(\ell)}\|$, define the quantities

$$M = \prod_{\ell=1}^{D+1} S_\ell, \quad M_+^i = \prod_{\ell=1}^{i} S_\ell, \quad M_-^i = \prod_{\ell=i}^{D+1} S_\ell. \tag{E.1}$$

Also suppose $\|\boldsymbol{x}\|_{\ell_2} \leq N$ for some $N \geq \sqrt{d}$ for all feasible inputs $\boldsymbol{x} \in \mathcal{X}$. Observe that, if the activations are zero-mean using $|\phi'_{\boldsymbol{\alpha}}| \leq B$,

$$\|\boldsymbol{h}^{(\ell)}\|_{\ell_2} \leq B^\ell M_+^\ell N.$$

Define $k_0 = d$ and let k_ℓ be the width of the ℓth layer. Define

$$\bar{S}_i = S_i + \max(1, \sqrt{k_i/k_{i-1}}).$$

In general, since $|\phi_{\boldsymbol{\alpha}}(0)| \leq B$, we have that

$$\|\boldsymbol{h}^{(\ell)}\|_{\ell_2} \leq B^\ell \bar{M}_+^\ell N. \tag{E.2}$$

where

$$\bar{M} = \prod_{\ell=1}^{D+1} \bar{S}_\ell, \quad \bar{M}_+^\ell = \prod_{i=1}^{\ell} \bar{S}_i, \quad \bar{M}_-^\ell = \prod_{i=\ell}^{D+1} \bar{S}_i.$$

E.2 Upper Bounding Jacobian

Before Lipschitzness, let us upper bound the Jacobian spectral norm. Clearly given two activation parameters $\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}$, noting $\mathcal{J}$ is concatenation of $\mathcal{J}^{(\ell)}$'s, Jacobian's obey

$$\|\mathcal{J}_{\boldsymbol{\alpha}}\| \leq \sqrt{\sum_{\ell=1}^{D+1} \|\mathcal{J}_{\boldsymbol{\alpha}}^{(\ell)}\|^2} \leq \sqrt{2D} \max_{1 \leq \ell \leq D+1} \|\mathcal{J}_{\boldsymbol{\alpha}}^{(\ell)}\| \tag{E.3}$$

$$\|\mathcal{J}_{\boldsymbol{\alpha}} - \mathcal{J}_{\bar{\boldsymbol{\alpha}}}\| \leq \sqrt{D+1} \max_{1 \leq \ell \leq D+1} \|\mathcal{J}_{\boldsymbol{\alpha}}^{(\ell)} - \mathcal{J}_{\bar{\boldsymbol{\alpha}}}^{(\ell)}\| \tag{E.4}$$

$$\leq \sqrt{2Dn} \max_{1 \leq \ell, i \leq D+1} \|\mathcal{J}_{i,\boldsymbol{\alpha}}^{(\ell)} - \mathcal{J}_{i,\bar{\boldsymbol{\alpha}}}^{(\ell)}\|. \tag{E.5}$$

Additionally denoting $\mathcal{J} := \mathcal{J}_{\boldsymbol{\alpha}}$ observe that

$$\|\mathcal{J}^{(\ell)}\| \leq \sqrt{n} B^{D-\ell+1} M_-^{\ell+1} \times \|\boldsymbol{h}^{(\ell-1)}\|_{\ell_2} \leq \sqrt{n} B^{D-\ell+1} M_-^{\ell+1} \times B^{\ell-1} \bar{M}_+^{\ell-1} N \tag{E.6}$$

$$\leq \sqrt{n} B^D \bar{M} N. \tag{E.7}$$

Thus, we find that

$$\|\mathcal{J}\| \leq \sqrt{2Dn} B^D \bar{M} N. \tag{E.8}$$

E.3 Lipschitzness of the Gram Matrix

Now that we have an expression for the Jacobian matrices with respect to different weights we wish to bound the Lipshitzness with respect to $\boldsymbol{\alpha}$. Let $\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}$ be two activation choices. Observe via (E.8) that

$$\|\mathcal{J}_{\boldsymbol{\alpha}} \mathcal{J}_{\boldsymbol{\alpha}}^T - \mathcal{J}_{\bar{\boldsymbol{\alpha}}} \mathcal{J}_{\bar{\boldsymbol{\alpha}}}^T\| \leq 2\sqrt{2Dn} B^D \bar{M} N \|\mathcal{J}_{\boldsymbol{\alpha}} - \mathcal{J}_{\bar{\boldsymbol{\alpha}}}\|. \tag{E.9}$$

The right side will be bounded via (E.5). Thus, to proceed, we will consider the derivative of $\mathcal{J}_i^{(\ell)}$ (fixing $\boldsymbol{\alpha}$). Using the chain rule we have

$$\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{\alpha}} = \sum_{s=\ell-1}^{D} \frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)} \frac{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}}$$

Thus, using the triangular inequality

$$\left\|\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{\alpha}}\right\| \leq \sum_{s=\ell-1}^{D} \left\|\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}\right\| \left\|\frac{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}}\right\| \leq D \max_{\ell-1\leq s\leq D} \left\|\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}\right\| \left\|\frac{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}}\right\|. \tag{E.10}$$

Now note that based on the structure of the Jacobian discussed above and using the fact that the first and second order derivatives of the activations are bounded by $B \geq 1$ we have

$$\left\|\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{h}^{(\ell-1)}(\boldsymbol{x}_i)}\right\| \leq B^{D-\ell+1} \prod_{u=\ell+1}^{D+1} \left\|\boldsymbol{W}^{(u)}\right\| \leq B^D \bar{M} N, \qquad \text{for} \qquad s = \ell - 1 \tag{E.11}$$

$$\left\|\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}\right\| \leq B^{D-\ell+1} \left(\prod_{u=\ell+1}^{D+1} \left\|\boldsymbol{W}^{(u)}\right\|\right) \left\|\boldsymbol{h}^{(\ell-1)}(\boldsymbol{x}_i)\right\|_{\ell_2} \leq B^D \bar{M} N \qquad \text{for} \qquad s \geq \ell, \tag{E.12}$$

$$\left\|\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}\right\| = 0 \qquad \text{for} \qquad s < \ell - 1. \tag{E.13}$$

To proceed, we need to bound the remaining term $\|\frac{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}}\|$. We can write

$$\|\frac{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}}\| \leq \sqrt{\sum_{\ell=1}^{s} \|\frac{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}^{(\ell)}}\|^2} \tag{E.14}$$

$$\leq \sqrt{D} \max_{1\leq \ell \leq s} \|\frac{\partial \boldsymbol{h}^{(s)}(\boldsymbol{x}_i)}{\partial \boldsymbol{h}^{(\ell)}(\boldsymbol{x}_i)}\| \|\frac{\partial \boldsymbol{h}^{(\ell)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}^{(\ell)}}\| \tag{E.15}$$

$$\leq \sqrt{D} B^D \bar{M} N \sqrt{h}. \tag{E.16}$$

To see this, define $\bar{\boldsymbol{h}}^{(\ell)} = \boldsymbol{W}^{(\ell)} \boldsymbol{h}^{(\ell-1)}$ and note $\boldsymbol{h}^{(\ell)} = \sigma_{\boldsymbol{\alpha}^{(\ell)}}(\bar{\boldsymbol{h}}^{(\ell)})$. Now observe that $\frac{\partial \boldsymbol{h}^{(\ell)}(\boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}^{(\ell)}}$ has the following clean form, with columns that are bounded via (E.2)

$$\frac{\partial \boldsymbol{h}^{(\ell)}(\boldsymbol{x})}{\partial \boldsymbol{\alpha}^{(\ell)}} = \left[\phi_1(\bar{\boldsymbol{h}}^{(\ell)}) \ \ldots \ \phi_h(\bar{\boldsymbol{h}}^{(\ell)})\right] \in \mathbb{R}^{k_\ell \times h}.$$

Combining (E.13) with (E.16) and plugging in (E.10), we obtain

$$\left\|\frac{\partial \mathcal{J}_i^{(\ell)}}{\partial \boldsymbol{\alpha}}\right\| \leq D B^D \bar{M} N \times \sqrt{D} B^D \bar{M} N \sqrt{h} \leq D^{3/2} (B^D \bar{M} N)^2 \sqrt{h}. \tag{E.17}$$

E.4 Finalizing the Proof of Lemma 3

We now put the bounds above together. Overall, combining (E.9), (E.5), (E.17), we find that

$$\frac{\|\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\bar{\boldsymbol{\alpha}}}\|}{\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_{\ell_2}} \leq 2\sqrt{2Dn} B^D \bar{M} N \times \sqrt{2Dn} \times D^{3/2} (B^D \bar{M} N)^2 \sqrt{h} \leq 4n\sqrt{h} (D B^D \bar{M} N)^3.$$

This is summarized in the following lemma.

Lemma 14 (Gram matrix Lipschitzness) *Suppose ith layer has k_i neurons with $k_0 = d$ and $\boldsymbol{W}^{(i)} \in \mathbb{R}^{k_i \times k_{i-1}}$. Suppose ϕ_i obeys $|\phi_i(0)|, |\phi_i'(x)|, |\phi_i''(x)| \le B$. Suppose input features obey $\|\boldsymbol{x}\|_{\ell_2} \le N$ for some $N \ge \sqrt{d}$. Also let S_i be an upper bound on $\|\boldsymbol{W}_i\|$ and define $\bar{M} = \prod_{i=1}^{D+1}(S_i + \max(1, \sqrt{k_i/k_{i-1}}))$. Suppose $\boldsymbol{\alpha} = [\boldsymbol{\alpha}^{(1)} \ \ldots \ \boldsymbol{\alpha}^{(D)}]$ where $\boldsymbol{\alpha}^{(i)}$ governs ith layer activation and $\|\boldsymbol{\alpha}^{(i)}\|_{\ell_1} \le 1$. We have that*

$$\|\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\bar{\boldsymbol{\alpha}}}\| \le 4n\sqrt{h}(DB^D\bar{M}N)^3\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_{\ell_2}. \tag{E.18}$$

The following lemma is a restatement of Lemma 3 (i.e. its more precise version) and essentially follows from the deterministic bound of Lemma 14.

Lemma 15 *Suppose ith layer has k_i neurons with $k_0 = d$ and $\boldsymbol{W}^{(i)} \in \mathbb{R}^{k_i \times k_{i-1}}$. Suppose ϕ_i obeys $|\phi_i(0)|, |\phi_i'(x)|, |\phi_i''(x)| \le B$. Suppose input features are normalized to ensure $\|\boldsymbol{x}\|_{\ell_2} \lesssim \sqrt{d}$. Fix constants $\rho \ge 1$, $\bar{c} > 0$. Suppose the aspect ratios obey $k_i/k_{i-1} + 1 \le \rho$ for $2 \le i \le D$. Suppose layer i is initialized as $\mathcal{N}(0, c_i)$ for $1 \le i \le D+1$. Additionally suppose*

$$c_i \le \begin{cases} \bar{c} & \text{if} \quad i = 1 \\ \bar{c}/k_{i-1} & \text{if} \quad i \ge 2 \end{cases}.$$

Set $k_{\min} = \min_{1 \le i \le D} k_i$. Then, there exists a constant $C > 0$ such that, with probability $1 - De^{-10k_{\min}}$, for all $\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}} \in \boldsymbol{\Delta}$

$$\frac{\|\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} - \widehat{\boldsymbol{K}}_{\bar{\boldsymbol{\alpha}}}\|}{\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_{\ell_2}} \le (CB\sqrt{\bar{c}\rho})^{3D} n\sqrt{h} D^3 (k_1/d + 1)^{3/2} d^3.$$

Similarly $\boldsymbol{K}_{\boldsymbol{\alpha}} = \mathbb{E}[\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}]$ is also $(CB\sqrt{\bar{c}\rho})^{3D} n\sqrt{h} D^3 (k_1/d + 1)^{3/2} d^3$-Lipschitz function of $\boldsymbol{\alpha}$.

Proof We just need to plug in the proper quantities to Lemma 14. Let $C > 0$ be an absolute constant to be determined. Let ρ_i be the ith layer aspect ratio $k_i/k_{i-1} + 1$. Let $\bar{c}_i = c_i k_{i-1}$. Observe that, using Gaussian tail bound and Lemma 5, for all $D + 1 \ge i \ge 1$,

$$\mathbb{P}\left(\sqrt{\frac{k_{i-1}}{\bar{c}_i}}\|\boldsymbol{W}^{(i)}\| \ge \sqrt{k_i} + \sqrt{k_i - 1} + t\right) \le e^{-t^2/2} \tag{E.19}$$

$$\mathbb{E}[\|\boldsymbol{W}^{(i)}\|^3] \le 16\bar{c}_i^{3/2}\left(\frac{\sqrt{k_i} + \sqrt{k_i - 1} + 5}{\sqrt{k_{i-1}}}\right)^3 \le (C^2\bar{c}_i\rho_i)^{3/2}. \tag{E.20}$$

Define $\Gamma_i = C\sqrt{\bar{c}_i\rho_i}$. This also implies that, with probability at least $1 - (D+1)e^{-10k_{\min}}$ (union bound over all layers),

$$\|\boldsymbol{W}^{(i)}\| \le \Gamma_i, \quad \mathbb{E}[\|\boldsymbol{W}^{(i)}\|^3]^{1/3} \le \Gamma_i.$$

Since $\bar{S}_i \le \|\boldsymbol{W}^{(i)}\| + \sqrt{\rho_i}$, this also implies (after adjusting the constant C)

$$\bar{S}_i \le \Gamma_i, \quad \mathbb{E}[\bar{S}_i^3]^{1/3} \le \Gamma_i.$$

Finally, we simply need to calculate $\bar{M}$. Noticing $\bar{c}_i\rho_i \le \bar{c}\rho$ for $i \ge 2$ and $\bar{c}_1\rho_1 \le \bar{c}d(k_1 + d)/d = \bar{c}(k_1 + d)$, we find

$$\bar{M} = \prod_{\ell=1}^{D+1} \bar{S}_i \le \prod_{\ell=1}^{D+1} \Gamma_i \le (C^2\sqrt{\bar{c}\rho})^D \sqrt{k_1 + d} \tag{E.21}$$

$$\mathbb{E}[\bar{M}^3] \le (C\sqrt{\bar{c}\rho})^{3D}(k_1 + d)^{3/2}. \tag{E.22}$$

We conclude with the result after multiplying the remaining terms $n\sqrt{h}D^3B^{3D}N^3$ from (E.18) with $N = \sqrt{d}$. ∎

F Properties of the Spectral Estimator

The following establishes an asymptotic guarantee for spectral estimator when learning an overparameterized rank-1 matrix. Unlike Theorem 5, this result allows for label noise.

Theorem 9 (Guarantees for the spectral estimator) *Let $(\boldsymbol{X}_i)_{i=1}^n \subset \mathbb{R}^{h\times p}$ be i.i.d. matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Let $y_i = \boldsymbol{\alpha}^T\boldsymbol{X}_i\boldsymbol{\theta}+\sigma z_i$ for a unit norm $\boldsymbol{\alpha}\in\mathbb{R}^h, \boldsymbol{\theta}\in\mathbb{R}^p$ and suppose the noise is $(z_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Form the cross-moment matrix*

$$\hat{\boldsymbol{M}} = \frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{X}_i.$$

Let $\widehat{\boldsymbol{\alpha}}$ be the top left singular vector of $\hat{\boldsymbol{M}}$. Let $\bar{p} = p/n$ and $\bar{h} = h/n$. Let $1\geq\rho\geq 0$ be the absolute correlation between $\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}$ i.e. $\rho = |\boldsymbol{\alpha}^T\widehat{\boldsymbol{\alpha}}|$. In the large dimensional limit $p,n,h\to\infty$ (while keeping $\bar{p},\bar{h}$ constant in the limit), with probability 1, we have

$$\frac{\rho}{1-\rho^2} \geq \frac{(1+\sigma^2)^{-1}/\sqrt{\bar{h}} - (2\sqrt{\bar{p}}+\sqrt{\bar{h}})}{2(\sqrt{\bar{p}}+1)}. \tag{F.1}$$

Specifically, assuming $(1+\sigma^2)\sqrt{\bar{p}\bar{h}}\leq 1/6$, we find $\rho^2\geq 1-64(1+\sigma^2)^2\bar{p}\bar{h}$.

Proof When an inequality (e.g. $\leq$) holds in the large dimensional limit, with probability 1, we use the P-overset notation (e.g. $\overset{P}{\leq}$). During the proof, $\text{vec}(\cdot)$ denotes the vectorization of a matrix obtained by putting all columns on top of each other. $\text{mtx}(\cdot)$ denotes the inverse operation that constructs a matrix from a vector.

Let $\boldsymbol{x}_i = \text{vec}(\boldsymbol{X}_i)$. Let $g_i = \boldsymbol{\theta}^T\boldsymbol{x}_i$ and $h_i\sim\mathcal{N}(0,1)$ independent of others. Decompose $\boldsymbol{x}_i = \boldsymbol{x}_i' + (g_i-h_i)\boldsymbol{\theta}$ where $\boldsymbol{x}_i' = (\boldsymbol{I}-\boldsymbol{\theta}\boldsymbol{\theta}^T)\boldsymbol{x}_i + h_i\boldsymbol{\theta} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is independent of g_i.

$$\text{vec}(\hat{\boldsymbol{M}}) = \frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{\theta}+\sigma z_i) \tag{F.2}$$

$$= \frac{1}{n}[\sum_{i=1}^n \boldsymbol{x}_i'(\boldsymbol{x}_i^T\boldsymbol{\theta}+\sigma z_i) + g_i^2\boldsymbol{\theta} - g_ih_i\boldsymbol{\theta} + \sigma(g_i-h_i)z_i\boldsymbol{\theta}] \tag{F.3}$$

$$= \boldsymbol{\theta} + \text{rest}_1 + \frac{\boldsymbol{\gamma}}{\sqrt{n}} \quad \text{where} \quad \|\|\text{rest}_1\|_{\ell_2}\|_{\psi_1} \lesssim \frac{1+\sigma}{\sqrt{n}}. \tag{F.4}$$

where $\boldsymbol{\gamma}\sim\mathcal{N}(0,\nu)$ where $\nu = 1+\sigma^2$ and rest_1 is allowed to have an adversarial direction. Concretely, $\boldsymbol{\gamma}$ (approximately) equal to the $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i'(\boldsymbol{x}_i^T\boldsymbol{\theta}+\sigma z_i)$ term.

Following this, we can rewrite as

$$\text{vec}(\boldsymbol{M}) = \frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i = \text{vec}(\bar{\boldsymbol{M}}) + \text{rest}_1\boldsymbol{p},$$

where $\bar{\boldsymbol{M}}, \boldsymbol{G}\in\mathbb{R}^{h\times p}$, $\boldsymbol{G}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\nu) = \text{mtx}(\boldsymbol{\gamma})$ and

$$\bar{\boldsymbol{M}} = \boldsymbol{\alpha}\boldsymbol{\theta}^T + \frac{\boldsymbol{G}}{\sqrt{n}}.$$

Let $\boldsymbol{g} = \boldsymbol{G\theta} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \nu)$. Consider the covariance matrix

$$\bar{\boldsymbol{M}}\bar{\boldsymbol{M}}^T = \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \frac{1}{n}\boldsymbol{G}\boldsymbol{G}^T + \frac{1}{\sqrt{n}}(\boldsymbol{\alpha}\boldsymbol{g}^T + \boldsymbol{g}\boldsymbol{\alpha}^T).$$

Observe that

$$\|\boldsymbol{\alpha}^T\bar{\boldsymbol{M}}\|_{\ell_2}^2 = 1 + \frac{\|\boldsymbol{G}^T\boldsymbol{\alpha}\|_{\ell_2}^2}{n} + \frac{2\boldsymbol{g}^T\boldsymbol{\alpha}}{\sqrt{n}}.$$

Set $\bar{p} = p/n$ and $\bar{h} = h/n$. Set $\gamma = \sqrt{\bar{p}\bar{h}}$. This implies that

$$\|\boldsymbol{\alpha}^T\bar{\boldsymbol{M}}\|_{\ell_2}^2 = 1 + \frac{\nu p}{n} + \text{rest}_2 = 1 + \nu\bar{p} + \text{rest}_2 \quad \text{where} \quad \|\text{rest}_2\|_{\psi_1} \lesssim \frac{\sqrt{\nu}\sqrt{n+p}}{n} = \frac{\sqrt{\nu}\sqrt{1+\bar{p}}}{\sqrt{n}}.$$

Thus in the high-dimensional limit (n, p sufficiently large), contributions of $\text{rest}_1, \text{rest}_2$ disappears and

$$\|\boldsymbol{\alpha}^T\bar{\boldsymbol{M}}\|_{\ell_2}^2 \xrightarrow{P} 1 + \nu\bar{p}. \tag{F.5}$$

Now, let $\boldsymbol{e} = \text{top\_eigvec}(\boldsymbol{G}\boldsymbol{G}^T)$ and $\bar{\boldsymbol{g}} = \boldsymbol{g}/\|\boldsymbol{g}\|_{\ell_2}$. Observe that $\boldsymbol{e}$ is generated uniformly randomly over the range space of $\boldsymbol{G}$. Thus, the inner products $\bar{\boldsymbol{g}}^T\boldsymbol{e}, \bar{\boldsymbol{g}}^T\boldsymbol{\alpha}, \boldsymbol{e}^T\boldsymbol{\alpha}$ all have subgaussian norm at most $1/\sqrt{n}$ and become orthogonal in the high-dimensional limit. Also note that since $\boldsymbol{e}$ is an eigenvector, $\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{e}$ is again asymptotically orthogonal to $\boldsymbol{\alpha}, \bar{\boldsymbol{g}}$ (as it is perfectly parallel to $\boldsymbol{e}$ which is orthogonal to $\boldsymbol{\alpha}, \bar{\boldsymbol{g}}$). Top eigenvalue obeys the Bai-Yin law,

$$\frac{\boldsymbol{e}^T\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{e}}{n} \xrightarrow{P} \nu\frac{(\sqrt{p}+\sqrt{h})^2}{n} = \nu(\sqrt{\bar{p}}+\sqrt{\bar{h}})^2 \tag{F.6}$$

Since $\boldsymbol{\alpha}$ is fixed (and independent of $\boldsymbol{G}$), we have that

$$\boldsymbol{\alpha}^T\frac{\boldsymbol{G}\boldsymbol{G}^T}{n} = \nu\bar{p}\boldsymbol{\alpha} + \nu\sqrt{\bar{p}\bar{h}}\bar{\boldsymbol{h}} + \text{rest}_3 \quad \text{where} \quad \|\|\text{rest}_3\|_{\ell_2}\|_{\psi_1} \lesssim \nu\frac{\sqrt{\bar{p}}}{\sqrt{n}},$$

where $\bar{\boldsymbol{h}}$ term is distributed as $\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n)$. Let $\boldsymbol{a}$ be the top eigenvector of $\bar{\boldsymbol{M}}\bar{\boldsymbol{M}}^T$ and $\boldsymbol{a} = a\boldsymbol{\alpha} + b\boldsymbol{v}$ where $\boldsymbol{\alpha}, \boldsymbol{v}$ are two orthogonal unit vectors and $a^2 + b^2 = 1$. Note that $\bar{\boldsymbol{h}}$ is uniformly generated in the range space of $\boldsymbol{G}^T$. Thus $\|\boldsymbol{G}\bar{\boldsymbol{h}}\|_{\ell_2}^2/n \xrightarrow{P} \nu(\bar{p}+\bar{h})$. It is also orthogonal to $\boldsymbol{e}, \boldsymbol{a}, \bar{\boldsymbol{g}}$ in the limit. Consequently, (in the limit) we have

$$\begin{aligned}(a\boldsymbol{\alpha} + b\boldsymbol{v})^T\frac{\boldsymbol{G}\boldsymbol{G}^T}{n}(a\boldsymbol{\alpha} + b\boldsymbol{v}) &\overset{P}{\leq} \nu[\bar{p}a^2 + 2ab\sqrt{\bar{p}\bar{h}}|\bar{\boldsymbol{h}}^T\boldsymbol{v}| + b^2(\sqrt{\bar{p}}+\sqrt{\bar{h}})^2] &\text{(F.7)}\\ &\overset{P}{\leq} \nu[\bar{p} + 2|ab|\gamma + 2b^2\gamma + b^2\bar{h}] &\text{(F.8)}\end{aligned}$$

We additionally have the bound

$$|\frac{2\boldsymbol{a}^T\boldsymbol{g}\boldsymbol{\alpha}^T\boldsymbol{a}}{\sqrt{n}}| \overset{P}{\leq} 2\sqrt{\nu}|ab|\sqrt{\bar{h}}.$$

Combining, in large dimensional limit, we find that

$$\|\boldsymbol{a}^T\bar{\boldsymbol{M}}\|_{\ell_2}^2 \overset{P}{\leq} a^2 + \nu\bar{p} + 2\nu|ab|(\gamma + \sqrt{\bar{h}}) + \nu b^2(2\gamma + \bar{h}). \tag{F.9}$$

Since $\boldsymbol{a}$ is the top eigenvector, using bounds (F.5) and (F.9), we obtain via $\|\boldsymbol{a}^T\bar{\boldsymbol{M}}\|_{\ell_2}^2 \geq \|\boldsymbol{\alpha}^T\bar{\boldsymbol{M}}\|_{\ell_2}^2$ that

$$a^2 + \nu\bar{p} + 2\nu|ab|(\gamma + \sqrt{\bar{h}}) + \nu b^2(2\gamma + \bar{h}) \geq 1 + \nu\bar{p} \tag{F.10}$$

$$\implies \nu\bar{p} + 2\nu|ab|(\gamma + \sqrt{\bar{h}}) + \nu b^2(2\gamma + \bar{h}) \geq b^2 \tag{F.11}$$

$$\implies 2\nu|ab|(\gamma + \sqrt{\bar{h}}) \geq (1 - \nu(2\gamma + \bar{h}))b^2 \tag{F.12}$$

$$\implies 2\nu|a|(\gamma + \sqrt{\bar{h}}) \geq (1 - \nu(2\gamma + \bar{h}))|b| \tag{F.13}$$

$$\implies \frac{|a|}{|b|} \geq \frac{1 - \nu(2\gamma + \bar{h})}{2\nu(\gamma + \sqrt{\bar{h}})}. \tag{F.14}$$

After simplification, the line above is identical to (F.1). To proceed, using the fact that $\bar{p} \geq \gamma \geq \bar{h}$ and assuming $\nu\gamma \leq 1/6$ we find

$$\frac{|a|}{|b|} \geq \frac{1}{8\nu\gamma}.$$

This also yields $|a|^2 \geq 1/(1 + 64\nu^2\gamma^2) \geq 1 - 64\nu^2\gamma^2$. ∎

G Finalizing the Proof of Theorem 5

The proof is completed in two stages. First, we prove (6.3). The statement is directly implied by Theorem 9 by setting the noise level to zero. This gives us $\widehat{\boldsymbol{\alpha}}$ which is ρ (absolute) correlated with $\boldsymbol{\alpha}_\star$ where ρ obeys (6.3). For the second statement, we assume that validation training is complete, (6.3) and focus on the empirical risk minimization over training data to bound the risk of overparameterized learning with $\boldsymbol{\theta}_\star$. Below, without losing generality, we assume the correlation between $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_\star$ are nonnegative (as the situation is symmetric).

Observe that when $\widehat{\boldsymbol{\alpha}}$ is fixed (i.e. conditioned on the outcome of the spectral estimator), we define the feature matrix for the ERM (6.4). Specifically, define the matrix $\boldsymbol{\Phi}_{\widehat{\boldsymbol{\alpha}}} \in \mathbb{R}^{n\times p}$ where ith row of $\boldsymbol{\Phi}_{\boldsymbol{\alpha}}$ is given by $\hat{\boldsymbol{x}}_i = \boldsymbol{X}_i^T\widehat{\boldsymbol{\alpha}}$. Also define the ideal feature matrix $\boldsymbol{\Phi}_{\boldsymbol{\alpha}_\star}$ where the ith row is given by $\boldsymbol{x}_i = \boldsymbol{X}_i^T\boldsymbol{\alpha}_\star$. Observe that $\hat{\boldsymbol{x}}_i$ are essentially noisy features that contain a mixture of the right features (i.e. $\boldsymbol{x}_i$) as well as the wrong features that are induced by the component of $\widehat{\boldsymbol{\alpha}}$ orthogonal to $\boldsymbol{\alpha}_\star$.

Specifically, decompose $\widehat{\boldsymbol{\alpha}} = \rho\boldsymbol{\alpha}_\star + \sqrt{1-\rho^2}\bar{\boldsymbol{\alpha}}$ where $\bar{\boldsymbol{\alpha}}$ is also unit norm. Then set $\boldsymbol{z}_i = \boldsymbol{X}_i^T\bar{\boldsymbol{\alpha}}$ and observe that

$$\hat{\boldsymbol{x}}_i = \rho\boldsymbol{x}_i + \sqrt{1-\rho^2}\boldsymbol{x}_i.$$

The outcome of Theorem 9 upper bounds the magnitude of these wrong features (by lower bounding ρ). To proceed, we shall establish the exact asymptotic risk when fitting these noisy features. Given this discussion, the proof of the risk bound (6.4) is essentially established via the following lemma. The first statement applies for any value of ρ and the second statement chooses ρ induced by the spectral estimator to conclude with the proof of (6.4).

Lemma 16 (Asymptotic risk of regression with suboptimal feature map) *Fix $\boldsymbol{\theta}_\star \in \mathbb{R}^p$ and $(\boldsymbol{x}_i, \boldsymbol{z}_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$. Set $y_i = \boldsymbol{x}_i^T\boldsymbol{\theta}_\star$ and define the noisy features $\hat{\boldsymbol{x}}_i = \rho\boldsymbol{x}_i + \sqrt{1-\rho^2}\boldsymbol{z}_i$. Solve the problem (which is same as (6.2))*

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \quad \textit{where} \quad \mathcal{L}(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{\Phi}_{\widehat{\boldsymbol{\alpha}}}\boldsymbol{\theta}\|_{\ell_2}^2.$$

Consider the double asymptotic regime with $p, n \to \infty$ and $p/n \to \bar{p} > 1$. We have that

$$\lim_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[(y - \boldsymbol{x}^T\hat{\boldsymbol{\theta}})^2] = \frac{\bar{p}^2 - 2\bar{p}\rho + 2\rho - \rho^2}{\bar{p}(\bar{p}-1)}.$$

Specifically, assume $\bar{p}\bar{h} \le c \le 1/6$ for sufficiently small constant $c > 0$. Recalling $\rho^2 \ge 1 - 64\bar{h}\bar{p}$ as given by Theorem 9 (where we set $\sigma = 0$), we find

$$\lim_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}) \le 1 - \frac{1}{\bar{p}} + \frac{200\bar{h}}{1 - 1/\bar{p}}. \tag{G.1}$$

Proof We remark that related results/analysis exist in the literature (in the context of overparameterized high-dimensional learning and the properties of the min-norm interpolating solutions) [31]. Our strategy uses the results from [17].

Define the vector $\boldsymbol{a} = \sqrt{1-\rho^2}\boldsymbol{x} - \rho\boldsymbol{z}$ and note that $\hat{\boldsymbol{x}}, \boldsymbol{a}$ are independent. Additionally, note that

$$y = \boldsymbol{x}^T\boldsymbol{\theta}_\star = \rho\hat{\boldsymbol{x}}^T\boldsymbol{\theta}_\star + \sqrt{1-\rho^2}\boldsymbol{a}^T\boldsymbol{\theta}_\star.$$

Set $w = \boldsymbol{a}^T\boldsymbol{\theta}_\star \sim \mathcal{N}(0,1)$ which corresponds to the noise level of the problem. The original data in terms of the noisy features can be written as follows

$$y_i = \rho\hat{\boldsymbol{x}}_i^T\boldsymbol{\theta}_\star + \sqrt{1-\rho^2}w_i. \tag{G.2}$$

Define the asymptotic risk $\mathrm{risk}(\rho,\bar{p}) = \lim_{n\to\infty}\mathbb{E}[(y - \boldsymbol{x}^T\hat{\boldsymbol{\theta}})^2] = \mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star\|_{\ell_2}^2]$. Let $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I}_p/p)$. Let us introduce the random vector

$$\bar{\boldsymbol{\theta}} \sim \frac{\rho}{\bar{p}}\boldsymbol{\theta}_\star + \sqrt{\frac{1-\rho^2}{\bar{p}-1} + \frac{(\bar{p}-1)\rho^2}{\bar{p}^2}}\boldsymbol{h} = \frac{\rho}{\bar{p}}\boldsymbol{\theta}_\star + \sqrt{\frac{\bar{p}^2 - \rho^2\bar{p}^2 + \bar{p}^2\rho^2 - 2\bar{p}\rho^2 + \rho^2}{\bar{p}^2(\bar{p}-1)}}\boldsymbol{h} = \frac{\rho}{\bar{p}}\boldsymbol{\theta}_\star + \sqrt{\frac{\bar{p}^2 - 2\bar{p}\rho^2 + \rho^2}{\bar{p}^2(\bar{p}-1)}}\boldsymbol{h}.$$

Specializing the results of [17] to identity covariance shows that, in the double asymptotic overdetermined regime ($\bar{p} = p/n > 1$),

$$\begin{aligned}
\mathrm{risk}(\rho,\bar{p}) &= \lim_{n\to\infty}\mathbb{E}[\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star\|_{\ell_2}^2] = (1 - \frac{\rho}{\bar{p}})^2 + \frac{\bar{p}^2 - 2\bar{p}\rho^2 + \rho^2}{\bar{p}^2(\bar{p}-1)} && \text{(G.3)}\\
&= \frac{\bar{p}^3 - 2\bar{p}^2\rho + \rho^2\bar{p} - \bar{p}^2 + 2\bar{p}\rho - \rho^2}{\bar{p}^2(\bar{p}-1)} + \frac{\bar{p}^2 - 2\bar{p}\rho^2 + \rho^2}{\bar{p}^2(\bar{p}-1)} && \text{(G.4)}\\
&= \frac{\bar{p}^2 - 2\bar{p}\rho + \rho^2 - \bar{p} + 2\rho}{\bar{p}(\bar{p}-1)} + \frac{\bar{p} - 2\rho^2}{\bar{p}(\bar{p}-1)} && \text{(G.5)}\\
&= \frac{\bar{p}^2 - 2\bar{p}\rho + 2\rho - \rho^2}{\bar{p}(\bar{p}-1)}. && \text{(G.6)}
\end{aligned}$$

This proves the first statement. Observe that in the special case of $\rho = 1$, the risk reduces to $\mathrm{risk}(\boldsymbol{\alpha}_\star) := \mathrm{risk}(\rho,\bar{p}) = \frac{\bar{p}^2 - 2\bar{p} + 1}{\bar{p}(\bar{p}-1)} = \frac{\bar{p}-1}{\bar{p}} = 1 - 1/\bar{p}$.

To bound the risk on $\widehat{\boldsymbol{\alpha}}$, we study the derivative at $\rho = 1$. Note that

$$\frac{\partial\mathrm{risk}(\rho,\bar{p})}{\partial\rho} = \frac{2 - 2\rho - 2\bar{p}}{\bar{p}(\bar{p}-1)} \implies \frac{\partial\mathrm{risk}(\rho,\bar{p})}{\partial\rho}\Big|_{\rho=1} = \frac{-2\bar{p}}{\bar{p}(\bar{p}-1)} = -\frac{2}{\bar{p}-1}.$$

This implies that if $1-\varepsilon \le \rho \le 1$ for sufficiently small $\varepsilon > 0$, we have that

$$\text{risk}(\rho, \bar{p}) \ge \text{risk}(1, \bar{p}) + \frac{3(1-\rho)}{\bar{p}-1}$$

Since $\bar{h}\bar{p} \le c$ by choosing c sufficiently small, we can ensure $\rho \ge \rho^2 \ge 1 - 64\bar{p}\bar{h} \ge 1-\varepsilon$. Plugging in ρ lower bound above yields

$$\text{risk}(\rho, \bar{p}) \ge \text{risk}(1, \bar{p}) + \frac{3(1-64\bar{p}\bar{h})}{\bar{p}-1} \le \text{risk}(1, \bar{p}) + \frac{200\bar{h}}{1-1/\bar{p}}$$

concluding the overall proof of (G.1) which is a restatement of (6.4). ∎

H Proofs for Shallow Neural Networks

We consider the NAS algorithm of Section 4.2 where the solution to the lower-level problem is obtained via gradient-based. First define the Jacobian of the network and NTK kernel at the random initialization. Given training dataset $\mathcal{T}$, define the Jacobian of the network

$$\boldsymbol{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}) = [\frac{f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}_1)}{\partial \boldsymbol{W}} \ \frac{f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}_2)}{\partial \boldsymbol{W}} \ \dots \ \frac{f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}_{n_{\mathcal{T}}})}{\partial \boldsymbol{W}}]^T \in \mathbb{R}^{n_{\mathcal{T}} \times p}.$$

The Neural Tangent Kernel with activation $\sigma_{\boldsymbol{\alpha}}$ has the following kernel matrix

$$\boldsymbol{K}_{\boldsymbol{\alpha}} = \mathbb{E}_{\boldsymbol{W}_0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)}[\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)]$$

We first introduce some short-hand notation. Set $\boldsymbol{\theta} = \text{vec}(\boldsymbol{W} - \boldsymbol{W}_0)$. When $\boldsymbol{\alpha}$ is clear from context, given weights $\boldsymbol{W}$ define the network via $f_{\text{nn}}^{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ and linearized network as

$$f_{\text{lin}}^{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{v}^T \left[\sigma'_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x}) \odot (\boldsymbol{W} - \boldsymbol{W}_0)\boldsymbol{x}\right].$$

Based on this, introduce the initial prediction vector

$$\boldsymbol{p} := \boldsymbol{p}_{\boldsymbol{\alpha}} = [f_{\text{nn}}(\boldsymbol{x}_1, \boldsymbol{W}_0) \ f_{\text{nn}}(\boldsymbol{x}_2, \boldsymbol{W}_0) \ \dots \ f_{\text{nn}}(\boldsymbol{x}_n, \boldsymbol{W}_0)].$$

We then define the linearized problem

$$\widehat{\mathcal{L}}_{\mathcal{T}}^{\text{lin}}(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{p} - \boldsymbol{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\boldsymbol{\theta}\|_{\ell_2}^2. \tag{H.1}$$

For the theorem below, we denote $\boldsymbol{\theta}_t = \text{vec}(\boldsymbol{W}_t - \boldsymbol{W}_0)$. We also denote $\tilde{\boldsymbol{\theta}}_t = \text{vec}(\tilde{\boldsymbol{W}}_t - \boldsymbol{W}_0)$ where $\tilde{\boldsymbol{W}}_t$ is the linearized iterations which are obtained by training on the linearized problem $\widehat{\mathcal{L}}_{\mathcal{T}}^{\text{lin}}$.

Theorem 10 (Shallow NAS Master Theorem) *Suppose input features and labels are normalized to $\|\boldsymbol{x}\|_{\ell_2} \le 1, |y| \le 1$. Fix $\boldsymbol{v}$ with half $\sqrt{c_0/k}$ and half $-\sqrt{c_0/k}$ entries for sufficiently small $c_0 > 0$. Initialize $\boldsymbol{W}_0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let $B > 0$ upper bound $|\sigma'_{\boldsymbol{\alpha}}|, |\sigma''_{\boldsymbol{\alpha}}|$. Suppose the loss ℓ is bounded by a constant and 1-Lipschitz and NTK lower bound Assumption 3 holds and set the normalized lower bound $\bar{\lambda}_0 = \lambda_0/c_0$. Suppose the network width obeys*

$$k \gtrsim k_0 := k_0(\varepsilon, \bar{\lambda}_0, n_{\mathcal{T}})$$

Additionally suppose $n_{\mathcal{V}} \gtrsim \widetilde{\mathcal{O}}(h)$ where $\widetilde{\mathcal{O}}(\cdot)$ hides the log terms. Suppose the following holds with probability $1-p_0$ (over the initialization $\boldsymbol{W}_0$) uniformly over all $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ and for all $T \ge T_0 := T_0(\varepsilon, \bar{\lambda}_0, n_{\mathcal{T}})$

1. $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[|\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x})|], \frac{1}{n_{\mathcal{V}}}\sum_{i=1}^{n_{\mathcal{V}}}|\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\widetilde{\boldsymbol{x}}_i)| \le \varepsilon_0$.
2. *T'th iterate $\boldsymbol{\theta}_T$ obeys* $\|\boldsymbol{W}_T - \tilde{\boldsymbol{W}}_\infty\|_F = \|\boldsymbol{\theta}_T - \tilde{\boldsymbol{\theta}}_\infty\|_{\ell_2} \le \varepsilon_1$
3. *Rows are bounded via* $\|\boldsymbol{W}_T - \boldsymbol{W}_0\|_{2,\infty} \le \sqrt{C_0/k}$.
4. *At initialization, the network prediction is at most ε_2 i.e.* $\|\boldsymbol{p}_{\boldsymbol{\alpha}}\|_{\ell_2} \le \varepsilon_2$.
5. *Initial Jacobians obey* $\frac{\boldsymbol{J}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\alpha}}^T}{c_0} \succeq \bar{\lambda}_0\boldsymbol{I}_{n_{\mathcal{T}}}/2$.
6. *Initial Jacobians obey* $\|(\boldsymbol{J}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\alpha}}^T)^{-1} - \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\| \le \varepsilon_3$. *(Via Lemma 10, this is implied by* $\|\boldsymbol{J}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\alpha}}^T - \boldsymbol{K}_{\boldsymbol{\alpha}}\| \le c_0^2\bar{\lambda}_0^2\varepsilon_3/2$.*)*

Fix $M = 120B^4\bar{\lambda}_0^{-2}\Gamma(n_{\mathcal{T}}^2 + n_{\mathcal{V}}^2)\|\boldsymbol{y}\|_{\ell_2}$. *Then, with probability* $1 - 4e^{-t} - p_0$, *δ-approximate NAS output obeys*

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \le \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} 2B\sqrt{\frac{c_0\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}}{n_{\mathcal{T}}}} + C\sqrt{\frac{h\log(M)+t}{n_{\mathcal{V}}}} + \varepsilon + \delta, \tag{H.2}$$

where $\varepsilon = 3(\varepsilon_0 + \sqrt{c_0}BC_0/\sqrt{k} + \sqrt{c_0}B\varepsilon_1 + 2B\varepsilon_2/\sqrt{\bar{\lambda}_0}) + B\sqrt{c_0\varepsilon_3}$. *Additionally, since hinge loss dominates the 0-1 loss (standard classification error), the bound above also applied for the 0-1 loss* $\mathcal{L}^{0-1}$.

Proof For the proof, we would like to employ Theorem 2. To this aim, we introduce the so-called ideal feature map regression problem. Unlike (H.1), ideal problem uses the exact labels $\boldsymbol{y}$ and solves

$$\widehat{\mathcal{L}}_{\mathcal{T}}^{\text{ideal}}(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\boldsymbol{\theta}\|_{\ell_2}^2. \tag{H.3}$$

We define the ideal model to be the pseudo-inverse

$$\boldsymbol{\theta}^{\text{ideal}} = \boldsymbol{J}_{\boldsymbol{\alpha}}^{\dagger}\boldsymbol{y}. \tag{H.4}$$

Note that the ideal problem is equivalent to the feature map regression task described in Definition 2 where feature maps are $\frac{\partial f_{\text{nn}}(\boldsymbol{x})}{\partial \boldsymbol{W}_0}$. Thus, we can also study the generalization risk of $\boldsymbol{\theta}^{\text{ideal}}$ on the new examples given by $\mathcal{L}(f_{\text{lin}}^{\boldsymbol{\theta}^{\text{ideal}}})$ where

$$\mathcal{L}(f_{\text{lin}}^{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{D}}[\ell(y, \boldsymbol{\theta}^T\frac{\partial f_{\text{nn}}(\boldsymbol{x})}{\partial \boldsymbol{W}_0})]. \tag{H.5}$$

To proceed with the proof, set $\varepsilon' = \varepsilon_0 + BC_0\sqrt{c_0/k} + \sqrt{c_0}B\varepsilon_1 + 2\sqrt{c_0}B\varepsilon_2/\sqrt{\lambda_0}$. Let $\widehat{\boldsymbol{\alpha}}$ be a δ-approximate solution of the NAS problem. We first apply a triangle inequality on Lemmas 17, 18. Specifically, with probability $1 - p_0$ over $\boldsymbol{W}_0$, for all $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, Lemmas 17, 18 hold. Fix $\tilde{\mathcal{L}} \in \{\mathcal{L}, \widehat{\mathcal{L}}_{\mathcal{V}}\}$ (i.e. either validation or population loss). Thus recalling $f_{\boldsymbol{\alpha}}^{\mathcal{T}} = f_{\text{nn},\boldsymbol{\alpha}}^{\boldsymbol{\theta}_T}$, we write

$$\begin{aligned}
&|\tilde{\mathcal{L}}(f_{\text{nn}}^{\boldsymbol{\theta}_T}) - \tilde{\mathcal{L}}(f_{\text{lin}}^{\tilde{\boldsymbol{\theta}}_\infty})| \le \varepsilon_0 + \frac{\sqrt{c_0}BC_0}{\sqrt{k}} + \sqrt{c_0}B\varepsilon_1 \\
&|\tilde{\mathcal{L}}(f_{\text{lin}}^{\tilde{\boldsymbol{\theta}}_\infty}) - \tilde{\mathcal{L}}(f_{\text{lin}}^{\boldsymbol{\theta}^{\text{ideal}}})| \le 2\sqrt{c_0}B\varepsilon_2/\sqrt{\lambda_0} \\
\implies &|\tilde{\mathcal{L}}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) - \tilde{\mathcal{L}}(f_{\text{lin}}^{\boldsymbol{\theta}^{\text{ideal}}})| \le \varepsilon'.
\end{aligned} \tag{H.6}$$

Thus any δ-approximate solution $\widehat{\boldsymbol{\alpha}}$ of the NAS problem ensures that

$$\widehat{\mathcal{L}}_{\mathcal{V}}(f^{\boldsymbol{\theta}^{\text{ideal}}_{\widehat{\boldsymbol{\alpha}}}}) \leq \widehat{\mathcal{L}}_{\mathcal{V}}(f^{\mathcal{T}}_{\widehat{\boldsymbol{\alpha}}}) + \varepsilon' \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \widehat{\mathcal{L}}_{\mathcal{V}}(f^{\mathcal{T}}_{\boldsymbol{\alpha}}) + \varepsilon' + \delta \leq \inf_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \widehat{\mathcal{L}}_{\mathcal{V}}(f^{\boldsymbol{\theta}^{\text{ideal}}_{\boldsymbol{\alpha}}}) + 2\varepsilon' + \delta.$$

Thus, $f^{\boldsymbol{\theta}^{\text{ideal}}_{\widehat{\boldsymbol{\alpha}}}}$ is a $(2\varepsilon'+\delta)$-approximate solution of the linearized feature map regression. To proceed, Lemma 19 establishes the generalization guarantee for such a $f^{\boldsymbol{\theta}^{\text{ideal}}_{\widehat{\boldsymbol{\alpha}}}}$ via

$$\mathcal{L}(f^{\boldsymbol{\theta}^{\text{ideal}}_{\widehat{\boldsymbol{\alpha}}}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} 2\sqrt{c_0}B\sqrt{\frac{\boldsymbol{y}^T(\boldsymbol{J}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\alpha}}^T)^{-1}\boldsymbol{y}}{n_{\mathcal{T}}}} + C\sqrt{\frac{h\log(M)+\tau}{n_{\mathcal{V}}}} + 2\varepsilon' + \delta.$$

Finally, we go back to neural net's generalization via setting $\tilde{\mathcal{L}} = \mathcal{L}$ in (H.6) which gives $|\mathcal{L}(f^{\mathcal{T}}_{\widehat{\boldsymbol{\alpha}}}) - \mathcal{L}(f^{\boldsymbol{\theta}^{\text{ideal}}_{\widehat{\boldsymbol{\alpha}}}}_{\text{lin}})| \leq \varepsilon'$ where we note that $f^{\mathcal{T}}_{\widehat{\boldsymbol{\alpha}}} = f^{\boldsymbol{\theta}_T}_{\text{nn},\widehat{\boldsymbol{\alpha}}}$. To conclude also plug in (H.7) to move to $\boldsymbol{K}_{\boldsymbol{\alpha}}$. These as a whole imply Theorem 10's statement (H.2) after (1) applying the change of variable $3\varepsilon' + \sqrt{c_0}B\sqrt{\varepsilon_3} \leftrightarrow \varepsilon$ and then applying the change of variable $\bar{\lambda}_0 = c_0\lambda_0$. ■

Lemma 17 *Consider the setup of Theorem 10, specifically the itemized assumptions involving the initialization $\boldsymbol{W}_0$ which holds with probability $1-p_0$. Let $\boldsymbol{\theta}_T, \tilde{\boldsymbol{\theta}}_T$ be the iterations induced by any fixed activation $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$. For $\tilde{\mathcal{L}} \in \{\mathcal{L}, \widehat{\mathcal{L}}_{\mathcal{V}}\}$ (i.e. for population or validation risk), we have that*

$$|\tilde{\mathcal{L}}(f^{\boldsymbol{\theta}_T}_{nn}) - \tilde{\mathcal{L}}(f^{\tilde{\boldsymbol{\theta}}_\infty}_{lin})| \leq \varepsilon_0 + \frac{\sqrt{c_0}BC_0}{\sqrt{k}} + \sqrt{c_0}B\varepsilon_1.$$

Proof Applying Lemma 20, we find that

$$|\tilde{\mathcal{L}}(f^{\boldsymbol{\theta}_T}_{\text{nn}}) - \tilde{\mathcal{L}}(f^{\boldsymbol{\theta}_T}_{\text{lin}})| \leq \varepsilon_0 + \frac{\sqrt{c_0}C_0B}{\sqrt{k}}.$$

Next, observe that for any input with $\|\boldsymbol{x}\|_{\ell_2} \leq 1$, the neural feature maps are bounded by

$$\|\frac{\partial f_{\text{nn}}(\boldsymbol{x})}{\partial \boldsymbol{W}_0}\|_{\ell_2} \leq \sqrt{c_0}B.$$

This means that

$$|\tilde{\mathcal{L}}(f^{\boldsymbol{\theta}_T}_{\text{lin}}) - \tilde{\mathcal{L}}(f^{\tilde{\boldsymbol{\theta}}_\infty}_{\text{lin}})| \leq \sqrt{c_0}B\|\tilde{\boldsymbol{\theta}}_\infty - \boldsymbol{\theta}_T\|_{\ell_2} \leq \sqrt{c_0}B\varepsilon_1.$$

To conclude use a triangle inequality to combine the bounds above. ■

Lemma 18 (Bounding the perturbation of the linearized model) *Consider the setup of Theorem 10, specifically the itemized assumptions involving the initialization $\boldsymbol{W}_0$ which holds with probability $1-p_0$. Recall $\boldsymbol{\theta}^{ideal}$ from (H.4) and that $\boldsymbol{p}$ is the prediction vector on $\mathcal{T}$ at $\boldsymbol{W}_0$ bounded as $\|\boldsymbol{p}\|_{\ell_2} \leq \varepsilon_2$. For $\tilde{\mathcal{L}} \in \{\mathcal{L}, \widehat{\mathcal{L}}_{\mathcal{V}}\}$, we have that*

$$\tilde{\mathcal{L}}(f^{\tilde{\boldsymbol{\theta}}_\infty}_{lin}) \leq \tilde{\mathcal{L}}(f^{\boldsymbol{\theta}^{ideal}}_{lin}) + 2\sqrt{c_0}B\varepsilon_2/\sqrt{\lambda_0}.$$

Additionally, the perturbation due to empirical vs population Jacobian is bounded via

$$\sqrt{\boldsymbol{y}^T(\boldsymbol{J}\boldsymbol{J}^T)^{-1}\boldsymbol{y}} \leq \sqrt{\varepsilon_3}\|\boldsymbol{y}\|_{\ell_2} + \sqrt{\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y}}. \tag{H.7}$$

Proof Recall that feature map norm is bounded by $\sqrt{c_0}B$ and thus

$$\tilde{\mathcal{L}}(f_{\text{lin}}^{\tilde{\boldsymbol{\theta}}_\infty}) \le \tilde{\mathcal{L}}(f_{\text{lin}}^{\boldsymbol{\theta}^{\text{ideal}}}) + \sqrt{c_0}B\|\boldsymbol{\theta}^{\text{ideal}} - \tilde{\boldsymbol{\theta}}_\infty\|_{\ell_2}.$$

We upper bound the right hand side via

$$\|\boldsymbol{\theta}^{\text{ideal}} - \tilde{\boldsymbol{\theta}}_\infty\|_{\ell_2} \le \|\boldsymbol{J}_{\boldsymbol{\alpha}}^\dagger \boldsymbol{y} - \boldsymbol{J}_{\boldsymbol{\alpha}}^\dagger(\boldsymbol{y}-\boldsymbol{p})\|_{\ell_2} \le 2\|\boldsymbol{p}\|_{\ell_2}/\sqrt{\lambda_0} \le 2\varepsilon_2/\sqrt{\lambda_0}.$$

For the next result let $\boldsymbol{P} = \boldsymbol{K}^{-1} - (\boldsymbol{J}\boldsymbol{J}^T)^{-1}$. Using $\|\boldsymbol{P}\| \le \varepsilon_3$, we have that

$$\sqrt{\boldsymbol{y}^T(\boldsymbol{J}\boldsymbol{J}^T)^{-1}\boldsymbol{y}} \le \sqrt{\boldsymbol{y}^T(\boldsymbol{K}^{-1}-\boldsymbol{P})\boldsymbol{y}} \le \sqrt{\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y}} + \sqrt{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}} \le \sqrt{\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y}} + \|\boldsymbol{y}\|_{\ell_2}\sqrt{\|\boldsymbol{P}\|}.$$

■

The next result shows a uniform upper bound on the ideal solutions $\boldsymbol{\theta}_{\boldsymbol{\alpha}}^{\text{ideal}}$ which solve the feature map regression with Jacobian matrix $\boldsymbol{J}_{\boldsymbol{\alpha}}$.

Lemma 19 *Fix* $M = 120B^4\bar{\lambda}_0^{-2}\Gamma(n_{\mathcal{T}}^2 + n_{\mathcal{V}}^2)\|\boldsymbol{y}\|_{\ell_2}$. *Let* $\widehat{\boldsymbol{\alpha}}$ *be a* δ*-approximate solution of* (TVO) *with linearized Jacobian feature map with labels* $\boldsymbol{y}$. *Set* $\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} = \boldsymbol{J}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\alpha}}^T$ *and suppose* $\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}} \succeq \lambda_0 \boldsymbol{I}_{n_{\mathcal{T}}}/2$ *for all* $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ *with* $\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \|\boldsymbol{\alpha}\|_{\ell_1} \le 1$. *Recall the definition* (H.5). *With probability at least* $1 - 2\mathrm{e}^{-\tau}$, *we have that*

$$\mathcal{L}^{\text{ideal}}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \le \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} 2\sqrt{c_0}B\sqrt{\frac{\boldsymbol{y}^T\widehat{\boldsymbol{K}}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}}{n_{\mathcal{T}}}} + C\sqrt{\frac{h\log(M)+\tau}{n_{\mathcal{V}}}} + \delta. \tag{H.8}$$

where $f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}$ *is the* δ*-approximate solution of* (TVO) *with the feature map regression problem* (H.3).

Proof We need to plug in the right quantities into Theorem 2. First note that we assumed $\|\boldsymbol{\alpha}\|_{\ell_1} = R = 1$. First observe that neural feature maps $\frac{\partial f_{\text{nn}}(\boldsymbol{x})}{\partial \boldsymbol{W}}$ are bounded by $\sqrt{c_0}B$ in Euclidean norm thus we substitute $B \leftrightarrow c_0B^2$. Secondly, the Jacobian feature matrix $\boldsymbol{J}_{\boldsymbol{\alpha}}$ obeys (4.3) with $\lambda_0/2$. Thus we also set $\lambda_0 \leftrightarrow \lambda_0/2$, $\Gamma = 1$ and $R = 1$. Finally, apply the change of variable to normalized λ_0 via $\bar{\lambda}_0 = \lambda_0/c_0$. Thus we exactly find (H.8) for $M = 120B^4\bar{\lambda}_0^{-2}\Gamma(n_{\mathcal{T}}^2 + n_{\mathcal{V}}^2)\|\boldsymbol{y}\|_{\ell_2}$. ■

Lemma 20 *Let* $\boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$. *Suppose* σ *is a function with second derivative bounded by* $B > 0$ *in absolute value. Let* $c_0, C_0 > 0$ *be scalars. Suppose* $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ *is such that* $\sup_{1\le i\le k} \|\boldsymbol{w}_i - \boldsymbol{w}_{0,i}\|_{\ell_2} \le \sqrt{C_0/k}$ *and* $\|\boldsymbol{v}\|_{\ell_\infty} \le \sqrt{c_0/k}$. *Define neural net* $f_{nn}(\boldsymbol{x}) = \boldsymbol{v}^T\sigma(\boldsymbol{W}\boldsymbol{x})$ *and its linearization*

$$f_{lin}(\boldsymbol{x}) = \boldsymbol{v}^T(\sigma'(\boldsymbol{W}_0\boldsymbol{x}) \odot (\boldsymbol{W}-\boldsymbol{W}_0)\boldsymbol{x}).$$

Suppose input space $\mathcal{X}$ *is subset of unit Euclidean ball and* $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[|\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x})|]$, $\frac{1}{n_{\mathcal{V}}}\sum_{i=1}^{n_{\mathcal{V}}} |\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\widetilde{\boldsymbol{x}}_i)| \le \varepsilon_0$. *Let* ℓ *be a* Γ*-Lipschitz loss. Then for* $\tilde{\mathcal{L}} \in \{\mathcal{L}, \widehat{\mathcal{L}}_{\mathcal{V}}\}$

$$|\tilde{\mathcal{L}}(f_{nn}) - \tilde{\mathcal{L}}(f_{lin})| \le \Gamma(\varepsilon_0 + \frac{\sqrt{c_0}C_0B}{\sqrt{k}}).$$

Proof Let $\bar{f}_{\text{nn}}(\boldsymbol{x}) = f_{\text{nn}}(\boldsymbol{x}) - \boldsymbol{v}^T\sigma(\boldsymbol{W}_0\boldsymbol{x})$. Via Taylor series expansion, for any $\|\boldsymbol{x}\|_{\ell_2} \leq 1$

$$\begin{aligned}
|\bar{f}_{\text{nn}}(\boldsymbol{x}) - f_{\text{lin}}(\boldsymbol{x})| &= \sum_{i=1}^{k} |\boldsymbol{v}_i\sigma''(\boldsymbol{w}_{0,i}^T\boldsymbol{x})((\boldsymbol{w}_i - \boldsymbol{w}_{0,i})^T\boldsymbol{x})^2| \\
&= \sum_{i=1}^{k} \|\boldsymbol{v}\|_{\ell_\infty} B\|\boldsymbol{w}_i - \boldsymbol{w}_{0,i}\|_{\ell_2}^2\|\boldsymbol{x}\|_{\ell_2}^2 \\
&\leq B\|\boldsymbol{x}\|_{\ell_2}^2 \sum_{i=1}^{k} \frac{\sqrt{c_0}}{\sqrt{k}}(\sqrt{\frac{C_0}{k}})^2 \\
&\leq \frac{\sqrt{c_0}C_0\|\boldsymbol{x}\|_{\ell_2}^2 B}{\sqrt{k}} \leq \frac{\sqrt{c_0}C_0 B}{\sqrt{k}}.
\end{aligned}$$

Since loss function is Γ Lipschitz, we obtain

$$|\tilde{\mathcal{L}}(\bar{f}_{\text{nn}}) - \tilde{\mathcal{L}}(f_{\text{lin}})| \leq \Gamma(\varepsilon_0 + \frac{\sqrt{c_0}C_0 B}{\sqrt{k}}).$$

We conclude via triangle inequality after using the condition of small $\tilde{\mathcal{L}}$ prediction at $\boldsymbol{W}_0$ (which bounds $|f_{\text{nn}} - \bar{f}_{\text{nn}}|$). ■

I Gradient Descent Analysis for Shallow Networks

This section only focuses on the training dataset $\mathcal{T}$. Thus, to keep notation more concise, throughout we suppose $\mathcal{T}$ is a dataset with n samples (i.e. we set $n_{\mathcal{T}} \leftarrow n$). Following Section 4.2, starting at a random initialization $\boldsymbol{W}_0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we optimize the training loss

$$\widehat{\mathcal{L}}_{\mathcal{T}}(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}_i, \boldsymbol{W}))^2 = \frac{1}{2}\|\boldsymbol{y} - f_{\boldsymbol{\alpha}}(\boldsymbol{W})\|_{\ell_2}^2,$$

via gradient updates $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\widehat{\mathcal{L}}_{\mathcal{T}}(\boldsymbol{W}_\tau)$ for T iterations. Here $\boldsymbol{y}$ is the concatenated label vector and $f_{\boldsymbol{\alpha}}(\boldsymbol{W})$ is the prediction vector with entries $f_{\text{nn},\boldsymbol{\alpha}}(\boldsymbol{x}_i, \boldsymbol{W})$. We will drop the subscript $\boldsymbol{\alpha}$ as the $\boldsymbol{\alpha}$-dependence is clear from context. Consider the mixture of activation functions given by

$$\sigma_{\boldsymbol{\alpha}}(z) = \sum_{r=1}^{h} \boldsymbol{\alpha}_r\sigma_r(z)$$

Throughout this section we assume $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$. Assume $\boldsymbol{\Delta}$ is subset of the unit ℓ_1 ball i.e. all $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ obeys $\|\boldsymbol{\alpha}\|_{\ell_1} \leq 1$. Next we define the neural tangent kernel.

Definition 4 (Neural tangent kernel and minimum eigenvalue) *Let $\boldsymbol{w} \in \mathbb{R}^d$ be a random vector with a $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ distribution. Also consider a set of n input data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ aggregated into the rows of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$. Associated to a network $\boldsymbol{x} \mapsto \boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ and the input data matrix $\boldsymbol{X}$ we define the neural tangent kernel matrix as*

$$\boldsymbol{K}_{\boldsymbol{\alpha}} = \mathbb{E}_{\boldsymbol{W}_0 \overset{i.i.d}{\sim} \mathcal{N}(0,1)}[\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)]$$

We also define the eigenvalue $\lambda_{\boldsymbol{\alpha}}(\boldsymbol{X})$ based on $\boldsymbol{K}_{\boldsymbol{\alpha}}(\boldsymbol{X})$ as

$$\lambda_{\boldsymbol{\alpha}}(\boldsymbol{X}) := \lambda_{\min}\left(\boldsymbol{K}_{\boldsymbol{\alpha}}(\boldsymbol{X})\right).$$

Assumption 10 *We assume*

$$\min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \lambda_{\boldsymbol{\alpha}}(\boldsymbol{X}) \geq \lambda_0(\boldsymbol{X})$$

Additionally define the invariant initialization-scale lower bound

$$\bar{\lambda}_0(\boldsymbol{X}) = \frac{\lambda_0(\boldsymbol{X})}{\|\boldsymbol{v}\|_{\ell_2}^2}$$

and state the bounds in terms of this quantity.

Theorem 11 *Consider a data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ aggregated as rows/entries of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a label vector $\boldsymbol{y} \in \mathbb{R}^n$. Without loss of generality we assume the dataset is normalized so that $\|\boldsymbol{x}_i\|_{\ell_2} = 1$. Also consider a one-hidden layer neural network with k hidden units and one output of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ with $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ the input-to-hidden and hidden-to-output weights. We assume the activations $\sigma_1, \sigma_2, \ldots, \sigma_h$ with $h \leq n$ has bounded derivatives i.e. $\left|\sigma_j'(z)\right| \leq B$ and $\left|\sigma_j''(z)\right| \leq B$ for all z. Also let $\bar{\lambda}_0(\boldsymbol{X})$ denote the minimum eigenvalue of the neural net covariance per Assumption 10. Furthermore, we fix $\boldsymbol{v}$ by setting half of the entries of $\boldsymbol{v} \in \mathbb{R}^k$ to $\frac{\sqrt{c_0}}{\sqrt{k}}$ and the other half to $-\frac{\sqrt{c_0}}{\sqrt{k}}$ with $\sqrt{c_0}$ obeying $\sqrt{c_0} \leq \frac{1}{4\sqrt{\log n}}$ and train only over $\boldsymbol{W}$. Starting from an initial weight matrix $\boldsymbol{W}_0$ selected at random with i.i.d. $\mathcal{N}(0,1)$ entries, we run Gradient Descent (GD) updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ with step size $\eta \leq \frac{1}{2c_0B^2\|\boldsymbol{X}\|^2}$. Then, as long as, for some $\gamma \leq 1$ and and $C > 0$ a fixed numerical constant, we have*

$$k \geq C\frac{1}{\gamma^4\bar{\lambda}_0^8(\boldsymbol{X})}(\log n)B^{16}\|\boldsymbol{X}\|^{16}h + C\frac{B^8 n\|\boldsymbol{X}\|^8}{c_0\gamma^2\bar{\lambda}_0^5}, \tag{I.1}$$

then there is an event of probability at least $1-\frac{4}{n^3}-4e^{-10h}$ such that on this event, for all activation choices $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, all GD iterates obey

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2}^2 \leq 4n\left(1 - \eta\frac{c_0\bar{\lambda}_0(\boldsymbol{X})}{8}\right)^\tau, \tag{I.2}$$

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \leq \frac{16\sqrt{n}}{\sqrt{c_0\bar{\lambda}_0(\boldsymbol{X})}},$$

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_{2,\infty} \leq \frac{32B\|\boldsymbol{X}\|}{\sqrt{k}\sqrt{c_0}\bar{\lambda}_0(\boldsymbol{X})}\sqrt{n},$$

$$\left\|\boldsymbol{W}_\tau - \widetilde{\boldsymbol{W}}_\infty\right\|_F \leq \frac{5}{2}\frac{\gamma}{\sqrt{c_0}B\|\boldsymbol{X}\|}\sqrt{n} + 4\left(1 - \frac{1}{4}\eta c_0\bar{\lambda}_0(\boldsymbol{X})\right)^t\frac{\sqrt{n}}{\sqrt{c_0\bar{\lambda}_0(\boldsymbol{X})}} \tag{I.3}$$

Furthermore, on the same event, we also have:
• (a) for any two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over the unit Euclidean ball of $\mathbb{R}^d$

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_1}[\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x})] \leq \sqrt{c_0}\left(1 + 3\sqrt{\log n}\right)B \quad \textit{and} \quad \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_2}[\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x})] \leq \sqrt{c_0}\left(1 + 3\sqrt{\log n}\right)B, \tag{I.4}$$

• *(b) and prediction at initialization obeys*

$$\left\|\sigma_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}_0^T\right)\boldsymbol{v}\right\|_{\ell_2} \leq \sqrt{c_0}\sqrt{n}\left(1+3\sqrt{\log n}\right)B, \tag{I.5}$$

• *(c) and the following bound on the Jacobian matrix*

$$\left\|\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0) - \boldsymbol{K}_{\boldsymbol{\alpha}}(\boldsymbol{X})\right\| \leq \varepsilon_0^2 \tag{I.6}$$

holds for $\varepsilon_0 = \frac{\gamma}{512}\frac{\sqrt{c_0}\bar{\lambda}_0^2(\boldsymbol{X})}{B^3\|\boldsymbol{X}\|^3}$.

I.1 Proof of Theorem 11

In order to prove this result we first need to state some auxiliary lemmas that characterize various properties of the Jacobian matrix. The first two concern the uniform concentration of the Jacobian matrix and uniform bound on the minimum eigenvalue at initialization and will be proven later on in this section.

Lemma 21 (Jacobian Concentration) *Consider a one-hidden layer neural network model of the form* $\boldsymbol{x} \mapsto \boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}\left(\boldsymbol{W}\boldsymbol{x}\right)$ *where the activations* $\sigma_1, \sigma_2, \ldots, \sigma_h$ *have bounded second derivatives obeying* $\left|\sigma_j''(z)\right| \leq B$. *Also assume we have* n *data points* $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ *with unit euclidean norm* ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$). *Then, as long as*

$$\frac{1}{\|\boldsymbol{v}\|_{\ell_4}^4} \geq \frac{C}{\varepsilon_0^4}(\log n)B^4\left\|\boldsymbol{X}\right\|^4 h,$$

the Jacobian matrix at a random point $\boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$ *with i.i.d.* $\mathcal{N}(0,1)$ *entries obeys*

$$\left\|\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0) - \boldsymbol{K}_{\boldsymbol{\alpha}}(\boldsymbol{X})\right\| \leq \varepsilon_0^2$$

holds simultaneously for all $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ *with probability at least* $1 - 4e^{-10h}$.

Lemma 22 (Minimum eigenvalue of the Jacobian at initialization) *Consider a one-hidden layer neural network model of the form* $\boldsymbol{x} \mapsto \boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}\left(\boldsymbol{W}\boldsymbol{x}\right)$ *where the activations* $\sigma_1, \sigma_2, \ldots, \sigma_h$ *have bounded derivatives obeying* $\left|\sigma_j'(z)\right| \leq B$. *Also assume we have* n *data points* $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ *with unit euclidean norm* ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$). *Then, as long as*

$$\frac{\|\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_\infty}} \geq \sqrt{30h\log(nk)}\frac{\|\boldsymbol{X}\|}{\sqrt{\bar{\lambda}_0(\boldsymbol{X})}}B,$$

the Jacobian matrix at a random point $\boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$ *with i.i.d.* $\mathcal{N}(0,1)$ *entries obeys*

$$\min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} \sigma_{\min}\left(\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\right) \geq \frac{1}{2}\sqrt{c_0\bar{\lambda}_0(\boldsymbol{X})},$$

with probability at least $1 - \frac{1}{n^3}$.

The next three lemmas are immediate consequences of similar results in [61] (specifically Lemmas 5.7, 5.8, 6.12 respectively) and we therefore state them without proof.

Lemma 23 (Spectral norm of the Jacobian) *Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ where the activations $\sigma_1, \sigma_2, \ldots, \sigma_h$ have bounded derivatives obeying $\left|\sigma_j'(z)\right| \leq B$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$. Then the Jacobian matrix with respect to the input-to-hidden weights obeys*

$$\max_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \|\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W})\| \leq \sqrt{k} B \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{X}\|.$$

Lemma 24 (Jacobian Lipschitzness) *Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ where the activations $\sigma_1, \sigma_2, \ldots, \sigma_h$ have bounded second order derivatives obeying $\left|\sigma_j''(z)\right| \leq M$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ with unit euclidean norm ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$). Then the Jacobian mapping with respect to the input-to-hidden weights obeys*

$$\max_{\boldsymbol{\alpha} \in \boldsymbol{\Delta}} \left\|\mathcal{J}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{W}}) - \mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W})\right\| \leq M \|\boldsymbol{v}\|_{\ell_\infty} \|\boldsymbol{X}\| \left\|\widetilde{\boldsymbol{W}} - \boldsymbol{W}\right\|_F \quad \text{for all} \quad \widetilde{\boldsymbol{W}}, \boldsymbol{W} \in \mathbb{R}^{k \times d}.$$

Lemma 25 (Upper bound on initial prediction) *Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ where the activations $\sigma_1, \sigma_2, \ldots, \sigma_h$ have bounded derivatives obeying $\left|\sigma_j'(z)\right| \leq B$ and $h \leq n$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ with unit euclidean norm ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$) aggregated as rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and the corresponding labels given by $\boldsymbol{y} \in \mathbb{R}^n$. Furthermore, assume we set half of the entries of $\boldsymbol{v} \in \mathbb{R}^k$ to $\frac{\sqrt{c_0}}{\sqrt{k}}$ and the other half to $-\frac{\sqrt{c_0}}{\sqrt{k}}$. Then for $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries*

$$\left\|\sigma_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2} \leq \sqrt{c_0}\sqrt{n}\left(1 + 3\sqrt{\log n}\right) B,$$

holds with probability at least $1 - \frac{1}{n^3}$. Additionally, let $\mathcal{D}$ be any distribution supported over the unit Euclidean ball. With probability at least $1 - \frac{1}{n^3}$, we have that

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} |\boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})| \leq \sqrt{c_0}\left(1 + 3\sqrt{\log n}\right) B.$$

Proof By the triangular inequality we have

$$\begin{aligned}\left\|\sigma_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2} &\leq \|\boldsymbol{\alpha}\|_{\ell_1} \max_j \left\|\sigma_j\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2} \\ &= \max_j \left\|\sigma_j\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}\right\|_{\ell_2}\end{aligned}$$

The result holds by applying the union bound to Lemma 6.12 of [61] with $\delta = 3\sqrt{\log n}$.

For the second result, observe that $\boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ is a $\sqrt{c_0}B$ Lipschitz function of $\boldsymbol{W}$ via

$$\begin{aligned}|\boldsymbol{v}^T \sigma_j(\boldsymbol{W}\boldsymbol{x}) - \boldsymbol{v}^T \sigma_j(\boldsymbol{W}'\boldsymbol{x})| &\leq \|\boldsymbol{v}\|_{\ell_2} \|\sigma_j(\boldsymbol{W}\boldsymbol{x}) - \sigma_j(\boldsymbol{W}'\boldsymbol{x})\|_{\ell_2} &\text{(I.7)} \\ &\leq \sqrt{c_0} B \|(\boldsymbol{W} - \boldsymbol{W}')\boldsymbol{x}\|_{\ell_2} \leq \sqrt{c_0} B. &\text{(I.8)}\end{aligned}$$

This means that the expectation function $f(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} |\boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})|$ is $\sqrt{c_0}B$ Lipschitz as well. Finally, let us find its expectation over $\boldsymbol{W} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ via

$$\mathbb{E}[f(\boldsymbol{W})] = \mathbb{E}[|\boldsymbol{v}^T \sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})|] = \sup_{1 \leq j \leq h} \mathbb{E}_{\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I})}[|\boldsymbol{v}^T \sigma_j(\boldsymbol{g})|] \leq B\, \mathbb{E}_{\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I})}[|\boldsymbol{v}^T \boldsymbol{g}|] \leq \sqrt{c_0} B. \quad \text{(I.9)}$$

Here, the final line follows from Gaussian contraction inequality. Overall, the Lipschitz tail bound yields $\mathbb{P}(f(\boldsymbol{W}) \geq (1+3\sqrt{\log n})\sqrt{c_0}B) \leq n^{-3}$. ■

With these auxiliary lemmas in place we now state a more general version of the main theorem proven later on in this section.

Theorem 12 (Meta theorem) *Consider a data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ aggregated as rows/entries of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a label vector $\boldsymbol{y} \in \mathbb{R}^n$. Without loss of generality we assume the dataset is normalized so that $\|\boldsymbol{x}_i\|_{\ell_2} = 1$. Also consider a one-hidden layer neural network with k hidden units and one output of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}\boldsymbol{x})$ with $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ the input-to-hidden and hidden-to-output weights. We assume the activations $, \sigma_2, \ldots, \sigma_h$ has bounded derivatives i.e. $\left|\sigma_j'(z)\right| \leq B$ and $\left|\sigma_j''(z)\right| \leq M$ for all z. Also let $\bar{\lambda}_0(\boldsymbol{X})$ denote the normalized minimum eigenvalue of the neural net covariance per Assumption 10. Furthermore, we fix $\boldsymbol{v}$ and train only over $\boldsymbol{W}$. Starting from an initial weight matrix $\boldsymbol{W}_0$ selected at random with i.i.d. $\mathcal{N}(0,1)$ entries we run Gradient Descent (GD) updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ with step size $\eta \leq \frac{1}{2kB^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2}$. Then, as long as*

$$\begin{aligned}
\frac{\|\boldsymbol{v}\|_{\ell_2}^2}{\|\boldsymbol{v}\|_{\ell_\infty}} \geq& 64M\frac{c_0\|\boldsymbol{X}\|}{\lambda_0(\boldsymbol{X})}\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2} \\
\frac{\|\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_\infty}} \geq& \sqrt{30h\log(nk)}\frac{\|\boldsymbol{X}\|}{\sqrt{\bar{\lambda}_0(\boldsymbol{X})}}B \\
\frac{\|\boldsymbol{v}\|_{\ell_2}^5}{k^{1.5}\|\boldsymbol{v}\|_{\ell_\infty}^4} \geq& \frac{9216}{\gamma}MB^3\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}\frac{\|\boldsymbol{X}\|^4}{\bar{\lambda}_0^{2.5}(\boldsymbol{X})}
\end{aligned} \tag{I.10}$$

and $c > 0$ a fixed numerical constant, then with probability at least $1 - \frac{1}{n^3} - 4e^{-10h}$, for all activation choices $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$, all GD iterates obey

$$\|f(\boldsymbol{W}_\tau)-\boldsymbol{y}\|_{\ell_2}^2 \leq \left(1-\eta\frac{\|\boldsymbol{v}\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})}{8}\right)^\tau\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}^2, \tag{I.11}$$

$$\|\boldsymbol{W}_\tau-\boldsymbol{W}_0\|_F \leq \frac{8}{\|\boldsymbol{v}\|_{\ell_2}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}}\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2} \tag{I.12}$$

$$\|\boldsymbol{W}_\tau-\boldsymbol{W}_0\|_{2,\infty} \leq \frac{16B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|}{\|\boldsymbol{v}\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})}\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2} \tag{I.13}$$

Furthermore, assume

$$\max_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\left\|\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)-K_{\boldsymbol{\alpha}}(\boldsymbol{X})\right\| \leq \varepsilon_0^2 \tag{I.14}$$

holds with $\varepsilon_0 \leq \frac{\gamma}{512}\frac{\|\boldsymbol{v}\|_{\ell_2}^4\bar{\lambda}_0^2(\boldsymbol{X})}{k^{1.5}\|\boldsymbol{v}\|_{\ell_\infty}^3B^3\|\boldsymbol{X}\|^3}$ for some $\gamma \leq 1$. Then

$$\left\|\boldsymbol{W}_t-\widetilde{\boldsymbol{W}}_\infty\right\|_F \leq \frac{5}{4}\frac{\gamma}{\sqrt{k}B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|}\|\boldsymbol{r}_0\|_{\ell_2} + 2\left(1-\frac{1}{4}\eta\|\boldsymbol{v}\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})\right)^t\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_2}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}} \tag{I.15}$$

holds with probability at least $1 - \frac{1}{n^3} - 4e^{-10h}$ on the same event.

Finalizing the Proof of Theorem 11

Proof We now demonstrate how Theorem 11 follows from the meta theorem above. To this aim first note that $\|\boldsymbol{v}\|_{\ell_2} = \sqrt{c_0}$ and $\|\boldsymbol{v}\|_{\ell_\infty} = \frac{\sqrt{c_0}}{\sqrt{k}}$. We also note that the choice of (I.1) (specifically the second summand involving c_0) implies (I.10) so that the above meta theorem applies (with probability at least $1-\frac{1}{n^3}-4e^{-10h}$).

We now proceed by proving the various identities.

Proof of (I.5):
This follows immediately from Lemma 25.

Proof of (I.6):
Note that by Lemma 21 equation (I.5) holds with $\varepsilon_0 = \frac{\gamma}{512}\frac{\sqrt{c_0}\bar{\lambda}_0^2(\boldsymbol{X})}{B^3\|\boldsymbol{X}\|^3}$ with probability at least $1-4e^{-10h}$ as long as we have

$$\begin{aligned}\frac{1}{\|\boldsymbol{v}\|_{\ell_4}^4} \geq \frac{C}{\varepsilon_0^4}(\log n)B^4\|\boldsymbol{X}\|^4 h \quad &\Leftrightarrow \quad k \geq C\frac{c_0^2}{\varepsilon_0^4}(\log n)B^4\|\boldsymbol{X}\|^4 h \\ &\Leftrightarrow \quad k \geq C\frac{1}{\gamma^4\bar{\lambda}_0^8(\boldsymbol{X})}(\log n)B^{16}\|\boldsymbol{X}\|^{16} h\end{aligned}$$

This is true as the latter is the same as (I.1).

Proofs of (I.2) and (I.3):
For this statement, the critical ingredient is the fact that $\varepsilon_0 \leq \frac{\gamma}{512}\frac{\sqrt{c_0}\bar{\lambda}_0^2(\boldsymbol{X})}{B^3\|\boldsymbol{X}\|^3}$ which follows from the proof of (I.6) (right above). With this in mind, the critical condition (I.14) holds and (I.15) is applicable. Thus, the proof of inequalities in (I.2) and (I.3) follow from their counter parts in Theorem 12 (more specifically equations (I.11), (I.12), (I.13), and (I.15)) by substituting the choice of $\boldsymbol{v}$ and then noting that by Lemma 25 and the upper bound on $\sqrt{c_0}$

$$\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2} = \left\|\sigma_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v}-\boldsymbol{y}\right\|_{\ell_2} \leq \|\boldsymbol{y}\|_{\ell_2} + \sqrt{c_0}\sqrt{n}\left(1+3\sqrt{\log n}\right)B \leq 2\sqrt{n}$$

holds with probability at least $1-\frac{1}{n^3}$.

Proof of (I.4):
This result is also a direct application of the second statement of Lemma 25 (i.e. bounding the expected prediction over a distribution). Since we have two distributions $(\mathcal{D}_1, \mathcal{D}_2)$, the probability of success is $1-2/n^3$.

Final step: Union bounding above results in an additional $\frac{3}{n^3}$ probability of error in the final statement to obtain an overall probability of success of $1-\frac{4}{n^3}-4e^{-10h}$. ■

I.2 Proof of Meta Theorem (Theorem 12)

Proof of (I.11) and (I.12):
The proof of equations (I.11) and (I.12) follow from Corollary 6.11 [61] by replacing the following quantities in Corollary 6.11 [61] using the auxiliary Lemmas 22, 23, and 24.

$$\alpha := \frac{1}{4}\|\boldsymbol{v}\|_{\ell_2}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}, \quad \beta := \sqrt{k}B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|, \quad L = M\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|$$

Proof of (I.13):
To prove this inequality note that

$$\begin{aligned}\|\text{mat}\left(\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W})\boldsymbol{r}\right)\|_{2,\infty} &= \|\text{diag}(\boldsymbol{v})\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{W}\boldsymbol{X}^T\right)\text{diag}(\boldsymbol{r})\boldsymbol{X}\|_{2,\infty}\\ &\leq \|\boldsymbol{v}\|_{\ell_\infty}\max_{1\leq\ell\leq k}\|\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{w}_\ell^T\boldsymbol{X}^T\right)\text{diag}(\boldsymbol{r})\boldsymbol{X}\|_{\ell_2}\\ &\leq \|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|\max_{1\leq\ell\leq k}\|\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{w}_\ell^T\boldsymbol{X}^T\right)\text{diag}(\boldsymbol{r})\|_{\ell_2}\\ &\leq B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|\|\boldsymbol{r}\|_{\ell_2}\end{aligned}$$

Furthermore, by the triangular inequality we have

$$\begin{aligned}\|\boldsymbol{W}_\tau-\boldsymbol{W}_0\|_{2,\infty} \leq& \sum_{t=0}^{\tau-1}\|\boldsymbol{W}_{t+1}-\boldsymbol{W}_t\|_{2,\infty}\\ =&\eta\sum_{t=0}^{\tau-1}\|\text{mat}\left(\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_t)\boldsymbol{r}_t\right)\|_{2,\infty}\\ \leq&\eta B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|\sum_{t=0}^{\tau-1}\|\boldsymbol{r}_t\|_{\ell_2}\\ \leq&\eta B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|\left(\sum_{t=0}^{\tau-1}\left(1-\eta\frac{\|\boldsymbol{v}\|_{\ell_2}^2\lambda_0(\boldsymbol{X})}{8}\right)^{\frac{t}{2}}\right)\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}\\ \leq&\frac{\eta B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|}{1-\sqrt{1-\eta\frac{\|\boldsymbol{v}\|_{\ell_2}^2 c_0\bar{\lambda}_0(\boldsymbol{X})}{8}}}\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}\\ \leq&\frac{16B\|\boldsymbol{v}\|_{\ell_\infty}\|\boldsymbol{X}\|}{\|\boldsymbol{v}\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})}\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}\end{aligned}$$

where in the last inequality we used the fact that for $0\leq x\leq 1$ we have $\frac{1}{1-\sqrt{1-x}}\leq\frac{2}{x}$.
Proof of (I.15):
To prove this inequality we utilize Theorem 4 of [32] with $\alpha:=\sqrt{2}\alpha$, $\beta:=\beta$, $\varepsilon_0:=2\varepsilon_0=\gamma\frac{\alpha^4}{\beta^3}$, $\varepsilon:=\gamma\frac{\alpha^4}{\beta^3}$. Assumption 1 of this theorem is satisfied by the definition of α and β. Also using (I.14) Assumption 2 of [32] holds with $2\varepsilon_0\leq\frac{\alpha^4}{\beta^3}$. Also in this case the value of R in this theorem becomes equal to

$$R:=2\left\|\boldsymbol{J}^\dagger\boldsymbol{r}_0\right\|_{\ell_2}+\frac{2.5\gamma}{\beta}\|\boldsymbol{r}_0\|_{\ell_2}\leq\frac{4.5}{\alpha}\|\boldsymbol{r}_0\|_{\ell_2}$$

Furthermore, note that as long as

$$9216MB^3\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{\gamma}\frac{\|\boldsymbol{X}\|^4}{\bar{\lambda}_0^{2.5}(\boldsymbol{X})}\leq\frac{\|\boldsymbol{v}\|_{\ell_2}^5}{k^{1.5}\|\boldsymbol{v}\|_{\ell_\infty}^4},$$

we have

$$9\|\boldsymbol{r}_0\|_{\ell_2}L\leq\frac{\gamma\alpha^5}{\beta^3}.$$

which implies that

$$RL \leq \frac{4.5}{\alpha} \left\| \boldsymbol{r}_0 \right\|_{\ell_2} L \leq \frac{\gamma \alpha^4}{2\beta^3} = \frac{\varepsilon}{2}$$

so that Assumption 3 of this theorem is also satisfied. Thus, equation (37) of [32] implies

$$\left\| \boldsymbol{W}_t - \widetilde{\boldsymbol{W}}_t \right\|_F \leq \frac{5}{4} \frac{\gamma}{\beta} \left\| \boldsymbol{r}_0 \right\|_{\ell_2}$$

which together with the triangular inequality implies

$$\left\| \boldsymbol{W}_t - \widetilde{\boldsymbol{W}}_\infty \right\|_F \leq \left\| \boldsymbol{W}_t - \widetilde{\boldsymbol{W}}_t \right\|_F + \left\| \widetilde{\boldsymbol{W}}_t - \widetilde{\boldsymbol{W}}_\infty \right\|_F \leq \frac{5}{4} \frac{\gamma}{\beta} \left\| \boldsymbol{r}_0 \right\|_{\ell_2} + \left\| \widetilde{\boldsymbol{W}}_t - \widetilde{\boldsymbol{W}}_\infty \right\|_F \tag{I.16}$$

Define $\widetilde{r}_\tau = \boldsymbol{J}\text{vect}(\widetilde{\boldsymbol{W}}_\tau) - \boldsymbol{y}$. All that remains to complete the proof is to bound the last term. To this aim consider the singular value decomposition of $\boldsymbol{J} = \boldsymbol{U_J} \boldsymbol{\Sigma_J} \boldsymbol{V_J}^T$ and note that

$$\begin{aligned}
\text{vect}(\widetilde{\boldsymbol{W}}_t) - \text{vect}(\widetilde{\boldsymbol{W}}_\infty) =& \eta \boldsymbol{J}^T \sum_{\tau=t}^{\infty} \widetilde{\boldsymbol{r}}_\tau \\
=& \eta \boldsymbol{J}^T \sum_{\tau=0}^{\infty} \widetilde{\boldsymbol{r}}_{\tau+t} \\
=& \eta \boldsymbol{J}^T \left(\boldsymbol{I} - \eta \boldsymbol{J} \boldsymbol{J}^T \right)^t \sum_{\tau=0}^{\infty} \widetilde{\boldsymbol{r}}_\tau \\
=& \eta \boldsymbol{V_J} \boldsymbol{\Sigma_J} \left(\boldsymbol{I} - \eta \boldsymbol{\Sigma_J}^2 \right)^t \boldsymbol{U_J}^T \sum_{\tau=0}^{\infty} \widetilde{\boldsymbol{r}}_\tau \\
=& \eta \boldsymbol{V_J} \left(\boldsymbol{I} - \eta \boldsymbol{\Sigma_J}^2 \right)^t \boldsymbol{\Sigma_J} \boldsymbol{U_J}^T \sum_{\tau=0}^{\infty} \widetilde{\boldsymbol{r}}_\tau \\
=& \eta \boldsymbol{V_J} \left(\boldsymbol{I} - \eta \boldsymbol{\Sigma_J}^2 \right)^t \boldsymbol{V_J}^T \boldsymbol{V_J} \boldsymbol{\Sigma_J} \boldsymbol{U_J}^T \sum_{\tau=0}^{\infty} \widetilde{\boldsymbol{r}}_\tau \\
=& \eta \boldsymbol{V_J} \left(\boldsymbol{I} - \eta \boldsymbol{\Sigma_J}^2 \right)^t \boldsymbol{V_J}^T \boldsymbol{J}^T \sum_{\tau=0}^{\infty} \widetilde{\boldsymbol{r}}_\tau \\
=& \boldsymbol{V_J} \left(\boldsymbol{I} - \eta \boldsymbol{\Sigma_J}^2 \right)^t \boldsymbol{V_J}^T \left(\text{vect}(\widetilde{\boldsymbol{W}}_0) - \text{vect}(\widetilde{\boldsymbol{W}}_\infty) \right).
\end{aligned}$$

Now using the fact that $\text{vect}(\widetilde{\boldsymbol{W}}_0) - \text{vect}(\widetilde{\boldsymbol{W}}_\infty)$ belongs to $\text{span}(\boldsymbol{J}^T)$ we conclude that

$$\begin{aligned}
\left\| \boldsymbol{W}_t - \widetilde{\boldsymbol{W}}_\infty \right\|_F \leq& \left(1 - 4\eta\alpha^2 \right)^t \left\| \boldsymbol{W}_0 - \widetilde{\boldsymbol{W}}_\infty \right\|_F \\
\leq& \left(1 - 4\eta\alpha^2 \right)^t \frac{\left\| \boldsymbol{r}_0 \right\|_{\ell_2}}{2\alpha} \\
=& 2 \left(1 - \frac{1}{4} \eta \left\| \boldsymbol{v} \right\|_{\ell_2}^2 \bar{\lambda}_0(\boldsymbol{X}) \right)^t \frac{\left\| \boldsymbol{r}_0 \right\|_{\ell_2}}{\left\| \boldsymbol{v} \right\|_{\ell_2} \sqrt{\bar{\lambda}_0(\boldsymbol{X})}}
\end{aligned}$$

Plugging the latter into (I.16) completes this proof.

I.3 Proofs for Jacobian concentration (Proof of Lemma 21)

Proof To lower bound the minimum eigenvalue of $\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)$ universally for all $\boldsymbol{\alpha}$, we focus on lower bounding the minimum eigenvalue of $\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)^T$ for a fixed $\boldsymbol{\alpha}$. To this aim we use the identity

$$\begin{aligned}\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W})\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}) &= \left(\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}^T\right)\operatorname{diag}\left(\boldsymbol{v}\right)\operatorname{diag}\left(\boldsymbol{v}\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ &= \left(\sum_{\ell=1}^k \boldsymbol{v}_\ell^2\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right),\end{aligned}$$

mentioned earlier to conclude that

$$\begin{aligned}\mathbb{E}\left[\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)^T\right] &= \|\boldsymbol{v}\|_{\ell_2}^2\left(\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}\right)^T\right]\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right),\\ &:= \boldsymbol{K}_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\right).\end{aligned}\tag{I.17}$$

Now define

$$\boldsymbol{S}_\ell(\boldsymbol{\alpha}) = \left[\boldsymbol{v}_\ell^2\left(\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right) - \boldsymbol{K}_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\right)\right]$$

and note that $\mathbb{E}[\boldsymbol{S}_\ell] = 0$. Thus,

$$\begin{aligned}\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0) - \boldsymbol{K}_{\boldsymbol{\alpha}}(\boldsymbol{X}) &= \sum_{\ell=1}^k\left[\boldsymbol{v}_\ell^2\left(\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right) - \boldsymbol{K}_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\right)\right]\\ &= \sum_{\ell=1}^k \boldsymbol{S}_\ell(\boldsymbol{\alpha})\end{aligned}$$

To show that the spectral norm is small we will use the matrix Hoeffding inequality. Next note that

$$\begin{aligned}\boldsymbol{S}_\ell(\boldsymbol{\alpha}) &\preceq \boldsymbol{v}_\ell^2\left(\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ &= \boldsymbol{v}_\ell^2\operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\right)\boldsymbol{X}\boldsymbol{X}^T\operatorname{diag}\left(\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\right)\\ &\preceq \boldsymbol{v}_\ell^2 B^2\boldsymbol{X}\boldsymbol{X}^T\end{aligned}$$

Similarly, $\boldsymbol{S}_\ell(\boldsymbol{\alpha}) \succeq -\boldsymbol{v}_\ell^2 B^2\boldsymbol{X}\boldsymbol{X}^T$. Thus, by matrix Hoeffding inequality we have

$$\mathbb{P}\left\{\left\|\sum_{\ell=1}^k \boldsymbol{S}_\ell(\boldsymbol{\alpha})\right\| \geq t\right\} \leq 2ne^{-\frac{t^2}{8\sigma^2}} \quad \text{where} \quad \sigma^2 := \left\|\sum_{\ell=1}^k \boldsymbol{v}_\ell^4 B^4\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\right\|$$

Thus, for a fixed $\boldsymbol{\alpha}$ we have

$$\mathbb{P}\left\{\left\|\sum_{\ell=1}^k \boldsymbol{S}_\ell(\boldsymbol{\alpha})\right\| \geq t\right\} \leq 2ne^{-\frac{t^2}{8\sigma^2}} \leq 2ne^{-\frac{t^2}{B^4\|\boldsymbol{v}\|_{\ell_4}^4\|\boldsymbol{X}\|^4}}.\tag{I.18}$$

Next note that for fixed $\boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\alpha}}$ we have

$$\begin{aligned}&\boldsymbol{S}_\ell(\boldsymbol{\alpha}) - \boldsymbol{S}_\ell(\widetilde{\boldsymbol{\alpha}})\\ &= \boldsymbol{v}_\ell^2\left[\left(\sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T + \sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T + \sigma'_{\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\right]\\ &\quad - \boldsymbol{v}_\ell^2\,\mathbb{E}\left[\left(\sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T + \sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T + \sigma'_{\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\right]\\ &\preceq 8\boldsymbol{v}_\ell^2 B^2\left\|\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}\right\|_{\ell_1}\|\boldsymbol{X}\|^2\boldsymbol{I}_n\end{aligned}$$

Thus, for fixed $\boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\alpha}}$ we have

$$\mathbb{P}\left\{\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha})-\boldsymbol{S}_\ell(\widetilde{\boldsymbol{\alpha}})\right\|\geq t\right\}\leq 2ne^{-\frac{t^2}{64B^4\|\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}\|_{\ell_1}^2\|\boldsymbol{v}\|_{\ell_4}^4\|\boldsymbol{X}\|^4}}.$$

Thus in turn implies that for fixed $\boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\alpha}}$ we have

$$\begin{aligned}\left\|\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha})\right\|-\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\widetilde{\boldsymbol{\alpha}})\right\|\right\|_{\psi_2}\leq&\left\|\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha})-\boldsymbol{S}_\ell(\widetilde{\boldsymbol{\alpha}})\right\|\right\|_{\psi_2}\\ \leq&c\sqrt{\log n}B^2\left\|\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}\right\|_{\ell_1}\|\boldsymbol{v}\|_{\ell_4}^2\|\boldsymbol{X}\|^2\end{aligned}$$

Thus using Talagrand's comparison inequality (Corollary 8.6.2 of [67] and Exercise) for a fixed $\boldsymbol{\alpha}_0\in\boldsymbol{\Delta}$ and all $\boldsymbol{\alpha}\in\boldsymbol{\Delta}$ we have

$$\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\left|\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha})\right\|-\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha}_0)\right\|\right|\leq c\sqrt{\log n}B^2\|\boldsymbol{v}\|_{\ell_4}^2\|\boldsymbol{X}\|^2\left(\sqrt{h}+2u\right)$$

holds with probability at least $1-2e^{-u^2}$. Plugging in $u=\sqrt{10h}$ we arrive at

$$\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\left|\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha})\right\|-\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha}_0)\right\|\right|\leq c\sqrt{\log n}B^2\|\boldsymbol{v}\|_{\ell_4}^2\|\boldsymbol{X}\|^2\sqrt{h}$$

holds with probability at least $1-2e^{-10h}$. Thus, using the triangular inequality combined with (I.18) we have

$$\begin{aligned}\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}}\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha})\right\|\leq&c\sqrt{\log n}B^2\|\boldsymbol{v}\|_{\ell_4}^2\|\boldsymbol{X}\|^2\sqrt{h}+\left\|\sum_{\ell=1}^{k}\boldsymbol{S}_\ell(\boldsymbol{\alpha}_0)\right\|\\ \leq&2c\sqrt{\log n}B^2\|\boldsymbol{v}\|_{\ell_4}^2\|\boldsymbol{X}\|^2\sqrt{h}\end{aligned}$$

holds with probability at least $1-4e^{-10h}$. ■

I.4 Proof for minimum eigenvalue of Jacobian at initialization (Proof of Lemma 22)

To lower bound the minimum eigenvalue of $\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)$ universally for all $\boldsymbol{\alpha}$, we focus on lower bounding the minimum eigenvalue of $\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)^T$ for a fixed $\boldsymbol{\alpha}$. To this aim we use the identity

$$\begin{aligned}\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W})\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W})=&\left(\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}^T\right)\operatorname{diag}\left(\boldsymbol{v}\right)\operatorname{diag}\left(\boldsymbol{v}\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{W}\boldsymbol{X}^T\right)\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right)\\ =&\left(\sum_{\ell=1}^{k}\boldsymbol{v}_\ell^2\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}_\ell\right)^T\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right),\end{aligned}$$

mentioned earlier to conclude that

$$\begin{aligned}\mathbb{E}\left[\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)^T\right]=&\|\boldsymbol{v}\|_{\ell_2}^2\left(\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}\right)\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{w}\right)^T\right]\right)\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right),\\ :=&\boldsymbol{K}_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\right).\end{aligned}\tag{I.19}$$

Thus

$$\lambda_{\min}\left(\mathbb{E}\left[\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)^T\right]\right) \geq \lambda_{\boldsymbol{\alpha}}(\boldsymbol{X}). \tag{I.20}$$

To relate the minimum eigenvalue of the expectation to that of $\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)^T$ we utilize the matrix Chernoff identity stated below.

Theorem 13 (Matrix Chernoff) *Consider a finite sequence $\boldsymbol{A}_\ell \in \mathbb{R}^{n\times n}$ of independent, random, Hermitian matrices with common dimension n. Assume that $\boldsymbol{0} \preceq \boldsymbol{A}_\ell \preceq R\boldsymbol{I}$ for $\ell = 1, 2, \ldots, k$. Then*

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{\ell=1}^{k}\boldsymbol{A}_\ell\right) \leq (1-\delta)\lambda_{\min}\left(\sum_{\ell=1}^{k}\mathbb{E}[\boldsymbol{A}_\ell]\right)\right\} \leq n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\lambda_{\min}\left(\sum_{\ell=1}^{k}\mathbb{E}[\boldsymbol{A}_\ell]\right)}{R}}$$

for $\delta \in [0, 1)$.

We shall apply this theorem with $\boldsymbol{A}_\ell := \mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{w}_\ell)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{w}_\ell) = \boldsymbol{v}_\ell^2\text{diag}(\sigma'_{\boldsymbol{\alpha}}(\boldsymbol{X}\boldsymbol{w}_\ell))\boldsymbol{X}\boldsymbol{X}^T\text{diag}(\sigma'_{\boldsymbol{\alpha}}(\boldsymbol{X}\boldsymbol{w}_\ell))$. To this aim note that

$$\boldsymbol{v}_\ell^2\text{diag}(\sigma'_{\boldsymbol{\alpha}}(\boldsymbol{X}\boldsymbol{w}_\ell))\boldsymbol{X}\boldsymbol{X}^T\text{diag}(\sigma'_{\boldsymbol{\alpha}}(\boldsymbol{X}\boldsymbol{w}_\ell)) \preceq B^2\left\|\boldsymbol{v}\right\|_{\ell_\infty}^2\left\|\boldsymbol{X}\right\|^2\boldsymbol{I},$$

so that we can use Chernoff Matrix with $R = B^2\left\|\boldsymbol{v}\right\|_{\ell_\infty}^2\left\|\boldsymbol{X}\right\|^2$ to conclude that

$$\begin{aligned}\mathbb{P}\Bigg\{\lambda_{\min}\left(\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)\right) \leq (1-\delta)\lambda_{\min}\left(\mathbb{E}\left[\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)\right]\right)\Bigg\}& \\ \leq n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\lambda_{\min}\left(\mathbb{E}\left[\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)\right]\right)}{B^2\left\|\boldsymbol{v}\right\|_{\ell_\infty}^2\left\|\boldsymbol{X}\right\|^2}}&.\end{aligned}$$

Thus using (I.20) in the above with $\delta = \frac{1}{2}$ we have

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\min}\left(\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)\right) \leq \frac{1}{2}\left\|\boldsymbol{v}\right\|_{\ell_2}^2\bar{\lambda}_0\left(\boldsymbol{X}\right)\right\} \leq& \mathbb{P}\left\{\lambda_{\min}\left(\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\mathcal{J}_{\boldsymbol{\alpha}}^T(\boldsymbol{W}_0)\right) \leq \frac{1}{2}\left\|\boldsymbol{v}\right\|_{\ell_2}^2\bar{\lambda}_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\right)\right\} \\ \leq& n\cdot e^{-\frac{1}{10}\frac{\left\|\boldsymbol{v}\right\|_{\ell_2}^2\bar{\lambda}_{\boldsymbol{\alpha}}(\boldsymbol{X})}{B^2\left\|\boldsymbol{v}\right\|_{\ell_\infty}^2\left\|\boldsymbol{X}\right\|^2}} \\ \leq& n\cdot e^{-\frac{1}{10}\frac{\left\|\boldsymbol{v}\right\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})}{B^2\left\|\boldsymbol{v}\right\|_{\ell_\infty}^2\left\|\boldsymbol{X}\right\|^2}}\end{aligned}$$

This proves the result for a fixed $\boldsymbol{\alpha}$. Now let $\mathcal{N}_\epsilon$ be an ϵ, ℓ_1 ball cover of $\boldsymbol{\Delta}$ with $\epsilon := \frac{\sqrt{2}-1}{2B\sqrt{kn}}\frac{\left\|\boldsymbol{v}\right\|_{\ell_2}}{\left\|\boldsymbol{v}\right\|_{\ell_\infty}}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}$. Then using the union bound we conclude that for all $\boldsymbol{\alpha} \in \mathcal{N}_\varepsilon$

$$\sigma_{\min}\left(\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\right) \geq \frac{\left\|\boldsymbol{v}\right\|_{\ell_2}}{\sqrt{2}}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}$$

holds with probability at least

$$\begin{aligned}1-|\mathcal{N}_\varepsilon|\, n\cdot e^{-\frac{1}{10}\frac{\|\boldsymbol{v}\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})}{B^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2}} \geq& 1-n\cdot\left(\frac{3}{\varepsilon}\right)^h e^{-\frac{1}{10}\frac{\|\boldsymbol{v}\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})}{B^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2}}\\ =&1-\frac{1}{n}e^{2\log n+h\log\left(\frac{6B\sqrt{kn}\|\boldsymbol{v}\|_{\ell_\infty}}{(\sqrt{2}-1)\|\boldsymbol{v}\|_{\ell_2}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}}\right)-\frac{1}{10}\frac{\|\boldsymbol{v}\|_{\ell_2}^2\bar{\lambda}_0(\boldsymbol{X})}{B^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2}}\\ \geq&1-\frac{1}{n^3}\end{aligned}$$

where the last line holds as long as

$$\left(\frac{\|\boldsymbol{v}\|_{\ell_2}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}}{B\|\boldsymbol{v}\|_{\ell_\infty}}\right)^2\geq 10\|\boldsymbol{X}\|^2\left(4\log n+h\log\left(\frac{6B\sqrt{kn}\|\boldsymbol{v}\|_{\ell_\infty}}{(\sqrt{2}-1)\|\boldsymbol{v}\|_{\ell_2}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}}\right)\right)$$

which in turn holds as long as

$$\frac{\|\boldsymbol{v}\|_{\ell_2}\sqrt{\bar{\lambda}_0(\boldsymbol{X})}}{B\|\boldsymbol{v}\|_{\ell_\infty}}\geq 30\|\boldsymbol{X}\|\sqrt{h\log(nk)}\geq 30\|\boldsymbol{X}\|\cdot\max\left(\sqrt{\log(n)},\sqrt{h\log(nk)}\right).$$

Next, we focus on the deviation with respect to the $\boldsymbol{\alpha}$ parameter

$$\mathcal{J}_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0)-\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)=\left(\mathrm{diag}(\boldsymbol{v})\left(\sigma'_{\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{W}_0^T\right)-\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}_0^T\right)\right)\right)*\boldsymbol{X}.$$

Now using the fact that $(\boldsymbol{A}*\boldsymbol{B})(\boldsymbol{A}*\boldsymbol{B})^T=\left(\boldsymbol{A}\boldsymbol{A}^T\right)\odot\left(\boldsymbol{B}\boldsymbol{B}^T\right)$ we conclude that

$$\begin{aligned}&\left(\mathcal{J}_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0)-\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\right)\left(\mathcal{J}_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0)-\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\right)^T\\ &\quad=\left(\left(\sigma'_{\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{W}_0^T\right)-\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}_0^T\right)\right)\mathrm{diag}(\boldsymbol{v})\mathrm{diag}(\boldsymbol{v})\left(\sigma'_{\widetilde{\boldsymbol{\alpha}}}\left(\boldsymbol{X}\boldsymbol{W}_0^T\right)-\sigma'_{\boldsymbol{\alpha}}\left(\boldsymbol{X}\boldsymbol{W}_0^T\right)\right)^T\right)\\ &\qquad\odot\left(\boldsymbol{X}\boldsymbol{X}^T\right).\end{aligned}\tag{I.21}$$

To continue further we use the fact that for to PSD matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we have $\lambda_{\max}(\boldsymbol{A}\odot\boldsymbol{B})\leq(\max_i\boldsymbol{A}_{ii})\lambda_{\max}(\boldsymbol{B})$ combined with (I.21) to conclude that

$$\begin{aligned}\|\mathcal{J}_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0)-\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\|^2\leq&\|\boldsymbol{X}\|^2\left(\max_i\left\|\mathrm{diag}(\boldsymbol{v})\left(\sigma'_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0\boldsymbol{x}_i)-\sigma'_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x}_i)\right)\right\|_{\ell_2}^2\right)\\ \leq&\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2\left(\max_i\left\|\sigma'_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0\boldsymbol{x}_i)-\sigma'_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x}_i)\right\|_{\ell_2}^2\right)\\ =&\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2\left(\max_i\left\|\sum_{j=1}^h(\widetilde{\boldsymbol{\alpha}}_j-\boldsymbol{\alpha}_j)\sigma'_j(\boldsymbol{W}_0\boldsymbol{x}_i)\right\|_{\ell_2}^2\right)\\ \leq&\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2\left(\max_i\max_j\left\|\sigma'_j(\boldsymbol{W}_0\boldsymbol{x}_i)\right\|_{\ell_2}^2\right)\|\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_{\ell_1}^2\\ \leq&kB^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|^2\|\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_{\ell_1}^2\\ \leq&kB^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\boldsymbol{X}\|_F^2\|\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_{\ell_1}^2\\ =&knB^2\|\boldsymbol{v}\|_{\ell_\infty}^2\|\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_{\ell_1}^2.\end{aligned}$$

Thus,

$$\|\mathcal{J}_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0) - \mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)\| \leq \sqrt{kn}B \|\boldsymbol{v}\|_{\ell_\infty} \|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\ell_1}$$

By the definition of the $\mathcal{N}_\varepsilon$ cover for any $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ there exists a $\widetilde{\boldsymbol{\alpha}} \in \mathcal{N}_\varepsilon$ obeying $\|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\ell_1} \leq \varepsilon$. Thus, using the above deviation inequality for any $\boldsymbol{\alpha} \in \boldsymbol{\Delta}$ we have

$$\begin{aligned}
\sigma_{\min}(\mathcal{J}_{\boldsymbol{\alpha}}(\boldsymbol{W}_0)) \geq & \sigma_{\min}(\mathcal{J}_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{W}_0)) - \sqrt{kn}B \|\boldsymbol{v}\|_{\ell_\infty} \|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\ell_1} \\
\geq & \frac{1}{\sqrt{2}} \|\boldsymbol{v}\|_{\ell_2} \sqrt{\bar{\lambda}_0(\boldsymbol{X})} - \sqrt{kn}B \|\boldsymbol{v}\|_{\ell_\infty} \varepsilon \\
= & \frac{1}{\sqrt{2}} \|\boldsymbol{v}\|_{\ell_2} \sqrt{\bar{\lambda}_0(\boldsymbol{X})} - \sqrt{kn}B \|\boldsymbol{v}\|_{\ell_\infty} \frac{\sqrt{2}-1}{2B\sqrt{kn}} \frac{\|\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_\infty}} \sqrt{\bar{\lambda}_0(\boldsymbol{X})} \\
= & \frac{1}{2} \|\boldsymbol{v}\|_{\ell_2} \sqrt{\bar{\lambda}_0(\boldsymbol{X})}
\end{aligned}$$

completing the proof.

J Proof of Theorem 3

The result follows by plugging the proper quantities in Theorem 10. Due to the output-layer scaling c_0, λ_0 grows proportional to the initialization c_0. Thus to state a bound invariant of the initialization, we define the invariant lower bound $\bar{\lambda}_0 = \lambda_0/c_0$ and state the bounds in terms of this quantity. We remark that this is consistent with the literature on neural tangent kernel analysis. Specifically, we show that, in Theorem 10, one can choose

- $k_0 \propto \frac{B^{16} h n_{\mathcal{T}}^8 \log n_{\mathcal{T}}}{\varepsilon^4 \bar{\lambda}_0^8}$
- $T_0 \propto \frac{B^2 n_{\mathcal{T}}}{\bar{\lambda}_0} \log\left(\frac{B\sqrt{n_{\mathcal{T}}}}{\varepsilon\sqrt{\bar{\lambda}_0}}\right)$.
- $p_0 = 4n_{\mathcal{T}}^{-3} + 4e^{-10h}$.
- $c_0 \propto \frac{\varepsilon^2 \bar{\lambda}_0}{B^4 n_{\mathcal{T}} (1+3\sqrt{\log n_{\mathcal{T}}})^2}$

to conclude with the proof of Theorem 3. The verification of this choice will be accomplished via Theorem 11. The following is a restatement (more precise version) of Theorem 3.

Theorem 14 (Neural activation search) *Suppose input features are normalized as $\|\boldsymbol{x}\|_{\ell_2} = 1$ and labels take values in $\{-1, 1\}$. Pick $\boldsymbol{\Delta}$ to be a subset of the unit ℓ_1 ball. Suppose Assumption 3 holds for $\boldsymbol{\theta}_0 \leftrightarrow \boldsymbol{W}_0$ and the candidate activations have first two derivatives ($|\sigma_i'|, |\sigma_i''|$) upper bounded by $B > 0$. Furthermore, fix $\boldsymbol{v}$ with half $\sqrt{c_0/k}$ and half $-\sqrt{c_0/k}$ entries for $c_0 \propto \frac{\varepsilon^2 \bar{\lambda}_0}{B^4 n_{\mathcal{T}} (1+3\sqrt{\log n_{\mathcal{T}}})^2}$ (see supplementary). Also define the initialization-invariant lower bound $\bar{\lambda}_0 = \lambda_0/c_0$. Finally, assume the network width obeys*

$$k \gtrsim \varepsilon^{-4} \bar{\lambda}_0^{-8} B^{16} h n_{\mathcal{T}}^8 \log(n_{\mathcal{T}}),$$

for a tolerance level $\varepsilon > 0$ and the size of the validation data obeys $n_{\mathcal{V}} \gtrsim \widetilde{\mathcal{O}}(h)$. Following the bilevel optimization scheme for the shallow activation search with learning rate $\eta = \frac{1}{2c_0 B^2 \|\boldsymbol{X}\|^2}$ choice and number of iterations obeying $T \gtrsim \frac{B^2 n_{\mathcal{T}}}{\bar{\lambda}_0} \log\left(\frac{B\sqrt{n_{\mathcal{T}}}}{\varepsilon\sqrt{\bar{\lambda}_0}}\right)$, the misclassification bound (0-1 loss)

$$\mathcal{L}^{0-1}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} 2B\sqrt{\frac{c_0 \boldsymbol{y}^T \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}}{n_{\mathcal{T}}}} + C\sqrt{\frac{\widetilde{\mathcal{O}}(h)+t}{n_{\mathcal{V}}}} + \varepsilon + \delta, \tag{J.1}$$

holds with probability at least $1-4(e^{-t}+n_{\mathcal{T}}^{-3}+e^{-10h})$ *(over the randomness in* $\boldsymbol{W}_0, \mathcal{T}, \mathcal{V}$*). Here,* $\boldsymbol{y} = [y_1\ y_2\ \ldots\ y_{n_{\mathcal{T}}}]$*. Finally, on the same event, for all* $\boldsymbol{\alpha}\in\boldsymbol{\Delta}$*, training classification error obeys* $\hat{\mathcal{L}}_{\mathcal{T}}^{0-1}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) \leq \varepsilon$*.*

J.1 Finalizing the proof by verifying the k_0, T_0, c_0, p_0 choices

The main strategy for the proof is combining Theorem 11 with Theorem 10. We ensure that all five summands in the error $3(\varepsilon_0+\sqrt{c_0}BC_0/\sqrt{k}+\sqrt{c_0}B\varepsilon_1+2B\varepsilon_2/\sqrt{\bar{\lambda}_0})+B\sqrt{c_0\varepsilon_3}$ in Theorem 10 is less or equal to $1/5$ of the tolerance error ε. To achieve this goal, we set

$$\gamma = \frac{\varepsilon\,\|\boldsymbol{X}\|}{75\sqrt{n_{\mathcal{T}}}} \quad \text{and} \quad c_0 = \Big(\frac{\varepsilon\sqrt{\bar{\lambda}_0}}{30B^2\sqrt{n_{\mathcal{T}}}(1+3\sqrt{\log n_{\mathcal{T}}})}\Big)^2.$$

in this section and prove the result (that these choices of k_0, T_0, c_0, p_0 are valid).

Step 0: Verifying $k \geq k_0$ satisfies conditions. First, recall from Theorem 11 that we need to verify

$$k \geq \frac{C(\log n_{\mathcal{T}})B^{16}\,\|\boldsymbol{X}\|^{16}\,h}{\gamma^4\bar{\lambda}_0^8} + C\frac{n_{\mathcal{T}}B^8\|\boldsymbol{X}\|^8}{c_0\gamma^2\bar{\lambda}_0^5} \tag{J.2}$$

Observe that the second summand is bounded as follows (plugging in our c_0 choice)

$$\begin{aligned}\frac{nB^8\|\boldsymbol{X}\|^8}{c_0\gamma^2\bar{\lambda}_0^5} &\propto \frac{n_{\mathcal{T}}B^8\|\boldsymbol{X}\|^8}{\gamma^2\bar{\lambda}_0^5} \times \frac{B^4 n_{\mathcal{T}}(1+3\sqrt{\log n_{\mathcal{T}}})^2}{\varepsilon^2\bar{\lambda}_0} \\ &\propto \frac{n_{\mathcal{T}}^2 B^{12}\|\boldsymbol{X}\|^8\log n_{\mathcal{T}}}{\varepsilon^2\gamma^2\bar{\lambda}_0^6}.\end{aligned}$$

To proceed, by plugging in the value of γ (and using $B^2 n_{\mathcal{T}} \geq \bar{\lambda}_0$) we find that $k \geq k_0$ implies (J.2) via the following list of implications

$$\begin{aligned}
&k \gtrsim \frac{B^{16}n_{\mathcal{T}}^2\,\|\boldsymbol{X}\|^{12}\,h\log n_{\mathcal{T}}}{\varepsilon^4\bar{\lambda}_0^8} + \frac{n_{\mathcal{T}}^3 B^{12}\|\boldsymbol{X}\|^6\log n_{\mathcal{T}}}{\varepsilon^4\bar{\lambda}_0^6} \\
&k \gtrsim \frac{B^{16}n_{\mathcal{T}}^8 h\log n_{\mathcal{T}}}{\varepsilon^4\bar{\lambda}_0^8} + \frac{n_{\mathcal{T}}^6 B^{12}\log n_{\mathcal{T}}}{\varepsilon^4\bar{\lambda}_0^6} \quad \Longleftarrow \\
&k \propto \frac{B^{16}n_{\mathcal{T}}^8 h\log n_{\mathcal{T}}}{\varepsilon^4\bar{\lambda}_0^8} \quad \Longleftarrow \\
&k \geq k_0.
\end{aligned}$$

Step 1: Verifying the choice of p_0 and obtaining the values of $\varepsilon_0, \varepsilon_1, C_0, \varepsilon_2, \varepsilon_3$ for the itemized list of Theorem 10. We will substitute the bounds (I.3), (I.4), (I.5), (I.6) from Theorem 11 in the itemized assumptions of Theorem 10. Thus, setting $p_0 = 4n_{\mathcal{T}}^{-3} + 4e^{-10h}$, the following conditions hold with probability $1-\frac{4}{n_{\mathcal{T}}^3}-4e^{-10h}$ (which is the success probability of Theorem 11)

1. $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[|\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\boldsymbol{x})|], \frac{1}{n_{\mathcal{V}}}\sum_{i=1}^{n_{\mathcal{V}}}|\boldsymbol{v}^T\sigma_{\boldsymbol{\alpha}}(\boldsymbol{W}_0\widetilde{\boldsymbol{x}}_i)| \leq \varepsilon_0$, where $\varepsilon_0 = \sqrt{c_0}B(1+3\sqrt{\log n_{\mathcal{T}}})$

2. T'th iterate $\boldsymbol{\theta}_T$ obeys

$$\|\boldsymbol{W}_T - \tilde{\boldsymbol{W}}_\infty\|_F = \|\boldsymbol{\theta}_T - \tilde{\boldsymbol{\theta}}_\infty\|_{\ell_2} \le \varepsilon_1 = \frac{5}{2}\frac{\gamma}{\sqrt{c_0}B\,\|\boldsymbol{X}\|}\sqrt{n_{\mathcal{T}}} + 4\left(1 - \frac{1}{4}\eta c_0 \bar{\lambda}_0(\boldsymbol{X})\right)^t \frac{\sqrt{n_{\mathcal{T}}}}{\sqrt{c_0\bar{\lambda}_0(\boldsymbol{X})}}.$$

3. Rows are bounded via $\|\boldsymbol{W}_T - \boldsymbol{W}_0\|_{2,\infty} \le \sqrt{C_0/k}$ and $C_0 = \frac{32^2 B^2 \|\boldsymbol{X}\|^2}{c_0 \bar{\lambda}_0^2}$

4. At initialization, the network prediction is at most ε_2 i.e. $\|\boldsymbol{p}_{\boldsymbol{\alpha}}\|_{\ell_2} \le \varepsilon_2 = \sqrt{c_0}\sqrt{n_{\mathcal{T}}}B(1 + 3\sqrt{\log n_{\mathcal{T}}})$

5. Initial Jacobians obey $\frac{\boldsymbol{J}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\alpha}}^T}{c_0} \succeq \bar{\lambda}_0 \boldsymbol{I}_{n_{\mathcal{T}}}/2$.

6. Initial Jacobians obey $\|(\boldsymbol{J}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\alpha}}^T)^{-1} - \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\| \le \varepsilon_3 = \frac{2\gamma^2\bar{\lambda}_0^2}{512^2 c_0 B^6 \|\boldsymbol{X}\|^6}$.

Step 1.1: Bounding ε_1 and verifying the choice of T_0. In the second itemized condition of Step 1, we apply $\log$ on the second summand,

$$\begin{aligned}
\log\left(4\left(1 - \frac{1}{4}\eta c_0\bar{\lambda}_0(\boldsymbol{X})\right)^t \frac{\sqrt{n_{\mathcal{T}}}}{\sqrt{c_0\bar{\lambda}_0(\boldsymbol{X})}}\right) &= t\log\left(1 - \frac{1}{4}\eta c_0\bar{\lambda}_0(\boldsymbol{X})\right) + \log 4\frac{\sqrt{n_{\mathcal{T}}}}{\sqrt{c_0\bar{\lambda}_0(\boldsymbol{X})}} \\
&= t\log\left(1 - \frac{1}{4}\frac{1}{2c_0B^2\|\boldsymbol{X}\|^2}c_0\bar{\lambda}_0(\boldsymbol{X})\right) + \log(4\frac{\sqrt{n_{\mathcal{T}}}}{\sqrt{c_0\bar{\lambda}_0(\boldsymbol{X})}}) \\
&\le t\log\left(1 - \frac{\bar{\lambda}_0}{8B^2n_{\mathcal{T}}}\right) + \log(4\frac{\sqrt{n_{\mathcal{T}}}}{\sqrt{c_0\bar{\lambda}_0}}) \\
&\le t\left(-\frac{\bar{\lambda}_0}{8B^2n_{\mathcal{T}}}\right) + \log(4\frac{\sqrt{n_{\mathcal{T}}}}{\sqrt{c_0\bar{\lambda}_0}}).
\end{aligned}$$

Here, we hope to ensure that $\varepsilon_1 \le 5\frac{\gamma}{\sqrt{c_0}B\|\boldsymbol{X}\|}\sqrt{n_{\mathcal{T}}}$, Thus, we can bound t as following.

$$\begin{aligned}
t\left(-\frac{\bar{\lambda}_0}{8B^2n_{\mathcal{T}}}\right) + \log 4\frac{\sqrt{n_{\mathcal{T}}}}{\sqrt{c_0\bar{\lambda}_0}} &\le \log(\frac{5}{2}\frac{\gamma}{\sqrt{c_0}B\,\|\boldsymbol{X}\|}\sqrt{n_{\mathcal{T}}}) \\
\left(-\frac{\bar{\lambda}_0}{8B^2n_{\mathcal{T}}}\right)t &\le \log(\frac{5}{8}\frac{\gamma\sqrt{\bar{\lambda}_0}}{B\,\|\boldsymbol{X}\|}) \\
t &\ge \frac{8B^2n_{\mathcal{T}}}{\bar{\lambda}_0}\log(\frac{120B\sqrt{n_{\mathcal{T}}}}{\varepsilon\sqrt{\bar{\lambda}_0}}).
\end{aligned}$$

Thus, as long as[5]

$$T > T_0 = \frac{16B^2n_{\mathcal{T}}}{\bar{\lambda}_0}\log(\frac{120B\sqrt{n_{\mathcal{T}}}}{\varepsilon\sqrt{\bar{\lambda}_0}}) \propto \frac{B^2n_{\mathcal{T}}}{\bar{\lambda}_0}\log(\frac{B\sqrt{nt}}{\varepsilon\sqrt{\bar{\lambda}_0}}).$$

We have

$$\varepsilon_1 \le 5\frac{\gamma}{\sqrt{c_0}B\,\|\boldsymbol{X}\|}\sqrt{n_{\mathcal{T}}}.$$

Step 1.2: Applying $\varepsilon_0, \varepsilon_1, C_0, \varepsilon_2, \varepsilon_3$ to the error in Theorem 10. Considering Theorem 10, we have the following overall error

$$\text{error} = 3(\varepsilon_0 + \sqrt{c_0}BC_0/\sqrt{k} + \sqrt{c_0}B\varepsilon_1 + 2B\varepsilon_2/\sqrt{\bar{\lambda}_0}) + \sqrt{c_0}B\sqrt{\varepsilon_3}. \tag{J.3}$$

[5]Here, we choose the coefficient of T_0 to be 16 rather than 8 which will help in (J.5).

We will now show the values of k_0, c_0, T_0 ensure that error $< \varepsilon$. Specifically, plugging in the value of $\varepsilon_0, \varepsilon_1, C_0, \varepsilon_2, \varepsilon_3$ to (J.3), we have

$$\begin{aligned}\text{error} \leq & 3\sqrt{c_0}B(1+3\sqrt{\log n_{\mathcal{T}}}) + 3\frac{32^2B^3\left\|\boldsymbol{X}\right\|^2}{\sqrt{c_0}\bar{\lambda}_0^2\sqrt{k}} + 15\frac{\gamma}{\left\|\boldsymbol{X}\right\|}\sqrt{n_{\mathcal{T}}} \\ & + 6\frac{\sqrt{c_0}B^2\sqrt{n_{\mathcal{T}}}(1+3\sqrt{\log n_{\mathcal{T}}})}{\sqrt{\bar{\lambda}_0}} + \frac{\sqrt{2}\gamma\bar{\lambda}_0}{512B^2\left\|\boldsymbol{X}\right\|^3}. \end{aligned} \tag{J.4}$$

Step 2: Verifying c_0 and γ satisfies conditions. First, plugging in c_0 and $k \geq k_0$ into the second summand of (J.4), we find

$$\frac{B^3\left\|\boldsymbol{X}\right\|^2}{\sqrt{c_0}\bar{\lambda}_0^2\sqrt{k}} \leq \frac{B^3\left\|\boldsymbol{X}\right\|^2}{\bar{\lambda}_0^2}\frac{B^2\sqrt{n_{\mathcal{T}}}(1+3\sqrt{\log n_{\mathcal{T}}})}{\varepsilon\sqrt{\bar{\lambda}_0}}\frac{\varepsilon^2\bar{\lambda}_0^4}{B^8n_{\mathcal{T}}^4\sqrt{h\log n_{\mathcal{T}}}} \leq c\varepsilon\frac{\bar{\lambda}_0^{3/2}}{B^3n_{\mathcal{T}}^{3/2}\sqrt{h}} \leq c\varepsilon.$$

where $c > 0$ can be made arbitrarily small by enlarging the constant multiplier of the k_0 lower bound on the width k.

Setting $\gamma = \frac{\varepsilon\|\boldsymbol{X}\|}{75\sqrt{n_{\mathcal{T}}}}$ and $c_0 = (\frac{\varepsilon\sqrt{\bar{\lambda}_0}}{30B^2\sqrt{n_{\mathcal{T}}}(1+3\sqrt{\log n_{\mathcal{T}}})})^2$ in (J.4) and using the $c\varepsilon$ bound above (set $c < (15\cdot 32^2)^{-1}$), we find that each term in (J.4) is less to or equal than $\frac{\varepsilon}{5}$.

$$\begin{aligned}\text{error} &\leq \frac{\varepsilon\sqrt{\bar{\lambda}_0}}{10B\sqrt{n_{\mathcal{T}}}} + 3\cdot 32^2c\varepsilon + \frac{\varepsilon}{5} + \frac{\varepsilon}{5} + \frac{\sqrt{2}\varepsilon\bar{\lambda}_0}{38400B^2\|\boldsymbol{X}\|^2\sqrt{n_{\mathcal{T}}}}. \\ \text{error} &\leq 5\frac{\varepsilon}{5} = \varepsilon.\end{aligned}$$

Combine this with Theorem 10, fix $M = 120B^4\bar{\lambda}_0^{-2}\Gamma(n_{\mathcal{T}}^2+n_{\mathcal{V}}^2)\|\boldsymbol{y}\|_{\ell_2}$, with probability $1 - 4e^{-t} - \frac{4}{n_{\mathcal{T}}^3} - 4e^{-10h}$, δ-approximate NAS output obeys

$$\mathcal{L}(f_{\widehat{\boldsymbol{\alpha}}}^{\mathcal{T}}) \leq \min_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}} 2\sqrt{c_0}B\sqrt{\frac{\boldsymbol{y}^T\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}}{n_{\mathcal{T}}}} + C\sqrt{\frac{h\log(M)+t}{n_{\mathcal{V}}}} + \text{error} + \delta,$$

with error $\leq \varepsilon$. This concludes the proof of the main claim (J.1).

Finally, to conclude with the claim on the training risk satisfying $\hat{\mathcal{L}}_{\mathcal{T}}^{0-1}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) \leq \varepsilon$. Here, we will employ the first statement (I.2) of Theorem 11. For all architectures $\boldsymbol{\alpha}$, using the fact that the least-squares dominate the 0-1 loss, for $T \geq T_0$, using $B^2n_{\mathcal{T}} \geq \lambda_0$, we can write

$$\begin{aligned}\hat{\mathcal{L}}^{0-1}(f_{\boldsymbol{\alpha}}^{\mathcal{T}}) &\leq \frac{1}{n}\|f(\boldsymbol{W}_T) - \boldsymbol{y}\|_{\ell_2}^2 \leq 4\left(1 - \eta\frac{c_0\bar{\lambda}_0(\boldsymbol{X})}{8}\right)^T \\ &\leq 4\exp(-\eta T_0\frac{c_0\bar{\lambda}_0(\boldsymbol{X})}{8}) \\ &\leq 4\exp(-\frac{1}{2c_0B^2\|\boldsymbol{X}\|^2}\frac{c_0\bar{\lambda}_0}{8}\frac{16B^2n_{\mathcal{T}}}{\bar{\lambda}_0}\log(\frac{120B\sqrt{n_{\mathcal{T}}}}{\varepsilon\sqrt{\bar{\lambda}_0}})) \\ &\leq 4\exp(-\frac{n_{\mathcal{T}}}{\|\boldsymbol{X}\|^2}\log(\frac{120B\sqrt{n_{\mathcal{T}}}}{\varepsilon\sqrt{\bar{\lambda}_0}})) \\ &\leq 4\times\frac{\varepsilon\sqrt{\bar{\lambda}_0}}{120B\sqrt{n_{\mathcal{T}}}} \leq \varepsilon.\end{aligned} \tag{J.5}$$