Robust Fine-Tuning of Deep Neural Networks with Hessian-based Generalization Guarantees

Haotian Ju [*1] **Dongyue Li** [*1] **Hongyang R. Zhang** [1]

Abstract

We consider transfer learning approaches that fine-tune a pretrained deep neural network on a target task. We investigate generalization properties of fine-tuning to understand the problem of overfitting, which often happens in practice. Previous works have shown that constraining the distance from the initialization of fine-tuning improves generalization. Using a PAC-Bayesian analysis, we observe that besides distance from initialization, Hessians affect generalization through the noise stability of deep neural networks against noise injections. Motivated by the observation, we develop Hessian distance-based generalization bounds for a wide range of fine-tuning methods. Next, we investigate the robustness of fine-tuning with noisy labels. We design an algorithm that incorporates consistent losses and distance-based regularization for fine-tuning. Additionally, we prove a generalization error bound of our algorithm under class conditional independent noise in the training dataset labels. We perform a detailed empirical study of our algorithm on various noisy environments and architectures. For example, on six image classification tasks whose training labels are generated with programmatic labeling, we show a 3.26% accuracy improvement over prior methods. Meanwhile, the Hessian distance measure of the fine-tuned network using our algorithm decreases by six times more than existing approaches.

1. Introduction

Fine-tuning a pretrained neural network is a commonly used approach to perform transfer learning. With the emergence of foundation models [HZR+16; DCL+18; BMR+20], fine-tuning has become an efficient approach in many settings, including multitask learning [Rud17], meta-learning [FAL17], and zero-shot learning [WIL+22]. While fine-tuning approaches improve over supervised learning in many cases [LHC+19; RSR+20; VWM+20], their test performance are often limited by overfitting [BHG+17; ZBH+17; LNR+20]. Understanding the cause of overfitting is a challenge as addressing the problem requires a precise measure of generalization in deep neural networks. In this work, we analyze the generalization error of fine-tuned deep models using a PAC-Bayesian approach. Based by the analysis, we design an algorithm to improve the robustness of fine-tuning with limited and noisy labels.

There is a large body of work concerning generalization in deep neural networks, whereas less is known for fine-tuned models [NSZ20]. A central result of the recent literature is to show norm or margin based bounds that improve over classical capacity bounds [BFT17; NBS18]. These results can be applied to the fine-tuning setting, following a "distance from initialization" perspective [NK19; LS20]. Gouk et al. [GHP21] and Li and Zhang [LZ21] show generalization bounds depending on various normed distance between fine-tuned and initialization models. These results highlight that distance from initialization crucially affect generalization for fine-tuning, and inform the design of distance-based regularization to mitigate overfitting, due to fine-tuning a large model on a small training dataset.

Our motivating observation is that in addition to distance from initialization, Hessians crucially affect generalization. We consider a PAC-Bayesian analysis approach. Previous work has explored the noise stability properties of deep networks [AGN+18] using Lipschitzness of the activation functions. We instead quantify the loss stability of a deep model against noise perturbations using Hessians. This is a practical approach as computational frameworks for deepnet Hessians are developed recently [YGK+20]. We show that incorporating Hessians into the distance-based measure accurately correlates with the generalization error of fine-tuning. See Figure 1 for an illustration.

Results. Our theoretical contribution is to develop generalization bounds for fine-tuned multilayer feedforward net-

[*]Equal contribution [1]Northeastern University, Boston MA, United States. Correspondence to: All authors <{h.ju, li.dongyu, ho.zhang}@northeastern.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

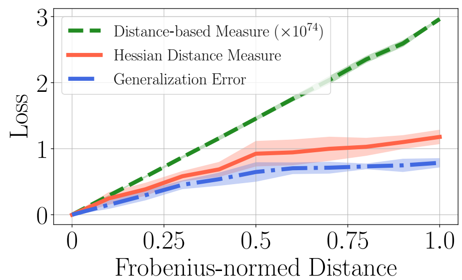


Figure 1: By incorporating Hessians with distance from initialization, the Hessian distance measure better correlates with the generalization error of fine-tuned models.

Method	Generalization bound
Fine-tuning	$\frac{\sum_{i=1}^{L}\sqrt{v_i^\top \mathbf{H}_i^+ v_i}}{\sqrt{n}}$
Distance-based reg.	$\frac{\sum_{i=1}^{L}\sqrt{\mathrm{Tr}[\mathbf{H}_i^+]\cdot\|v_i\|^2}}{\sqrt{n}}$
Consistent loss w/ reg.	$\frac{\sum_{i=1}^{L}\sqrt{\|(F^{-1})^\top\|_{1,\infty}\cdot\lvert\mathrm{Tr}[\mathbf{H}_i]\rvert\cdot\|v_i\|^2}}{\sqrt{n}}$

Table 1: Summary of our theoretical results: By measuring Hessians, our technique yields practical bounds for various fine-tuning methods. See also Section 2.1 for the definition of the notations.

works using Hessians. Our result applies to a wide range of fine-tuning approaches, including vanilla fine-tuning (with or without early stopping), distance-based regularization, and fine-tuning with weighted losses. To describe our result, we introduce a few notations. Let f_W be an L layer feedforward neural network with weight matrices $\{W_i\}$, for i from 1 to L. Given a pretrained initialization $f_{W^{(s)}}$ with weight matrices $W_1^{(s)}, \dots, W_L^{(s)}$, let v_i be a flatten vector of $W_i - W_i^{(s)}$. For every $i = 1, \dots, L$, let $\mathbf{H}_i$ be the *loss* Hessian over W_i. For technical reasons, denote $\mathbf{H}_i^+$ as a truncation of $\mathbf{H}_i$ with only nonnegative eigenvalues. We prove that on a training dataset of size n sampled from a distribution $\mathcal{D}$, with probability at least 0.99, the generalization error of a fine-tuned L-layer network f_W scales as

$$\mathrm{O}\left(\frac{\sum_{i=1}^{L}\sqrt{\max_{(x,y)\sim\mathcal{D}} v_i^\top \mathbf{H}_i^+[\ell(f_W(x),y)]v_i}}{\sqrt{n}}\right). \quad (1)$$

See Theorem 2.1 for the complete statement. Based on this result, one can derive a generalization bound for distance-based regularization of fine-tuning, assuming constrains on the norm of v_i for every i.

Next, we consider the robustness of fine-tuning against label noise, by considering a class conditional label noise setting [NDR+13]. Given training labels that are independently flipped with some probability according to a confusion matrix F, we extend our result by incorporating statistically-consistent losses. Table 1 summarizes our theoretical results. Compared to the result of Arora et al. [AGN+18], our proof involves a perturbation analysis of the loss Hessian, which may be of independent interest.

Our analysis leads to the design of a new algorithm with practical interest. We propose to *incorporate consistent losses with distance-based regularization* to improve the robustness of fine-tuning with a small amount of labels. We evaluate our algorithm on various tasks with noisy labels. First, we show that under class conditional label noise, our algorithm performs robustly compared to the baselines, even when the noise rate is as high as 60%. Second, on six image classification tasks whose labels are generated by programmatic labeling [MCS+21], our approach improves the prediction accuracy of fine-tuned models by 3.26% on average compared to existing approaches. Third, we extend our result for fine-tuning vision transformers and language models on the same six tasks, showing similar results. In ablation studies, we find that our algorithm reduces the Hessian distance measure by six times more than previous fine-tuning methods.

1.1. Related Work

The PAC-Bayesian analysis is an approach to deriving the generalization error of a machine learning model. It was introduced by McAllester [McA99b] and is used to analyze generalization in model averaging [McA99a] and co-training [DLM01]. In recent years, this approach has been used to show generalization bounds for deep networks [AGN+18; NK18; NBS18]. Besides, the PAC-Bayesian analysis has been used to derive the generalization error of black-box models, including graph neural networks [LUZ21] and data augmentation [CPD+21]. Empirically, PAC-Bayesian bounds can be optimized to achieve nonvacuous generalization bounds in real-world settings [DR17; ZVA+19; DHG+21]. We refer interested readers to the surveys of Guedj [Gue19] and Alquier [Alq21] for references.

Recent work finds that generalization measures that correlate with empirical performance rely on data-dependent measures [DR17; FKM+21]. Jiang et al. [JNM+20] show that sharpness-based generalization measures correlate with empirical performance better than norm or margin bounds. Tsuzuku et al. [TSS20] propose a matrix-normalized sharpness measure. These works focus on the numerical optimization of the PAC-Bayes bound. To our knowledge, provable bounds that capture notions such as sharpness have not been developed. Our Hessian distance-based measure is different from sharpness-based measures since the Hessian measures average perturbation whereas sharpness measures adversarial perturbation. It is conceivable that the techniques we have developed might be helpful for understanding sharpness. This is an interesting direction for future work. Our experiments with multiple datasets and architectures suggest that the Hessian distance measure is a practical tool for measuring generalization for fine-tuning.

While our work focuses on the PAC-Bayesian approach to understand generalization for fine-tuning, it is worth mentioning that there are many theoretical works concerning generalization in the broader space of transfer learning. The seminal work of Ben-David et al. [BBC+10] (and the earlier work of Crammer et al. [CKW08]) derives generalization bounds for learning from multiple sources. Recent works have proposed new theoretical frameworks and techniques to explain transfer including transfer exponent [HK19], task diversity [TJJ20], random matrix theory [WZR20; YZW+21], and minimax estimators [LHL21]. Shachaf et al. [SBG21] consider a deep linear network and the impact of the similarity between the source (on which a model is pretrained) and target domains on transfer.

There is a significant body of work on learning with noisy labels, so it is beyond the scope of this work to provide a comprehensive discussion. See, e.g., Song et al. [SKP+22] for a recent survey. Some works that are related to ours involve the design of robust losses [LT15; LNR+20], the estimation of the confusion matrix [PRK+17; XLW+19; YLH+20; ZLA21], the implicit regularization of early stopping, [YRC07; LSO20], and label smoothing [LBM+20]. The work most related to ours within this literature is Natarajan et al. [NDR+13], which provides Rademacher complexity-based learning bounds under class conditional label noise. Our work differs from this work in two aspects. First, we use the PAC-Bayesian approach to study generalization. Second, we consider deep neural networks as the hypothesis class. Our work highlights the importance of explicit regularization for improving generalization given limited and noisy labels.

Organization. The rest of this paper is organized as follows. In Section 2, we first introduce the problem setup for fine-tuning with a deep model. Then, we state our main results including PAC-Bayesian bounds for fine-tuned models using Hessians. In Section 3, we describe experiments to validate the algorithmic implications of our PAC-Bayesian analysis. In Section 4, we summarize this paper and discuss several questions for future work. We present a detailed proof of our results in Appendix A. In Appendix B, we fill in missing details in the implementation of the experiments.

2. Generalization Bounds for Fine-Tuning using Hessians

This section presents our approach to understand generalization in fine-tuning. After setting up the problem formally, we first present PAC-Bayesian bounds for (vanilla) fine-tuning and distance-based regularization. Then, we consider the robustness of fine-tuning against label noise, and extend our result by incorporate consistent losses. Lastly, we give an overview of the proofs.

2.1. Problem setup

Suppose we want to solve a target task given a training dataset of size n. Denote the feature vectors and labels as x_i and y_i, for every $i = 1, \dots, n$, in which x_i is d-dimensional and y_i is a class label between 1 to k. Assume that the training examples are drawn independently from an unknown distribution $\mathcal{D}$. Let $\mathcal{X}$ be the support of the features vectors of $\mathcal{D}$.

Given a pretrained L-layer feedforward network with weight matrices $\hat{W}_i^{(s)}$, for $i = 1, \dots, L$, we fine-tune the weights to solve the target task. Let f_W be an L-layer network initialized with $\hat{W}^{(s)}$. The dimension of layer i, W_i, is d_i by d_{i-1}. Let d_i be the output dimension of layer i. Thus, d_0 is equal to the input dimension d and d_L is equal to the output dimension k. Let $\phi_i(\cdot)$ be the activation function of layer i. Given a feature vector $x \in \mathcal{X}$, the output of f_W is

$$f_W(x) = \phi_L\left(W_L \cdot \phi_{L-1}\Big(W_{L-1} \cdots \phi_1(W_1 \cdot x)\Big)\right). \qquad (2)$$

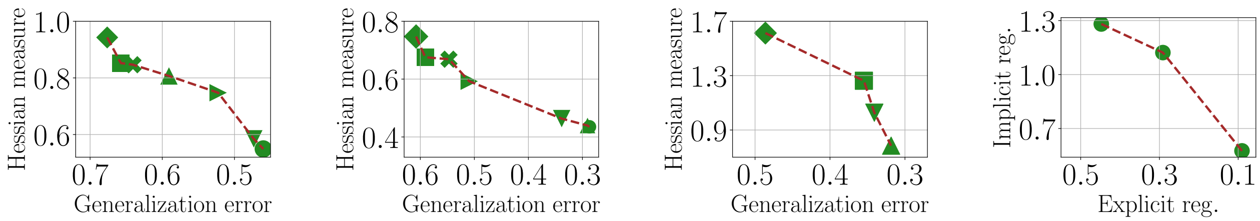
Given a loss function $\ell : \mathbb{R}^k \times \{1, \dots, k\} \to \mathbb{R}$, let $\ell(f_W(x), y)$ be the loss of f_W. The empirical loss of f_W, denoted by $\hat{\mathcal{L}}(f_W)$, is the loss of f_W averaged among the training examples. The expected loss of f_W, denoted by $\mathcal{L}(f_W)$, is the expectation of $\ell(f_W(x), y)$ over x sampled from $\mathcal{D}$ with label y. The generalization error of f_W is defined as its expected loss minus its empirical loss.

Notations. For any vector v, let $\|v\|$ be the Euclidean norm of v. For any matrix $X \in \mathbb{R}^{m\times n}$, let $\|X\|_F$ be the Frobenius norm and $\|X\|_2$ be the spectral norm of X. Let $\|X\|_{1,\infty}$ be defined as $\max_{1\le j\le n} \sum_{i=1}^m |X_{i,j}|$. If X is a squared matrix, then the trace of X, $\text{Tr}[X]$, is equal to the sum of X's diagonal entries. For two matrices X and Y with the same dimension, let $\langle X, Y\rangle = \text{Tr}\left[X^\top Y\right]$ be the matrix inner product of X and Y. For every i from 1 to L, let $\mathbf{H}_i[\ell(f_W(x), y)]$ be a $d_i d_{i-1}$ by $d_i d_{i-1}$ Hessian matrix of the loss $\ell(f_W(x), y)$ over W_i. Given an eigendecomposition of the Hessian matrix, $UDU^\top$, let $\mathbf{H}_i^+[\ell(f_W(x), y)] = U\max(D, 0)U^\top$ be a truncation into the positive eigen space. Let v_i be a flatten vector of the matrix $W_i - \hat{W}_i^{(s)}$.

For two functions $f(n)$ and $g(n)$, we write $g(n) = O(f(n))$ if there exists a fixed value C that does not grow with n such that $g(n) \le C \cdot f(n)$ when n is large enough.

2.2. PAC-Bayesian bounds for fine-tuned models

To better understand what impacts the performance of fine-tuning, we investigate it with a PAC-Bayesian approach



(a) ResNet-50 (b) ResNet-50 (c) BERT-Base (d) Average $\mathcal{D}_i$

Figure 2: 2a, 2b, 2c: The Hessian measures accurately correlate with empirical generalization errors for seven fine-tuning methods. 2d: $\mathcal{D}_i$ is smaller for fine-tuning with explicit distance-based regularization compared to implicit early-stopped regularization. Each dot represents one model. See Section B for the experimental setup.

[AGN+18; NBS18; LUZ21]. We consider a prior distribution $\mathcal{P}$ defined as a *noisy perturbation* of the *pretrained* weight matrices $\hat{W}^{(s)}$. We also consider a posterior distribution $\mathcal{Q}$ defined as a *noisy perturbation* of the *fine-tuned* weights $\hat{W}$. From a PAC-Bayesian perspective, the generalization performance of a deep network depends on how such noisy perturbations affect its predictions.

Previous work [AGN+18] has explored the noise stability properties of deep networks and showed how to derive stronger generalization bounds that depend on properties of the network. Let $\ell_{\mathcal{Q}}(f_W(x), y)$ be the loss of $f_W(x)$, after a noisy perturbation following $\mathcal{Q}$. Denote $\ell_{\mathcal{Q}}(f_W(x), y)$ minus $\ell(f_W(x), y)$ as $\mathcal{I}(f_W(x), y)$: A lower value of $\mathcal{I}(f_W(x), y)$ means that the network is more error resilient. The result of Arora et al. [AGN+18] shows how to quantify the error resilience of f_W using Lipschitzness and smoothness of the activation functions across different layers. However, there is still a large gap between the generalization bounds achieved by Arora et al. [AGN+18] and the empirical generalization errors. Thus, a natural question is whether one can achieve generalization bounds that better capture empirical performance.

The key idea of our approach is to measure generalization using the Hessian of $\ell(f_W(x), y)$. From a technical perspective, this goes beyond the previous works by leveraging Lipschitzness and smoothness of the derivatives of f_W. Besides, computational frameworks [YGK+20] are developed for efficiently computing deep net Hessians, such as Hessian vector products [SBL16; Pap18; GKX19; Pap19]. Our central observation is that incorporating Hessians with distance from initialization leads to provable bounds that accurately correlate with empirical performance.

How can we connect Hessians with generalization? Consider a Taylor's expansion of the loss $\ell(f_{W+U}(x), y)$ at W, for some small perturbation U over W. Suppose U has mean zero and covariance Σ. Then, $\mathcal{I}(f_W(x), y)$ is equal to $\langle \Sigma, \mathbf{H}[\ell(f_W(x), y)] \rangle$ plus higher-order expansion terms. While this expansion applies to every sample, the next question is how to argue the uniform convergence of the Hessians. This requires a perturbation analysis of the Hessian of all layers, but can be achieved assuming Lipschitness and smoothness of the derivatives of the activation functions. We state our result formally below.

Theorem 2.1. *Assume the activation functions $\phi_i(\cdot)$ for all $i = 1, \dots, L$ and the loss function $\ell(\cdot, \cdot)$ are all twice-differentiable, and their first-order and second-order derivatives are all Lipschitz-continuous. Suppose the loss function $\ell(x, y)$ is bounded by a fixed value C for any $x \in \mathcal{X}$ with class label y. Given an L-layer network $f_{\hat{W}}$, with probability at least 0.99, for any fixed ϵ close to zero, we have*

$$\mathcal{L}(f_{\hat{W}}) \leq (1+\epsilon)\hat{\mathcal{L}}(f_{\hat{W}}) + \frac{(1+\epsilon)\sqrt{C}\sum_{i=1}^{L}\sqrt{\mathcal{H}_i}}{\sqrt{n}} + \xi, \quad (3)$$

where $\mathcal{H}_i$ is any value greater than $\max_{(x,y)\in\mathcal{D}} v_i^\top \mathbf{H}_i^+[\ell(f_{\hat{W}}(x), y)] v_i$, for all $i = 1, \dots, L$, and $\xi = O(n^{-3/4})$ represents an error term from the Taylor's expansion.

Theorem 2.1 applies to vanilla fine-tuning (with or without early stopping). A corollary of this result can be derived for distance-based regularization, which restricts the distance between W_i and $\hat{W}_i^{(s)}$ for every layer [LXW+18; LGD18; GHP21]:

$$\|W_i - \hat{W}_i^{(s)}\|_F \leq \alpha_i, \ \forall\, i = 1, \dots, L. \quad (4)$$

Since $\mathcal{H}_i \leq \alpha_i^2 \max_{(x,y)\in\mathcal{X}} \mathrm{Tr}[\mathbf{H}_i^+(\ell(f_{\hat{W}}(x), y))]$, we thus obtain a PAC-Bayes bound for distance-based regularization of fine-tuning.

We demonstrate that the generalization bound of Theorem 2.1 accurately correlate with empirical generalization errors. We experiment with seven methods, including fine-tuning with and without early stopping, distance-based regularization, label smoothing, mixup, etc. Figure 2 shows that the Hessian measure (i.e. second part of equation (3)) correlate with the generalization errors of these methods. Second, we plot the value of $\mathcal{D}_i$ (averaged over all layers) between fine-tuning with implicit (early stopping) regularization and explicit (distance-based) regularization. We fine that $\mathcal{D}_i$ is smaller for explicit regularization compared to implicit regularization.

Algorithm 1 Consistent loss reweighting with layerwise projection

Input: Training dataset $\{(x_i, \tilde{y}_i)\}_{i=1}^n$ with input feature x_i and noisy label $\tilde{y}_i$, for $i = 1, \dots, n$
Require: Pretrained network $f_{\hat{W}^{(s)}}$, layerwise distance α_i for $i = 1, \dots, L$, number of epochs T, learning rate η, and confusion matrix F
Output: A trained model $f_{W^{(T)}}$

1: Let $\Lambda = F^{-1}$
2: At $t = 0$, initialize model parameters with pretrained weight matrices $W^{(0)} = \hat{W}^{(s)}$
3: **while** $0 < t < T$ **do**
4: Let $\bar{\mathcal{L}}(f_{W^{(t-1)}}) = \frac{1}{n}\sum_{i=1}^n \sum_{c=1}^k \Lambda_{\tilde{y}_i,c} \cdot \ell(f_{W^{(t-1)}}(x_i), c)$
5: Update $W^{(t)} \leftarrow W^{(t-1)} - \eta \cdot \nabla\bar{\mathcal{L}}(f_{W^{(t-1)}})$ using SGD
6: Project $W_i^{(t)}$ in constrained region: $W_i^{(t)} \leftarrow \min\left(1, \frac{\alpha_i}{\|W_i^{(t)} - W_i^{(0)}\|_F}\right)\left(W_i^{(t)} - W_i^{(0)}\right) + W_i^{(0)}, \forall i = 1, \dots, L$
7: **end while**

2.3. Incorporating consistent losses with distance-based regularization

Next, we consider the robustness of fine-tuning with noisy labels. We consider a random classification noise model where in the training dataset, each label y_i is independently flipped to $1, \dots, k$ with some probability [AL88]. Denote the noisy label by $\tilde{y}_i$. With class conditional label noise [NDR+17; KLA18], $\tilde{y}$ is equal to z with probability $F_{y,z}$, for any $z = 1, \dots, k$, where F is a k by k confusion matrix.

Previous works [NDR+13; PRK+17] have suggested minimizing statistically-consistent losses for learning with noisy labels. Let $\bar{\ell}$ be a weighted loss parametrized by a k by k weight matrix Λ,

$$\bar{\ell}(f_W(x), \tilde{y}) = \sum_{i=1}^{k} \Lambda_{\tilde{y},i} \cdot \ell(f_W(x), i). \tag{5}$$

It is known that for $\Lambda = F^{-1}$ (recall that F is the confusion matrix), the weighted loss $\bar{\ell}$ is unbiased in the sense that the expectation of $\bar{\ell}(f_W(x), \tilde{y})$ over the labeling noise is $\ell(f_W(x), y)$:

$$\mathop{\mathbb{E}}_{\tilde{y}}\left[\bar{\ell}(f_W(x), \tilde{y}) \mid y\right] = \ell(f_W(x), y). \tag{6}$$

See Lemma 1 in Natarajan et al. [NDR+13] for the binary setting and Theorem 1 in Patrini et al. [PRK+17] for the multiclass setting. As a result, a natural approach to design theoretically grounded fine-tuning algorithm is to minimize weighted losses with $\Lambda = F^{-1}$. However, this approach does not take overfitting into consideration.

Our proposed approach involves incorporating consistent losses with distance-based regularization. Let $\bar{\mathcal{L}}(f_W)$ be the average of the weighted loss $\bar{\ell}(f_W(x_i), \tilde{y}_i)$ over $i = 1, \dots, n$. We extend Theorem 2.1 to this setting as follows.

Theorem 2.2. *Assume the activation functions $\{\phi_i(\cdot)\}_{i=1}^L$ and the loss function $\ell(\cdot,\cdot)$ satisfy the assumptions stated in Theorem 2.1. Suppose the noisy labels are independent of the feature vector conditional on the class label. Suppose the loss function $\ell(x, z)$ is bounded by C for any $x \in \mathcal{X}$ and any $z \in \{1, \dots, k\}$. Let $\Lambda = F^{-1}$. With probability 0.99, for any fixed ϵ close to zero, we have*

$$\mathcal{L}(f_{\hat{W}}) \leq (1+\epsilon)\bar{\mathcal{L}}(f_{\hat{W}}) + \frac{(1+\epsilon)\sqrt{C\,\|\Lambda^\top\|_{1,\infty}}\sum_{i=1}^L \sqrt{\alpha_i^2 \mathscr{H}_i}}{\sqrt{n}} + \xi,$$

where α_i is any value greater than $\|\hat{W}_i - \hat{W}_i^{(s)}\|_F$, for any $i = 1, \dots, L$, $\xi = O\big(n^{-3/4}\big)$ represents an error term from the Taylor's expansion, and

$$\mathscr{H}_i = \max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left|\mathrm{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x), y))]\right|.$$

Theorem 2.2 shows that combining consistent losses and leads to a method that is provably robust under class conditional label noise. Notice that this result uses the trace of the Hessian $\mathbf{H}_i$ as opposed to the truncated Hessian $\mathbf{H}_i^+$. The reason is that the reweighing matrix Λ might include negative coefficients, and $\mathbf{H}_i^+$ deals with this problem. Based on the theory, we propose Algorithm 1, which instantiates the idea. In Section 3, we evaluate the robustness of Algorithm 1 for both image and text classification tasks.

2.4. Proof overview

We give an overview of the proof of Theorems 2.1 and 2.2. First, we show that the noise stability of an L-layer network f_W admits a layerwise Hessian approximation. Let $U \sim \mathcal{Q}$ be a random variable drawn from a posterior distribution $\mathcal{Q}$. We are interested in the perturbed loss, $\ell_{\mathcal{Q}}(f_U(x), y)$, which is the expectation of $\ell(f_U(x), y)$ over U.

Lemma 2.3. *In the setting of Theorem 2.1, for any $i = 1, 2, \cdots, L$, let $U_i \in \mathbb{R}^{d_i d_{i-1}}$ be a random vector sampled from a Gaussian distribution with mean zero and variance Σ_i. Let the posterior distribution $\mathcal{Q}$ be centered at W_i and perturbed with an appropriately reshaped U_i at every layer. Then, there exists a fixed value $C_1 > 0$ that does not*

Table 2: Showing the relative residual sum of squares (RSS) of the Hessian approximation (7) under isotropic noise perturbation of variance σ^2. The noise stability results are estimated by the average of 500 noise injections. All the experiments are done with the CIFAR-100 dataset.

σ	ResNet-18		ResNet-50	
	Noise Stability	Hessian Approx.	Noise Stability	Hessian Approx.
0.01	0.86 ± 0.19	1.07	0.39 ± 0.10	1.24
0.011	1.13 ± 0.24	1.29	1.12 ± 0.20	1.50
0.012	1.45 ± 0.29	1.54	1.56 ± 0.26	1.79
0.013	1.82 ± 0.35	1.80	2.10 ± 0.33	2.10
0.014	2.23 ± 0.40	2.09	2.71 ± 0.38	2.43
0.015	2.65 ± 0.43	2.34	3.30 ± 0.40	2.79
0.016	3.07 ± 0.45	2.73	3.80 ± 0.40	3.18
0.017	3.47 ± 0.47	3.08	4.17 ± 0.40	3.59
0.018	3.84 ± 0.49	3.46	4.60 ± 0.44	4.02
0.019	4.15 ± 0.51	3.85	4.77 ± 0.44	4.48
0.020	4.43 ± 0.55	4.27	5.03 ± 0.82	4.97
Relative RSS	**0.75%**		**2.98%**	

grow with n, such that for any $x \in \mathcal{X}$ and $y \in \{1, \dots, k\}$, $\ell_{\mathcal{Q}}(f_W(x), y) - \ell(f_W(x), y)$ is upper bounded by

$$\sum_{i=1}^{L} \left(\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_W(x), y)] \right\rangle + C_1 \left\|\Sigma_i\right\|_F^{3/2} \right). \quad (7)$$

See Appendix A.3 for the proof of Lemma 2.3. Interestingly, we find that the Hessian estimate is remarkably accurate in practice. We use ResNet-18 and ResNet-50 fine-tuned on the CIFAR-100 dataset. We estimate the noise stability in equation (7) by randomly sampling 500 isotropic Gaussian perturbations and average the perturbed losses. Then, we compute the Hessian term in right hand side of equation (7) by measuring the traces of the loss's Hessian at each layer of the network and average over the same data samples. Table 2 shows the result. The relative error of the Hessian estimate is within 3% on ResNet.

Based on the Hessian approximation, consider applying a PAC-Bayes bound over the L-layer network f_W (see also Catoni [Cat07], Guedj [Gue19], and Dziugaite et al. [DHG+21]). The KL divergence between the prior distribution and the posterior distribution is equal to

$$\sum_{i=1}^{L} \left\langle \Sigma_i^{-1}, v_i v_i^\top \right\rangle. \quad (8)$$

Minimizing the sum of the Hessian estimate and the above KL divergence (8) in the PAC-Bayes bound leads to a covariance matrix per layer, which depends on v_i and $\mathbf{H}_i^+$:

$$\sum_{i=1}^{L} \left(\langle \Sigma_i, \mathbf{H}_i \rangle + \frac{1}{n} \langle \Sigma_i^{-1}, v_i v_i^\top \rangle \right) \leq \sum_{i=1}^{L} \left(\langle \Sigma, \mathbf{H}_i^+ \rangle + \frac{1}{n} \langle \Sigma_i^{-1}, v_i v_i^\top \rangle \right),$$

where the $1/n$ factor comes from the PAC-Bayes bound (cf. Theorem A.1). Thus, minimizing the above over the covariance matrices leads to a Hessian distance based generalization bound of Theorem 2.1.

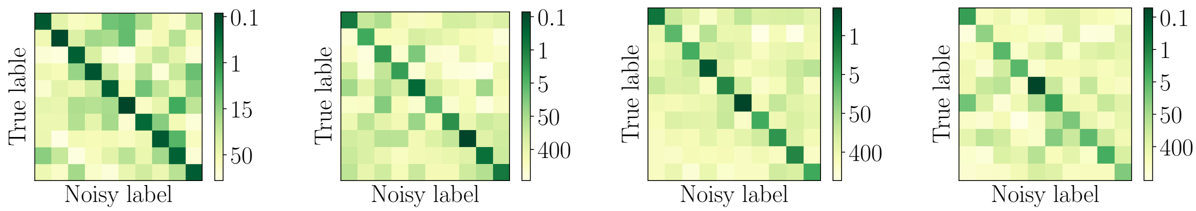
The above sketch highlights the crux of our result. The rigorous proof, on the other hand, is significantly more involved. One technical challenge is to argue the uniform convergence of the Hessian operator $\mathbf{H}_i$ between the empirical loss and the expected loss over $\mathcal{D}$. This requires a perturbation analysis of the Hessian, and uses the assumption that the second-order derivatives of the activation functions (and the loss) are Lipschitz-continuous.

Lemma 2.4. *In the setting of Theorem 2.1, there exist some fixed values C_2, C_3 that do not grow with n and $1/\delta$, such that with probability at least $1 - \delta$ over the randomness of the training set,*

$$\left\| \frac{1}{n} \sum_{j=1}^{n} \mathbf{H}_i[\ell(f_W(x_j), y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\mathbf{H}_i[\ell(f_W(x), y)]] \right\| \leq \frac{C_2 \sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}, \quad (9)$$

for any $i = 1, \dots, L$, where $C_2 = 4\mathcal{H}_i \sqrt{\sum_{i=1}^{L} d_i d_{i-1}}$, $C_3 = 5L\left(\sqrt{\sum_{i=1}^{L} \|W_i\|_F^2}\right) G/\mathcal{H}_i$, and G is a fixed value that depends on $\prod_{i=1}^{L} \|W_i\|$ and the Lipschitness of the activation functions and their first-order and second-order derivatives (cf. equation (29)).

See Appendix A.4 for the proof of Lemma 2.4, which is based on an ϵ-net argument. The rest of the proof of Theorem 2.1 can be found in Appendix A.1.



(a) ResNet-18 on CIFAR-10 (b) ResNet-50 on CIFAR-10 (c) ResNet-18 on CIFAR-100 (d) ResNet-50 on CIFAR-100

Figure 3: Illustrating the intuition behind Theorem 2.2: With consistent losses, the noisy stability of the loss Hessian with noisy labels is equal to the noise stability of the loss Hessian with class labels in expectation. Thus, the unbiased weighted losses enjoy similar noise stability properties as training with class labels. Further, the unbiased loss favors stable models as the trace of the Hessian is smallest when $\tilde{y} = y$.

The key to handling noisy labels is a consistency condition for the weighted loss $\bar{\ell}(f_W(x), \tilde{y})$ with noisy label $\tilde{y}$:

$$\mathcal{L}(f_W) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{\tilde{y}|y} \left[\bar{\ell}(f_W(x), \tilde{y})\right].$$

Thus, consider a noisy data distribution $\tilde{D}$, with the label y of every x flipped independently to $\tilde{y}$. Our key observation is that $\bar{\ell}(f_W(x), \tilde{y})$ enjoys similar noise stability properties as $\ell(f_W(x), y)$. See Figure 3 for an illustration. The detailed proof of Theorem 2.2 can be found in Appendix A.2.

3. Experiments

This section conducts experiments to complement our theory. We evaluate the robustness of Algorithm 1 under various noisy environments and architectures. First, we validate that when the training labels are corrupted with class conditional label noise, our approach is robust even under up to 60% of label noise. Second, on six weakly-supervised image classification tasks, our algorithm improves the prediction accuracy over prior approaches by 3.26% on average. Lastly, we present a detailed analysis that validates the theory, including comparing our generalization bound with previous results and comparing the Hessian distance measure for various fine-tuning methods.

3.1. Experimental setup

Datasets. We evaluate the robustness of our approach in image and text classification tasks. For image classification, we use six domains of object classification tasks from the DomainNet [PBX+19] dataset. We consider two kinds of label noise. First, we generate synthetic random noise by randomly flipping the labels of training samples uniformly with a given noise rate. Second, we create labels using weak supervision approaches [MCS+21]. The statistics of the six datasets are described in Table 6 from Section B.2. We refer the reader to Mazzetto et al. [MCS+21] for more details.

For text classification, we use the MRPC dataset from the GLUE benchmark [WSM+18], which is to predict whether two sentences are semantically equivalent.

Models. In fine-tuning from noisy labels, we use pretrained ResNet-18 and ResNet-101 [HZR+16] models on image classification tasks and extend our results to Vision Transformer (ViT) model [DBK+20]. We use the RoBERTa-Base [LOG+19] model on text classification tasks.

Baselines. We show that our fine-tuning algorithm is competitive with or even outperforms baseline methods on datasets with noisy labels. We consider the following three kinds of baselines: (i) Regularization: Fine-tuning with early stop (Early stopping), label smoothing [MKH19], Mixup [ZCD+18; WZV+20], and Early-Learning Regularization (ELR) [LNR+20]; (iii) Self-training: FixMatch [SBL+20] and Self-Adaptive Training (SAT) [HZZ20]; (ii) Robust loss function: Active Passive Loss (APL) [MHW+20], Generalized Jensen-Shannon Divergence (GJS) [EA21], dual transition estimator (DualT) [YLH+20], Supervised Contrastive learning (SupCon) [GDC+20], and Sharpness-Aware Minimization (SAM) [FKM+21].

For the implementation of our algorithm, we use the confusion matrix F estimated by the method from the work of Yao et al. [YLH+20]. We use layerwise distances [LZ21] for applying the regularization constraints in Equation (4). We describe the implementation details and hyper-parameters in Section B.2.

3.2. Experimental results

Improved robustness against independent label noise. Our result in Section 2.3 illustrates why our algorithm might be expected to achieve provable robustness under class conditional label noise. To validate its effectiveness, we apply our algorithm in both image and text classification tasks with different levels of synthetic random labels. Table 3 reports the test accuracy.

Table 3: Test accuracy of fine-tuning with noisy labels. We consider two settings: one where the training labels are independently flipped to incorrect labels, and another where the training labels are created using programmatic labeling approaches. The reported results are averaged over 10 random seeds.

Independent label noise	Sketch, ResNet-18		Sketch, ResNet-101		MRPC, RoBERTa-Base	
	40%	60%	40%	60%	20%	40%
Early Stopping	72.41±3.53	53.84±3.09	77.14±3.09	61.39±1.28	81.39±0.73	66.05±0.63
Label Smooth [MKH19]	74.69±1.97	55.35±1.60	81.47±1.36	64.90±2.93	80.00±0.62	65.87±0.95
Mixup [ZCD+18]	70.65±1.85	58.49±3.25	76.04±2.29	60.12±2.37	80.62±0.14	68.37±1.33
FixMatch [SBL+20]	73.35±3.15	61.51±2.17	76.19±1.39	62.19±1.57	81.10±1.76	68.48±1.43
ELR [LNR+20]	74.29±2.52	63.14±2.05	80.00±1.45	64.35±3.98	82.78±0.86	67.86±1.44
APL [MHW+20]	75.63±1.81	64.69±2.72	78.69±2.45	64.82±2.41	80.49±0.24	66.49±0.93
SAT [HZZ20]	75.18±1.54	62.33±2.24	80.00±2.96	65.58±2.91	82.80±0.42	67.50±1.00
GJS [EA21]	73.22±2.34	59.63±5.15	77.14±2.46	63.27±2.37	81.42±1.03	67.57±1.53
DualT [YLH+20]	72.49±3.17	59.59±3.44	77.96±0.33	62.31±3.98	82.49±0.53	66.49±0.93
SupCon [GDC+20]	75.14±1.73	61.06±3.20	78.86±1.80	63.92±2.15	82.30±1.80	68.32±1.16
SAM [FKM+21]	77.63±2.16	64.53±2.84	80.57±2.70	67.27±2.39	82.61±0.91	69.06±1.41
Ours	**81.96±0.98**	**70.00±1.71**	**85.44±1.26**	**71.84±2.72**	**83.55±0.63**	**72.64±1.77**

Correlated label noise	DomainNet, ResNet-18					
	Clipart 41.47%	Infograph 63.29%	Painting 44.50%	Quickdraw 60.54%	Real 34.64%	Sketch 47.68%
Early Stopping	73.88±2.04	38.82±2.59	69.69±1.35	44.16±1.92	78.52±1.03	61.84±3.67
Label Smooth [MKH19]	74.56±2.30	38.40±2.67	70.76±1.74	46.50±3.03	81.39±0.93	62.29±2.48
Mixup [ZCD+18]	72.88±0.94	39.27±3.10	69.28±3.18	47.66±3.20	80.17±2.05	62.08±3.06
FixMatch [SBL+20]	77.04±2.52	41.95±1.52	73.31±2.10	48.74±2.08	86.33±2.54	64.61±3.28
ELR [LNR+20]	76.08±2.03	40.14±2.74	72.06±1.73	47.40±3.09	83.64±2.09	65.76±3.19
APL [MHW+20]	77.40±2.33	41.22±2.58	73.61±3.12	49.88±3.24	85.79±1.59	64.69±2.30
SAT [HZZ20]	75.24±2.79	39.58±1.47	70.69±2.69	48.18±2.95	81.90±1.07	65.39±2.77
GJS [EA21]	77.20±2.59	40.94±2.19	72.51±2.87	48.14±3.40	85.05±1.94	65.43±3.35
DualT [YLH+20]	75.24±2.02	38.75±2.12	70.27±2.24	46.62±3.16	83.33±3.01	65.47±1.91
SupCon [GDC+20]	76.56±3.53	40.38±1.94	72.51±2.45	49.20±2.63	81.87±0.84	65.67±2.90
SAM [FKM+21]	79.04±1.57	41.50±1.94	73.23±2.29	50.10±1.66	84.61±2.04	66.73±2.88
Ours	**83.28±1.64**	**43.38±2.45**	**76.32±1.08**	**50.32±2.74**	**92.36±0.78**	**66.86±3.29**

- First, we fine-tune ResNet-18 on the Sketch domain from the DomainNet dataset with 40% and 60% synthetic noise. We observe that our algorithm improves upon the baseline methods by **6.43%** on average. The results corroborate our theory that provides generalization guarantees for our algorithm.
- Second, to evaluate our algorithm with deeper models, we apply ResNet-101 on the same label noise settings. We observe a higher average performance with ResNet-101, and our algorithm achieves a similar performance boost of **6.03%** on average over the previous methods.
- Third, we fine-tune RoBERTa-Base model on MRPC dataset with 20% and 40% synthetic noise. Our algorithm outperforms previous methods by **2.91%** on average. These results show that our algorithm is robust to label noise across various tasks and architectures.

Improved robustness against correlated label noise. Next, we show that our algorithm can also achieve competitive performance under real-world label noise. We evaluate our algorithm on weakly-supervised image classification tasks. We fine-tune ResNet-18 on the six domains from the DomainNet dataset and report the test accuracy in Table 3. We find that Algorithm 1 outperforms previous methods by **3.26%**, averaged over the six tasks.

Extension to other architectures. Lastly, we expect our results under correlated label noise to hold with various model architectures. As an extension of the above experiment, we fine-tune Vision Transformer [DBK+20] models on the same set of tasks. Table 4 reports the results on the Clipart and Sketch domain (the results for the other domains are similar and thus omitted). We notice that using ViT as the base model boosts the performance across all approaches significantly (e.g., over 4% for early stopping). We find that our approach still outperforms baseline methods by **2.42%** on average.

3.3. Detailed analysis

How does our generalization bound compare to previous results? Our result suggests that Hessian-based generalization bounds correlate with empirical performance of fine-tuning. Next, we further compare our result with previous generalization bounds, including norm bounds [LS20; GHP21; LZ21], and margin bounds [PDV17; AGN+18; NBS18]. See Appendix B.1 for a precise description of the measured results. For our result, we numerically calculate $\sum_{i=1}^{L} \sqrt{C \cdot \mathcal{H}_i}/\sqrt{n} + \xi$ from equation 3. We compute $\mathcal{H}_i$ using Hessian vector product with PyHessian [YGK+20]. We take the maximum over the training and test dataset. Following prior works, we evaluate the results using the CIFAR-

Figure 4: Our algorithm reduces the Hessian distance measure more effectively than baseline methods.

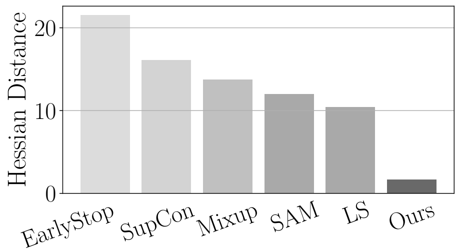


Table 5: Numerical comparison between our generalization bound using Hessians and previous results. The base models include ResNet-50 and BERT-Base.

Generalization bound	CIFAR-10	CIFAR-100	MRPC	SST2
Pitas et al. [PDV17]	5.51E+10	3.13E+12	/	/
Arora et al. [AGN+18]	1.62E+06	9.66E+07	/	/
Long and Sedghi [LS20]	6.30E+13	8.32E+13	/	/
Gouk et al. [GHP21]	2.04E+69	2.72E+69	/	/
Li and Zhang [LZ21]	1.63E+27	1.31E+29	/	/
Neyshabur et al. [NBS18]	3.13E+11	1.62E+13	/	/
Our result	2.26	7.23	3.83	9.71

10 and CIFAR-100 datasets [KH+09]. We use ResNet-50 as the base model for fine-tuning. Additionally, we evaluate our results on text classification tasks, including MRPC and SST-2 from the GLUE benchmark. We use BERT-Base-Uncased [DCL+18] as the base model for fine-tuning.

Table 5 shows the result. We find that our bound is orders of magnitude smaller than previous results. This shows our theory leads to practical bounds for real-world settings.

How much does the Hessian distance measure drop? The design of our algorithm is motivated by our Hessian-based generalization guarantees. We hypothesize that our algorithm can effectively reduce the Hessian distance measure $\sum_{i=1}^{L} \sqrt{\mathcal{H}_i}$ (cf. Equation 3) of fine-tuned models. We compare the Hessian quantity of models fine-tuned by different algorithms. We select five baseline methods to compare with our algorithm and expect similar results for comparison with other baselines. Figure 4 shows the results on ResNet-18 models fine-tuned on the Clipart dataset with weak supervision noise. We notice that our algorithm reduces the quantity by six times more than previous fine-tuning algorithms.

Both components of Algorithm 1 contribute to final results. We study the influence of each component of our algorithm: the weighted loss scheme and the regularization constraints. We ablate the effect of our weighted loss and regularization constraints, showing that both are crucial to the final performance. We run the same experiments with only one component from the algorithm. As shown in Table 7 from Section B.2, removing either component affects the performance negatively. This suggests that both the weighted loss and the regularization constraints play an important role in the final performance.

Table 4: Test accuracy of fine-tuning vision transformers with correlated noisy labels on Clipart and Sketch domains from the DomainNet dataset. Results are averaged over 10 random seeds.

Correlated label noise	DomainNet, ViT-Base	
	Clipart	Sketch
Early Stopping	80.48±2.65	66.29±3.82
Label Smooth [MKH19]	80.00±3.23	67.02±1.38
Mixup [ZCD+18]	81.44±2.33	69.06±1.05
FixMatch [SBL+20]	81.92±1.35	68.08±2.52
ELR [LNR+20]	82.48±2.32	68.33±3.27
APL [MHW+20]	79.12±3.75	68.33±2.39
SAT [HZZ20]	81.20±2.35	68.57±2.70
GJS [EA21]	80.48±2.61	66.94±2.98
DualT [YLH+20]	82.24±3.55	67.92±2.97
SupCon [GDC+20]	82.64±2.80	65.22±2.18
SAM [FKM+21]	80.16±2.00	68.24±2.36
Ours	**84.40±1.94**	**71.02±1.81**

4. Conclusion

This work approached generalization for fine-tuning deep networks using a PAC-Bayesian analysis. The analysis shows that besides distance from initialization, Hessians play a crucial role in measuring generalization. Our theoretical contribution involves Hessian distance based generalization bounds for a variety of fine-tuning methods on deep networks. We empirically show that the Hessian distance measure accurately correlates with the generalization error of fine-tuning in practice. Our theory implies an algorithm for fine-tuning with noisy labels. Experiments demonstrate the robustness of our algorithm under both synthetic and real-world label noise.

We describe two questions for future work. First, is it possible to extend our result to networks with non-differentiable activation functions? In light of the wide use of ReLU activations, it is conceivable that non-differentiable points rarely happen in practice. One approach would be to use the smoothed analysis framework [ST04] and argue that such non-differentiable points occur with negligible probability. Second, it would be interesting to apply Hessians to understand generalization in other black-box models.

Acknowledgement

Thanks to the anonymous referees for their feedback. Thanks to Rajmohan Rajaraman for helpful discussions. This work is in part supported by a seed/proof-of-concept grant from the Khoury College of Computer Sciences, Northeastern University.

References

[ASY+19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. "Optuna: A next-generation hyperparameter optimization framework". In: *KDD*. 2019.

[Alq21] Pierre Alquier. "User-friendly introduction to PAC-Bayes bounds". In: *arXiv preprint arXiv:2110.11216* (2021).

[AL88] Dana Angluin and Philip Laird. "Learning from noisy examples". In: *Machine Learning* (1988).

[AGN+18] 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*. 2018.

[BHG+17] Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. "Regularization techniques for fine-tuning in neural machine translation". In: *EMNLP* (2017).

[BFT17] Peter Bartlett, Dylan J Foster, and Matus Telgarsky. "Spectrally-normalized margin bounds for neural networks". In: *Neural Information Processing Systems* (2017).

[BBC+10] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. "A theory of learning from different domains". In: *Machine learning* (2010).

[BMR+20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners". In: *NeurIPS* (2020).

[Cat07] Olivier Catoni. "PAC-Bayesian supervised classification: the thermodynamics of statistical learning". In: *arXiv preprint arXiv:0712.0248* (2007).

[CPD+21] Evangelos Chatzipantazis, Stefanos Pertigkiozoglou, Edgar Dobriban, and Kostas Daniilidis. "Learning Augmentation Distributions using Transformed Risk Minimization". In: *arXiv preprint arXiv:2111.08190* (2021).

[CKW08] Koby Crammer, Michael Kearns, and Jennifer Wortman. "Learning from Multiple Sources". In: *Journal of Machine Learning Research* 9.8 (2008).

[DLM01] Sanjoy Dasgupta, Michael Littman, and David McAllester. "PAC generalization bounds for co-training". In: *Advances in neural information processing systems* 14 (2001).

[DCL+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *ACL* (2018).

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[DHG+21] Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. "On the role of data in PAC-Bayes". In: *International Conference on Artificial Intelligence and Statistics*. 2021.

[DR17] Gintare Karolina Dziugaite and Daniel M Roy. "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data". In: *UAI* (2017).

[EA21] Erik Englesson and Hossein Azizpour. "Generalized Jensen-Shannon Divergence Loss for Learning with Noisy Labels". In: *Neural Information Processing Systems* (2021).

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1126–1135.

[FKM+21] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. "Sharpness-aware minimization for efficiently improving generalization". In: *ICLR* (2021).

[GKX19] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. "An investigation into neural net optimization via hessian eigenvalue density". In: *International Conference on Machine Learning*. 2019.

[GHP21] Henry Gouk, Timothy M Hospedales, and Massimiliano Pontil. "Distance-Based Regularisation of Deep Networks for Fine-Tuning". In: *ICLR* (2021).

[Gue19] Benjamin Guedj. "A Primer on PAC-Bayesian Learning". In: *Proceedings of the French Mathematical Society*. Vol. 33.

Société Mathématique de France. 2019, pp. 391–414.

[GDC+20] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. “Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning”. In: *International Conference on Learning Representations*. 2020.

[HK19] Steve Hanneke and Samory Kpotufe. “On the value of target data in transfer learning”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[HZR+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[HG16] Dan Hendrycks and Kevin Gimpel. “Gaussian error linear units (gelus)”. In: *arXiv preprint arXiv:1606.08415* (2016).

[HZZ20] Lang Huang, Chao Zhang, and Hongyang Zhang. “Self-adaptive training: beyond empirical risk minimization”. In: *NeurIPS* (2020).

[JNM+20] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. “Fantastic generalization measures and where to find them”. In: *ICLR* (2020).

[JNG+19] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. “A short note on concentration inequalities for random vectors with subgaussian norm”. In: *arXiv preprint arXiv:1902.03736* (2019).

[KLA18] Ashish Khetan, Zachary C Lipton, and Animashree Anandkumar. “Learning From Noisy Singly-labeled Data”. In: *ICLR*. 2018.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. “Learning multiple layers of features from tiny images”. In: (2009).

[LHL21] Qi Lei, Wei Hu, and Jason Lee. “Near-optimal linear regression under distribution shift”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 6164–6174.

[LZ21] Dongyue Li and Hongyang Zhang. “Improved Regularization and Robustness for Fine-tuning in Neural Networks”. In: *NeurIPS* (2021).

[LSO20] 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*. PMLR. 2020, pp. 4313–4324.

[LXW+18] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. “Delta: Deep learning transfer using feature map with attention for convolutional networks”. In: *International Conference on Learning Representations*. 2018.

[LGD18] Xuhong Li, Yves Grandvalet, and Franck Davoine. “Explicit inductive bias for transfer learning with convolutional networks”. In: *ICML* (2018).

[LUZ21] Renjie Liao, Raquel Urtasun, and Richard Zemel. “A PAC-Bayesian Approach to Generalization Bounds for Graph Neural Networks”. In: *ICLR* (2021).

[LNR+20] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. “Early-learning regularization prevents memorization of noisy labels”. In: *NeurIPS* (2020).

[LT15] Tongliang Liu and Dacheng Tao. “Classification with noisy labels by importance reweighting”. In: *IEEE Transactions on pattern analysis and machine intelligence* (2015).

[LHC+19] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. “Multi-task deep neural networks for natural language understanding”. In: *arXiv preprint arXiv:1901.11504* (2019).

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[LS20] Philip M Long and Hanie Sedghi. “Generalization bounds for deep convolutional neural networks”. In: *ICLR* (2020).

[LBM+20] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. “Does label smoothing mitigate label noise?” In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6448–6458.

[Ma21] Tengyu Ma. *Lecture Notes for Machine Learning Theory (CS229M/STATS214)*. 2021. URL: https://raw.githubusercontent.com/tengyuma/cs229m_notes/main/master.pdf.

[MHW+20] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. "Normalized loss functions for deep learning with noisy labels". In: *International Conference on Machine Learning*. 2020.

[MCS+21] Alessio Mazzetto, Cyrus Cousins, Dylan Sam, Stephen H Bach, and Eli Upfal. "Adversarial Multi Class Learning under Weak Supervision with Performance Guarantees". In: *International Conference on Machine Learning*. 2021.

[McA13] David McAllester. "A PAC-Bayesian tutorial with a dropout bound". In: *arXiv preprint arXiv:1307.2118* (2013).

[McA99a] David A McAllester. "PAC-Bayesian model averaging". In: *Proceedings of the twelfth annual conference on Computational learning theory*. 1999.

[McA99b] David A McAllester. "Some pac-bayesian theorems". In: *Machine Learning* (1999).

[MKH19] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. "When does label smoothing help?" In: *NeurIPS* (2019).

[NK19] Vaishnavh Nagarajan and J Zico Kolter. "Generalization in deep networks: The role of distance from initialization". In: *arXiv preprint arXiv:1901.01672* (2019).

[NK18] Vaishnavh Nagarajan and Zico Kolter. "Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience". In: *International Conference on Learning Representations*. 2018.

[NDR+17] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. "Cost-Sensitive Learning with Noisy Labels." In: *J. Mach. Learn. Res.* (2017).

[NDR+13] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. "Learning with noisy labels". In: *Advances in neural information processing systems* (2013).

[NBS18] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. "A pac-bayesian approach to spectrally-normalized margin bounds for neural networks". In: *ICLR* (2018).

[NSZ20] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. "What is being transferred in transfer learning?" In: *NeurIPS* (2020).

[Pap18] Vardan Papyan. "The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size". In: *arXiv preprint arXiv:1811.07062* (2018).

[Pap19] Vardan Papyan. "Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians". In: *ICML* (2019).

[PRK+17] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. "Making deep neural networks robust to label noise: A loss correction approach". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1944–1952.

[PBX+19] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. "Moment matching for multi-source domain adaptation". In: *The International Conference on Computer Vision*. 2019.

[PDV17] Konstantinos Pitas, Mike Davies, and Pierre Vandergheynst. "Pac-bayesian margin bounds for convolutional neural networks". In: *arXiv preprint arXiv:1801.00171* (2017).

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* (2020).

[Rud17] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[SBL16] Levent Sagun, Leon Bottou, and Yann LeCun. "Eigenvalues of the hessian in deep learning: Singularity and beyond". In: *arXiv preprint arXiv:1611.07476* (2016).

[SBG21] Gal Shachaf, Alon Brutzkus, and Amir Globerson. "A Theoretical Analysis of Fine-tuning with Linear Teachers". In: *Advances in Neural Information Processing Systems* (2021).

[SBL+20] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *NeurIPS* (2020).

[SKP+22] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. "Learning from noisy labels with deep neural networks: A survey". In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[ST04] Daniel A Spielman and Shang-Hua Teng. "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time". In: *Journal of the ACM (JACM)* 51.3 (2004), pp. 385–463.

[TJJ20] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. "On the theory of transfer learning: The importance of task diversity". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7852–7862.

[Tro15] Joel A Tropp. "An introduction to matrix concentration inequalities". In: *arXiv preprint arXiv:1501.01571* (2015).

[TSS20] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. "Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis". In: *International Conference on Machine Learning*. 2020.

[VWM+20] Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. "Exploring and predicting transferability across NLP tasks". In: *EMNLP* (2020).

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[WSM+18] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018.

[WIL+22] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. "Robust fine-tuning of zero-shot models". In: *CVPR* (2022).

[WZV+20] Sen Wu, Hongyang Zhang, Gregory Valiant, and Christopher Ré. "On the generalization effects of linear transformations in data augmentation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10410–10420.

[WZR20] Sen Wu, Hongyang R Zhang, and Christopher Ré. "Understanding and improving information transfer in multi-task learning". In: *ICLR* (2020).

[XLW+19] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. "Are anchor points really indispensable in label-noise learning?" In: *Advances in Neural Information Processing Systems* 32 (2019).

[YZW+21] Fan Yang, Hongyang R Zhang, Sen Wu, Weijie J Su, and Christopher Ré. "Analysis of Information Transfer from Heterogeneous Sources via Precise High-dimensional Asymptotics". In: *arXiv preprint arXiv:2010.11750* (2021).

[YLH+20] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. "Dual t: Reducing estimation error for transition matrix in label-noise learning". In: *NeurIPS* (2020).

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. "On early stopping in gradient descent learning". In: *Constructive Approximation* 26.2 (2007), pp. 289–315.

[YGK+20] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. "Pyhessian: Neural networks through the lens of the hessian". In: *IEEE International Conference on Big Data*. 2020.

[ZBH+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *ICLR* (2017).

[ZCD+18] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization". In: *ICLR* (2018).

[ZLA21] Mingyuan Zhang, Jane Lee, and Shivani Agarwal. "Learning from noisy labels with no change to the training process". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12468–12478.

[ZVA+19] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. "Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach". In: *ICLR* (2019).

A. Proofs

We will analyze the generalization performance of fine-tuning using PAC-Bayesian analysis. One chooses a *prior* distribution, denoted as $\mathcal{P}$, and a *posterior* distribution, denoted as $\mathcal{Q}$. The *perturbed* empirical loss of f_W, denoted by $\hat{\mathcal{L}}_{\mathcal{Q}}(f_W)$, is the average of $\ell_{\mathcal{Q}}(f_U(x), y)$ over the training dataset. The *perturbed* expected loss of f_W, denoted by $\mathcal{L}_{\mathcal{Q}}(f_W)$, is the expectation of $\ell_{\mathcal{Q}}(f_W(x), y)$, for a sample x drawn from $\mathcal{D}$ with class label y. We will use the following PAC-Bayes bound from Theorem 2, McAllester [McA13].

Theorem A.1. *Suppose the loss function $\ell(x, y)$ lies in a bounded range $[0, C]$ given any $x \in \mathcal{X}$ with label y. For any $\beta \in (0, 1)$ and $\delta \in (0, 1]$, with probability at least $1 - \delta$, the following holds*

$$\mathcal{L}_{\mathcal{Q}}(f_W) \leq \frac{1}{\beta}\hat{\mathcal{L}}_{\mathcal{Q}}(f_W) + \frac{C\Big(KL(\mathcal{Q}||\mathcal{P}) + \log\frac{1}{\delta}\Big)}{2\beta(1-\beta)n}. \tag{10}$$

This result provides a flexibility in setting β. Our results will set β to balance the perturbation error of $\mathcal{Q}$ and the KL divergence between $\mathcal{P}$ and $\mathcal{Q}$. We will need the KL divergence between the prior $\mathcal{P}$ and the posterior $\mathcal{Q}$ in the PAC-Bayesian analysis. This is stated in the following result.

Proposition A.2. *Suppose the noise perturbation at layer i is drawn from a Gaussian distribution with mean zero and covariance Σ_i, for every $i = 1, \dots, L$. Then, the KL divergence between $\mathcal{P}$ and $\mathcal{Q}$ is equal to*

$$KL(\mathcal{Q}||\mathcal{P}) = \frac{1}{2}\sum_{i=1}^{L} \mathbf{vec}(W_i - \hat{W}_i^{(s)})^\top \Sigma_i^{-1} \mathbf{vec}(W_i - \hat{W}_i^{(s)}).$$

As a corollary, if the population covariance of the noise distribution at every layer is isotropic, i.e., $\Sigma_i = \sigma_i^2 \operatorname{Id}$, for any $i = 1, \dots, L$, then the KL divergence between $\mathcal{P}$ and $\mathcal{Q}$ is equal to

$$KL(\mathcal{Q}||\mathcal{P}) = \sum_{i=1}^{L} \frac{\|W_i - \hat{W}_i^{(s)}\|_F^2}{2\sigma_i^2}.$$

Proof. The proof is based on the definition of multivariate normal distributions. Let Z_i be the weight matrix of every layer i in the posterior distribution, for i from 1 to L. By the definition of the KL divergence between two distributions, we have

$$\begin{aligned}
KL(\mathcal{Q}||\mathcal{P}) &= \mathop{\mathbb{E}}_{Z\sim\mathcal{Q}}\left[\log\left(\frac{\mathcal{Q}(Z)}{\mathcal{P}(Z)}\right)\right] = \mathop{\mathbb{E}}_{Z\sim\mathcal{Q}}\left[\log \mathcal{Q}(Z) - \log \mathcal{P}(Z)\right] \\
&= \mathop{\mathbb{E}}_{Z\sim\mathcal{Q}}\left[\sum_{i=1}^{L} -\frac{1}{2}\mathbf{vec}(Z_i - W_i)^\top \Sigma_i^{-1}\mathbf{vec}(Z_i - W_i) + \frac{1}{2}\mathbf{vec}(Z_i - \hat{W}_i^{(s)})^\top \Sigma_i^{-1}\mathbf{vec}(Z_i - \hat{W}_i^{(s)})\right] \\
&= -\frac{1}{2}\mathop{\mathbb{E}}_{Z\sim\mathcal{Q}}\left[\sum_{i=1}^{L} \operatorname{Tr}\left[\mathbf{vec}(Z_i - W_i)\mathbf{vec}(Z_i - W_i)^\top \Sigma_i^{-1}\right] - \operatorname{Tr}\left[\mathbf{vec}(Z_i - \hat{W}_i^{(s)})\mathbf{vec}(Z_i - \hat{W}_i^{(s)})^\top \Sigma_i^{-1}\right]\right].
\end{aligned}$$

In the above equation, we recall that the expectation of Z_i is W_i. Additionally, the population covariance of Z_i is Σ_i. Therefore, after cancelling out common terms, we get

$$KL(\mathcal{Q}||\mathcal{P}) = \sum_{i=1}^{L} \frac{1}{2}\mathbf{vec}(W_i - \hat{W}_i^{(s)})^\top \Sigma_i^{-1} \mathbf{vec}(W_i - \hat{W}_i^{(s)}).$$

The proof is complete. □

As stated in Theorem 2.1, we require the activation functions $\phi_i(\cdot)$ (for any $1 \leq i \leq L$) and loss function $\ell(\cdot, \cdot)$ to be twice-differentiable. Additionally, we require that they are Lipschitz-continuous. For brevity, We restate these conditions below.

Assumption A.3. *Assume that the activation functions $\phi_i(\cdot)$ (for any $1 \leq i \leq L$) and the loss function $\ell(\cdot, \cdot)$ over the first argument are twice-differentiable and κ_0-Lipschitz. Their first-order derivatives are κ_1-Lipschitz and their second-order derivatives are κ_2-Lipschitz.*

We give several examples of these Lipschitz constants for commonly used activation functions.

Example A.4. *For the Sigmoid function $S(x)$, we have $S(x) \in (0,1)$ for any $x \in \mathbb{R}$. Since the derivative of the Sigmoid function $S'(x) = S(x)(1 - S(x))$, we get that κ_0, κ_1, and κ_2 are at most 1/4.*

For the Tanh function $Tanh(x)$, we have $Tanh(x) \in (-1,1)$ for any $x \in \mathbb{R}$. Since the derivative of the Tanh function satisfies $Tanh'(x) = 1 - (Tanh(x))^2$, we get that κ_0, κ_1, and κ_2 are at most 1.

For the GELU function $GELU(x)$ [HG16], we have $GELU(x) = x \cdot F(x)$ where $F(x)$ be the cumulative density function (CDF) of the standard normal distribution. Let $p(x)$ be the probability density function (PDF) of the standard normal distribution. We have that $F'(x) = p(x)$ and $p'(x) = -xp(x)$. Then, we can get that κ_0, κ_1, and κ_2 at most $1 + e^{-1/2}/\sqrt{2\pi} \approx 1.242$.

A.1. Proof of Theorem 2.1

Proof of Theorem 2.1. First, we separate the gap of $\mathcal{L}(f_{\hat{W}})$ and $\frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}})$ into three parts:

$$\mathcal{L}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}}) = \mathcal{L}(f_{\hat{W}}) - \mathcal{L}_{\mathcal{Q}}(f_{\hat{W}}) + \mathcal{L}_{\mathcal{Q}}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}}) + \frac{1}{\beta}\hat{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}}).$$

By the Taylor's expansion of Lemma 2.3, we can bound the noise stability of $f_{\hat{W}}(x)$ with respect to the empirical loss and the expected loss:

$$\begin{aligned}\mathcal{L}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}}) \leq & \left(- \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{L}\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right\rangle\right] + \sum_{i=1}^{L} C_1 \left\|\Sigma_i\right\|_F^{3/2}\right) \\ & + \left(\mathcal{L}_{\mathcal{Q}}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}})\right) + \frac{1}{\beta}\left(\frac{1}{n}\sum_{i=1}^{L}\sum_{j=1}^{n}\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x_j), y_j)]\right\rangle + \sum_{i=1}^{L} C_1 \left\|\Sigma_i\right\|_F^{3/2}\right),\end{aligned}$$

which is equivalent to the following equation:

$$\begin{aligned}\mathcal{L}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}}) \leq & \left(- \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{L}\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right\rangle\right] + \frac{1}{n\beta}\sum_{i=1}^{L}\sum_{j=1}^{n}\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x_j), y_j)]\right\rangle\right) \\ & + \left(\frac{1}{\beta} + 1\right)\sum_{i=1}^{L} C_1 \left\|\Sigma_i\right\|_F^{3/2} + \left(\mathcal{L}_{\mathcal{Q}}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}})\right).\end{aligned} \tag{11}$$

Next, we combine the upper bound of the noise stability of $f_{\hat{W}}(x)$ with respect to the empirical loss and the expected loss:

$$\begin{aligned} & \frac{1}{n\beta}\sum_{i=1}^{L}\sum_{j=1}^{n}\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x_j), y_j)]\right\rangle - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{L}\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right\rangle\right] \\ = & \sum_{i=1}^{L}\left(\frac{1}{n}\sum_{j=1}^{n}\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x_j), y_j)]\Big\rangle - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\Big\rangle\right]\right) \\ & + \left(\frac{1}{\beta} - 1\right)\frac{1}{n}\sum_{i=1}^{L}\sum_{j=1}^{n}\left\langle \Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x_j), y_j)]\right\rangle.\end{aligned} \tag{12}$$

We use the uniform convergence result of Lemma 2.4 to bound equation (12), leading to a concentration error term of

$$\begin{aligned} & \sum_{i=1}^{L}\left\langle \Sigma_i, \frac{1}{n}\sum_{j=1}^{n}\mathbf{H}_i(\ell[f_{\hat{W}}(x_j), y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right] \right\rangle \\ \leq & \sum_{i=1}^{L}\left\|\Sigma_i\right\|_F \left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{H}_i[\ell(f_{\hat{W}}(x_j), y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right]\right\|_F \\ \leq & \frac{C_2\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}\sum_{i=1}^{L}\left\|\Sigma_i\right\|_F.\end{aligned} \tag{13}$$

Recall that v_i is a flatten vector of the matrix $\hat{W}_i - \hat{W}_i^{(s)}$. By the PAC-Bayes bound of Theorem A.1 and the KL divergence result of Proposition A.2,

$$\mathcal{L}_{\mathcal{Q}}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}}) \leq \frac{C\big(KL(\mathcal{Q}||\mathcal{P}) + \log\frac{1}{\delta}\big)}{2\beta(1-\beta)n} \leq \frac{C\Big(\frac{1}{2}\sum_{i=1}^{L}\langle v_i, \Sigma_i^{-1}v_i\rangle + \log\frac{1}{\delta}\Big)}{2\beta(1-\beta)n}. \tag{14}$$

Combining equations (11), (13), (14) with probability at least $1-2\delta$, we get

$$\mathcal{L}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}}) \leq \frac{C_2\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}\sum_{i=1}^{L}\|\Sigma_i\|_F + \Big(\frac{1}{\beta}-1\Big)\frac{1}{n}\sum_{i=1}^{L}\sum_{j=1}^{n}\Big\langle\Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x_j), y_j)]\Big\rangle + \Big(\frac{1}{\beta}+1\Big)C_1\sum_{j=1}^{L}\|\Sigma_i\|_F^{3/2} + \frac{C(\frac{1}{2}\sum_{i=1}^{L}\langle v_i, \Sigma_i^{-1}v_i\rangle + \log\frac{1}{\delta})}{2\beta(1-\beta)n}.$$

Recall the truncated Hessian $\mathbf{H}_i^{+}[\ell(f_{\hat{W}}(x), y)]$ is equal to $U_i\max(D_i, 0)U_i^T$, where $U_iD_iU_i^T$ is the eigendecomposition of $\mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]$. We have

$$\big\langle\Sigma_i, \mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\big\rangle \leq \big\langle\Sigma_i, \mathbf{H}_i^{+}[\ell(f_{\hat{W}}(x), y)]\big\rangle.$$

Next, the upper bound of $\mathcal{L}(f_{\hat{W}}) - \frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}})$ relies on Σ_i and β. Our goal is to select some Σ_i and $\beta \in (0,1)$ to minimize this quantity. For any $x \in \mathcal{X}$, let Σ_i be the matrix that satisfies

$$\Big(\frac{1}{\beta}-1\Big)\big\langle\Sigma_i, \mathbf{H}_i^{+}[\ell(f_{\hat{W}}(x), y)]\big\rangle = \frac{C\langle v_i, \Sigma_i^{-1}v_i\rangle}{4\beta(1-\beta)n}, \tag{15}$$

which implies

$$\Sigma_i = \sqrt{\frac{C}{4(1-\beta)^2 n\|v_i\|^2}}\mathbf{H}_i^{+}[\ell(f_{\hat{W}}(x), y)]^{-\frac{1}{2}}v_iv_i^{\top}. \tag{16}$$

If the truncated Hessian is zero, then Σ_i will be equal to zero. Next, we substitute Σ_i into equation (15) and get

$$\begin{aligned}\Big(\frac{1}{\beta}-1\Big)\big\langle\Sigma_i, \mathbf{H}_i^{+}(\ell(f_{\hat{W}}(x), y))\big\rangle = \frac{C\langle v_i, \Sigma_i^{-1}v_i\rangle}{4\beta(1-\beta)n} &= \sqrt{\frac{C}{4\beta^2 n\|v_i\|^2}}\Big\langle\mathbf{H}_i^{+}[\ell(f_{\hat{W}}(x), y)]^{\frac{1}{2}}, v_iv_i^{\top}\Big\rangle \\ &\leq \sqrt{\frac{C}{4\beta^2 n\|v_i\|^2}}\Big\|\mathbf{H}_i^{+}[\ell(f_{\hat{W}}(x), y)]^{\frac{1}{2}}v_i\Big\| \cdot \|v_i\| \\ &= \sqrt{\frac{C\cdot v_i^{\top}\mathbf{H}_i^{+}[\ell(f_{\hat{W}}(x), y)]v_i}{4\beta^2 n}} \leq \sqrt{\frac{C\mathcal{H}_i}{4\beta^2 n}}.\end{aligned}$$

Thus, we have shown that equation (15) is less than $\sqrt{\frac{C\mathcal{H}_i}{4\beta^2 n}}$. Next, the gap between $\mathcal{L}(f_{\hat{W}})$ and $\frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}})$ is

$$\begin{aligned}\mathcal{L}(f_{\hat{W}}) \leq& \left(\frac{1}{\beta}\hat{\mathcal{L}}(f_{\hat{W}}) + \sum_{i=1}^{L}\sqrt{\frac{C\mathcal{H}_i}{\beta^2 n}}\right) \\ &+ \left(\frac{C_2\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}\sum_{i=1}^{L}\|\Sigma_i\|_F + \Big(1+\frac{1}{\beta}\Big)C_1\sum_{i=1}^{L}\|\Sigma_i\|_F^{3/2} + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta}\right).\end{aligned} \tag{17}$$

Based on equation (17), we set $\epsilon = (1-\beta)/\beta$. Then we have

$$\begin{aligned}\mathcal{L}(f_{\hat{W}}) \leq& (1+\epsilon)\Big(\hat{\mathcal{L}}(f_{\hat{W}}) + \sum_{i=1}^{L}\sqrt{\frac{C\mathcal{H}_i}{n}}\Big) \\ &+ \left(\frac{C_2\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}\sum_{i=1}^{L}\|\Sigma_i\|_F + \Big(1+\frac{1}{\beta}\Big)C_1\sum_{i=1}^{L}\|\Sigma_i\|_F^{3/2} + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta}\right).\end{aligned}$$

Let ξ be defined as follows. We have

$$\xi = \frac{C_2\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}} \sum_{i=1}^{L} \|\Sigma_i\|_F + \Big(1+\frac{1}{\beta}\Big) C_1 \sum_{i=1}^{L} \|\Sigma_i\|_F^{3/2} + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta} = O\Big(n^{-3/4}\Big),$$

where we use that $\|\Sigma_i\|_F = O\left(n^{-1/2}\right)$ from equation (16). Hence, we conclude that

$$\mathcal{L}(f_{\hat{W}}) \le (1+\epsilon)\hat{\mathcal{L}}(f_{\hat{W}}) + (1+\epsilon)\sum_{i=1}^{L}\sqrt{\frac{C\mathcal{H}_i}{n}} + \xi.$$

Thus, we have finished the proof. □

A.2. Proof of Theorem 2.2

Proof of Theorem 2.2. By equation (6), we have the expected loss of $f_{\hat{W}}$ using $\bar{\ell}$ as

$$\mathcal{L}(f_{\hat{W}}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\bar{\ell}(f_{\hat{W}}(x), y)\right] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}, \tilde{y}|y}\left[\bar{\ell}(f_{\hat{W}}(x), \tilde{y})\right].$$

Denote the empirical loss of $f_{\hat{W}}$ using $\bar{\ell}$ as

$$\bar{\mathcal{L}}(f_{\hat{W}}) = \frac{1}{n}\sum_{i=1}^{n}\bar{\ell}(f_{\hat{W}}(x_i), \tilde{y}_i).$$

We separate the gap between $\mathcal{L}\big(f_{\hat{W}}\big)$ and $\frac{1}{\beta}\bar{\mathcal{L}}\big(f_{\hat{W}}\big)$ into three parts:

$$\mathcal{L}\big(f_{\hat{W}}\big) - \frac{1}{\beta}\bar{\mathcal{L}}\big(f_{\hat{W}}\big) \le \mathcal{L}\big(f_{\hat{W}}\big) - \mathcal{L}_{\mathcal{Q}}\big(f_{\hat{W}}\big) + \mathcal{L}_{\mathcal{Q}}\big(f_{\hat{W}}\big) - \frac{1}{\beta}\bar{\mathcal{L}}_{\mathcal{Q}}\big(f_{\hat{W}}\big) + \frac{1}{\beta}\bar{\mathcal{L}}_{\mathcal{Q}}\big(f_{\hat{W}}\big) - \frac{1}{\beta}\bar{\mathcal{L}}\big(f_{\hat{W}}\big).$$

We denote $\bar{\ell}_{\mathcal{Q}}(f_{\hat{W}}(x), \tilde{y})$ as an expectation of $\bar{\ell}(f_{\hat{W}}(x), \tilde{y})$ under the posterior $\mathcal{Q}$. Define $\mathcal{L}_{\mathcal{Q}}\big(f_{\hat{W}}\big)$ as $\mathbb{E}_{(x,y)\sim\mathcal{D},\tilde{y}|y}\left[\bar{\ell}_{\mathcal{Q}}(f_{\hat{W}}(x), \tilde{y})\right]$. Define $\bar{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}})$ as $\frac{1}{n}\sum_{i=1}^{n}\bar{\ell}_{\mathcal{Q}}(f_{\hat{W}}(x_i), \tilde{y}_i)$. By equation (6), we have

$$\begin{aligned}
\mathcal{L}\big(f_{\hat{W}}\big) - \mathcal{L}_{\mathcal{Q}}\big(f_{\hat{W}}\big) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D},\tilde{y}|y}\left[\bar{\ell}_{\mathcal{Q}}(f_{\hat{W}}(x), \tilde{y}) - \bar{\ell}(f_{\hat{W}}(x), \tilde{y})\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{\tilde{y}|y}\left[\bar{\ell}_{\mathcal{Q}}(f_{\hat{W}}(x), \tilde{y}) - \bar{\ell}(f_{\hat{W}}(x), \tilde{y})\right]\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell_{\mathcal{Q}}(f_{\hat{W}}(x), y) - \ell(f_{\hat{W}}(x), y)\right].
\end{aligned}$$

Then, by equation (5), the noise stability of $f_{\hat{W}}$ with respect to the empirical loss is equal to

$$\begin{aligned}
\frac{1}{\beta}\bar{\mathcal{L}}_{\mathcal{Q}}\big(f_{\hat{W}}\big) - \frac{1}{\beta}\bar{\mathcal{L}}\big(f_{\hat{W}}\big) =& \frac{1}{n\beta}\sum_{j=1}^{n}\bar{\ell}_{\mathcal{Q}}(f_{\hat{W}}(x_j), \tilde{y}_j) - \frac{1}{n\beta}\sum_{j=1}^{n}\bar{\ell}(f_{\hat{W}}(x_j), \tilde{y}_j) \\
=& \frac{1}{n\beta}\sum_{l=1}^{k}\Lambda_{\tilde{y},l}\left(\sum_{j=1}^{n}\left(\ell_{\mathcal{Q}}(f_{\hat{W}}(x_j), l) - \ell(f_{\hat{W}}(x_j), l)\right)\right).
\end{aligned}$$

By the Taylor's expansion of Lemma 2.3, we can bound the noise stability of $f_{\hat{W}}$ with respect to the empirical loss and the expected loss:

$$\begin{aligned}
&\mathcal{L}\big(f_{\hat{W}}\big) - \frac{1}{\beta}\bar{\mathcal{L}}\big(f_{\hat{W}}\big) \\
\le& \Bigg(-\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{L}\sigma_i^2 \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right]\right] + \sum_{i=1}^{L} C_1\sigma_i^3\Bigg) + \Big(\mathcal{L}_{\mathcal{Q}}\big(f_{\hat{W}}\big) - \frac{1}{\beta}\bar{\mathcal{L}}_{\mathcal{Q}}\big(f_{\hat{W}}\big)\Big) \\
&+ \frac{1}{\beta}\Bigg(\frac{1}{n}\sum_{i=1}^{L}\sigma_i^2\Big(\sum_{l=1}^{k}\Lambda_{\tilde{y},l}\big(\sum_{j=1}^{n}\operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x_j), l)]\right]\big)\Big) + \sum_{l=1}^{k}\Lambda_{\tilde{y},l}\sum_{i=1}^{L} C_1\sigma_i^3\Bigg),
\end{aligned}$$

which is equivalent to the following equation:

$$\begin{aligned}
&\mathcal{L}(f_{\hat{W}}) - \frac{1}{\beta}\bar{\mathcal{L}}(f_{\hat{W}}) \\
\leq &\Bigg(- \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sum_{i=1}^{L} \sigma_i^2 \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right] \right] + \frac{1}{n\beta} \sum_{i=1}^{L} \sigma_i^2 \Big(\sum_{l=1}^{k} \Lambda_{\tilde{y},l} \big(\sum_{j=1}^{n} \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x_j), l)]\right] \big) \Big) \Bigg) \\
&+ \left(\sum_{i=1}^{L} C_1 \sigma_i^3 + \frac{1}{\beta} \sum_{l=1}^{k} \Lambda_{\tilde{y},l} \sum_{i=1}^{L} C_1 \sigma_i^3 \right) + \Big(\mathcal{L}_{\mathcal{Q}}(f_{\hat{W}}) - \frac{1}{\beta}\bar{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}}) \Big). \qquad (18)
\end{aligned}$$

Next, we combine the upper bound of the noise stability of $f_{\hat{W}}(x)$ with respect to the empirical loss and the expected loss:

$$\begin{aligned}
&\frac{1}{n\beta} \sum_{i=1}^{L} \sigma_i^2 \Big(\sum_{l=1}^{k} \Lambda_{\tilde{y},l} \big(\sum_{j=1}^{n} \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x_j), l)]\right] \big) \Big) - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sum_{i=1}^{L} \sigma_i^2 \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right] \right] \\
= &\sum_{i=1}^{L} \sigma_i^2 \left(\sum_{l=1}^{k} \Lambda_{\tilde{y},l} \Big(\frac{1}{n} \sum_{j=1}^{n} \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x_j), l)]\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x), y)]\right] \right] \Big) \right) \qquad (19) \\
&+ \Big(\frac{1}{\beta} - 1 \Big) \sum_{i=1}^{L} \sigma_i^2 \left(\sum_{l=1}^{k} \Lambda_{\tilde{y},l} \Big(\frac{1}{n} \sum_{j=1}^{n} \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x_j), l)]\right] \Big) \right). \qquad (20)
\end{aligned}$$

We use the uniform convergence result of Lemma 2.4 to bound equation (19), leading to a concentration error term of

$$\sum_{i=1}^{L} \sigma_i^2 \max_{\tilde{y}} \sum_{j=1}^{L} |\Lambda_{\tilde{y},j}| \frac{C_2 \sqrt{\log(C_3 n/\delta)}}{\sqrt{n}} = \frac{C_2 \sqrt{\log(C_3 n/\delta)}}{\sqrt{n}} \sum_{i=1}^{L} \sigma_i^2 \rho. \qquad (21)$$

For equation (20), the upper bound will be

$$\begin{aligned}
&\Big(\frac{1}{\beta} - 1 \Big) \sum_{i=1}^{L} \sigma_i^2 \left(\sum_{l=1}^{k} \Lambda_{\tilde{y},l} \Big(\frac{1}{n} \sum_{j=1}^{n} \operatorname{Tr}\left[\mathbf{H}_i[\ell(f_{\hat{W}}(x_j), l)]\right] \Big) \right) \\
\leq &\Big(\frac{1}{\beta} - 1 \Big) \sum_{i=1}^{L} \sigma_i^2 \rho \max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left| \operatorname{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x), y))] \right|. \qquad (22)
\end{aligned}$$

where $\rho = \|\Lambda^\top\|_{1,\infty}$. Recall that v_i is a flatten vector of the matrix $\hat{W}_i - \hat{W}_i^{(s)}$. By the PAC-Bayes bound of Theorem A.1 and Proposition A.2,

$$\begin{aligned}
\mathcal{L}_{\mathcal{Q}}(f_{\hat{W}}) - \frac{1}{\beta}\bar{\mathcal{L}}_{\mathcal{Q}}(f_{\hat{W}}) &\leq \frac{C\Big(KL(\mathcal{Q}||\mathcal{P}) + \log\frac{1}{\delta}\Big)}{2\beta(1-\beta)n} \\
&\leq \frac{C\Big(\sum_{i=1}^{L} \frac{\|v_i\|^2}{2\sigma_i^2} + \log\frac{1}{\delta}\Big)}{2\beta(1-\beta)n}. \qquad (23)
\end{aligned}$$

Combining equations (18), (22), (23), with probability at least $1 - 2\delta$, we get

$$\begin{aligned}
\mathcal{L}(f_{\hat{W}}) - \frac{1}{\beta}\bar{\mathcal{L}}(f_{\hat{W}}) \leq &\frac{C_2 \rho \sqrt{\log(C_3 n/\delta)}}{\sqrt{n}} \sum_{i=1}^{L} \sigma_i^2 + \Big(\frac{1}{\beta} - 1 \Big) \rho \sum_{i=1}^{L} \sigma_i^2 \max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left| \operatorname{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x), y))] \right| \\
&+ \Big(\frac{\rho}{\beta} + 1 \Big) \sum_{j=1}^{L} C_1 \sigma_i^3 + \frac{C(\sum_{i=1}^{L} \frac{\|v_i\|^2}{2\sigma_i^2} + \log\frac{1}{\delta})}{2\beta(1-\beta)n}. \qquad (24)
\end{aligned}$$

In equation (24), the upper bound of $\mathcal{L}\big(f_{\hat{W}}\big) - \frac{1}{\beta}\bar{\mathcal{L}}\big(f_{\hat{W}}\big)$ relies on σ_i and β. Our goal is to select $\sigma_i > 0$ and $\beta \in (0,1)$ to minimize equation (24). This is achieved when

$$\sigma_i^2 = \sqrt{\frac{C\left\|v_i\right\|^2}{4\rho(1-\beta)^2 n \cdot \max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left|\mathrm{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x),y))]\right|}}, \text{ for any } 1\le i\le L, \tag{25}$$

following the condition

$$\Big(\frac{1}{\beta}-1\Big)\rho\sigma_i^2 \max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left|\mathrm{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x),y))]\right| = \frac{C}{2\beta(1-\beta)n}\frac{\left\|v_i\right\|^2}{2\sigma_i^2}.$$

Then, equation (24) becomes

$$\mathcal{L}\big(f_{\hat{W}}\big) \le \Bigg(\frac{1}{\beta}\bar{\mathcal{L}}\big(f_{\hat{W}}\big) + \sum_{i=1}^{L}\sqrt{\frac{C\rho\max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left|\mathrm{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x),y))]\right|\cdot\left\|v_i\right\|^2}{\beta^2 n}}\Bigg) \tag{26}$$

$$+\Bigg(\frac{C_2\rho\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}\sum_{i=1}^{L}\sigma_i^2 + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta} + \Big(\frac{\rho}{\beta}+1\Big)C_1\sum_{i=1}^{L}\sigma_i^3\Bigg). \tag{27}$$

Set β as a fixed value close to 1 that does not grow with n and $1/\delta$. Let $\epsilon = (1-\beta)/\beta$. We get

$$\mathcal{L}\big(f_{\hat{W}}\big) \le (1+\epsilon)\Bigg(\bar{\mathcal{L}}\big(f_{\hat{W}}\big) + \sum_{i=1}^{L}\sqrt{\frac{C\rho\max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left|\mathrm{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x),y))]\right|\cdot\left\|v_i\right\|^2}{n}}\Bigg)$$
$$+\Bigg(\frac{C_2\rho\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}\sum_{i=1}^{L}\sigma_i^2 + \Big(\frac{\rho}{\beta}+1\Big)C_1\sum_{i=1}^{L}\sigma_i^3 + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta}\Bigg).$$

Let ξ be defined as follows. We have

$$\xi = \frac{C_2\rho\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}\sum_{i=1}^{L}\sigma_i^2 + \Big(\frac{\rho}{\beta}+1\Big)C_1\sum_{i=1}^{L}\sigma_i^3 + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta} = O\Big(n^{-3/4}\Big),$$

where we notice that $\sigma_i^2 = O\left(n^{-1/2}\right)$ because of equation (25). Thus, we conclude that

$$\mathcal{L}\big(f_{\hat{W}}\big) \le (1+\epsilon)\bar{\mathcal{L}}\big(f_{\hat{W}}\big) + (1+\epsilon)\sum_{i=1}^{L}\sqrt{\frac{C\rho\max_{x\in\mathcal{X}, y\in\{1,\dots,k\}} \left|\mathrm{Tr}[\mathbf{H}_i(\ell(f_{\hat{W}}(x),y))]\right|\cdot\left\|v_i\right\|^2}{n}} + \xi.$$

Since $\|v_i\| \le \alpha_i$, for any $i = 1,\dots,L$, we have finished the proof. □

A.3. Proof of Lemma 2.4

In this section, we provide the proof of Lemma 2.4. We need a perturbation analysis of the Hessian as follows. Given an L-layer neural network $f_W : \mathcal{X} \to \mathbb{R}^{d_L}$ and any input $z^0 \in \mathcal{X}$, let $z_W^i = \phi_i(W_i z_W^{i-1})$ be the activated output vector from layer i, for any $i = 1, 2, \dots, L$.

Proposition A.5. *Suppose that Assumption A.3 holds. The change in the Hessian of output of the loss function $\ell(f_W(x), y)$ with respect to W_i under perturbation U can be bounded as follows:*

$$\|\mathbf{H}_i[\ell(f_W(x),y)]\,|_{W_i+U_i,\forall 1\le i\le L} - \mathbf{H}_i[\ell(f_W(x),y)]\|_F \le G\|U\|_F, \tag{28}$$

where G has the following equation:

$$G = \frac{3}{2}(L+1)^2 e^6 \kappa_2\kappa_1^2\kappa_0^{3(L+1)}\Big(\max_{x\in\mathcal{X}}\|x\|\prod_{l=0}^{L} d_l \prod_{h=1}^{L}\|W_h\|_2\Big)^3\Big(\max_{1\le i\le L}\frac{1}{\|W_i\|_2^2}\sum_{h=1}^{L}\frac{1}{\|W_h\|_2}\Big). \tag{29}$$

where $\kappa_0, \kappa_1, \kappa_2 \ge 1$ is the maximum over the Lipschitz constants, first-order derivative Lipschitz constants, and second-order derivative Lipschitz constants of all the activation functions and the loss function.

We divide the proof of Proposition A.5 into several parts. First, we derive the perturbation of the network output.

Claim A.6. *Suppose that Assumption A.3 holds. For any $j = 1, 2, \cdots, L$, the change in the output of the j layer network z_W^j with perturbation U can be bounded as follows:*

$$\left\| z_{W+U}^j - z_W^j \right\| \leq \left(1 + \frac{1}{L}\right)^j \kappa_0^j \left(\max_{x \in \mathcal{X}} \|x\| \prod_{h=1}^{j} \|W_h\|_2 \sum_{h=1}^{j} \frac{\|U_h\|_2}{\|W_h\|_2} \right). \tag{30}$$

where $\kappa_0 \geq 1$ is the maximum over the Lipschitz constants of all the activation functions and the loss function.

See Lemma 2 in Neyshabur et al. [NBS18] for a proof. Second, we derive the perturbation of the derivative of the network.

Claim A.7. *Suppose that Assumption A.3 holds. The change in the Jacobian of output of the j layer network z_W^j with respect to W_i under perturbation U can be bounded as follows:*

$$\left\| \frac{\partial}{\partial W_i} z_W^j \Big|_{W_k + U_k, \forall 1 \leq k \leq j} - \frac{\partial}{\partial W_i} z_W^j \right\|_F \leq A_{i,j} \left(\max_{x \in \mathcal{X}} \|x\| \prod_{h=1}^{j} \|W_h\|_2 \sum_{h=1}^{j} \frac{\|U_h\|_2}{\|W_h\|_2} \right). \tag{31}$$

For any $j = 1, 2, \cdots, L$, $A_{i,j}$ has the following equation:

$$A_{i,j} = (j - i + 2)\left(1 + \frac{1}{L}\right)^{3j} \kappa_1 \kappa_0^{2j-1} \left(\max_{x \in \mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{l=i-1}^{j} d_l \prod_{h=1}^{j} \|W_h\|_2 \right),$$

where $\kappa_0, \kappa_1 \geq 1$ is the maximum over the Lipschitz constants and first-order derivative Lipschitz constants of all the activation functions and the loss function.

Proof. For any fixed $i = 1, 2, \cdots, L$, we will prove using induction with respect to j. If $j < i$, we have $\frac{\partial}{\partial W_i} z_W^j = 0$ since z_W^j remains unchanged regardless of any change of W_i. When $j = i$, we have

$$\begin{aligned} &\left\| \frac{\partial}{\partial W_i^{p,q}} z_W^i \Big|_{W_k + U_k, \forall 1 \leq k \leq i} - \frac{\partial}{\partial W_i^{p,q}} z_W^i \right\| = \left\| \frac{\partial}{\partial W_i^{p,q}} \phi_i(W_i z_W^{i-1}) \Big|_{W_k + U_k, \forall 1 \leq k \leq i} - \frac{\partial}{\partial W_i^{p,q}} \phi_i(W_i z_W^{i-1}) \right\| \\ =& \left| [z_{W+U}^{i-1}]^q \phi_i' \left(\sum_{r=1}^{d_{i-1}} \left((W_i^{p,r} + U_i^{p,r}) [z_{W+U}^{i-1}]^r \right) \right) - [z_W^{i-1}]^q \phi_i' \left(\sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right) \right|, \end{aligned}$$

where $W_i^{p,q}, U_i^{p,q}$ are the (p, q)'th entry of W_i and U_i. $[z_W^{i-1}]^q$ is the q'th entry of z_W^{i-1}. Next, by the triangle inequality, the above equation is smaller than

$$\left| [z_{W+U}^{i-1}]^q \right| \left| \phi_i' \left(\sum_{r=1}^{d_{i-1}} \left((W_i^{p,r} + U_i^{p,r}) [z_{W+U}^{i-1}]^r \right) - \phi_i' \left(\sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right) \right| \tag{32}$$

$$+ \left| [z_{W+U}^{i-1}]^q - [z_W^{i-1}]^q \right| \left| \phi_i' \left(\sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right) \right|. \tag{33}$$

Since ϕ_i' is κ_1-Lipschitz by Assumption A.3, equation (32) is at most

$$\left| [z_{W+U}^{i-1}]^q \right| \cdot \kappa_1 \left| \sum_{r=1}^{d_{i-1}} (W_i^{p,r} + U_i^{p,r}) [z_{W+U}^{i-1}]^r - \sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right|. \tag{34}$$

Since Assumption A.3 holds, we have $|\phi_i'(\cdot)| \leq \kappa_0$, by the Lipschitz condition on the first-order derivatives. Thus, equation (33) is at most

$$\left| [z_{W+U}^{i-1}]^q - [z_W^{i-1}]^q \right| \cdot \kappa_0. \tag{35}$$

Taking equation (34) and (35) together, the Frobenius norm of the Jacobian is

$$\left\|\frac{\partial}{\partial W_i} z_W^i \Big|_{W_k+U_k,\forall 1\le k\le i} - \frac{\partial}{\partial W_i} z_W^i\right\|_F = \sqrt{\sum_{p=1}^{d_i}\sum_{q=1}^{d_{i-1}} \left\|\frac{\partial}{\partial W_i^{p,q}} z_W^i \Big|_{W_k+U_k,\forall 1\le k\le i} - \frac{\partial}{\partial W_i^{p,q}} z_W^i\right\|^2}$$
$$\le d_i \left\|z_{W+U}^{i-1}\right\| \cdot k_1 d_{i-1} \left\|(W_i+U_i) z_{W+U}^{i-1} - W_i z_W^{i-1}\right\| + \kappa_0 d_i \left\|z_{W+U}^{i-1} - z_W^{i-1}\right\|.$$

By the Lipschitz continuity of z_{W+U}^{i-1} and equation (30) in Claim A.6, we have

$$\left\|z_{W+U}^{i-1}\right\| \le \left(1+\frac{1}{L}\right)^{i-1} \kappa_0^{i-1} \max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{i-1} \|W_h\|_2. \tag{36}$$

$$\left\|z_{W+U}^{i-1} - z_W^{i-1}\right\| \le \left(1+\frac{1}{L}\right)^{i-1} \kappa_0^{i-1} \max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{i-1} \|W_h\|_2 \sum_{h=1}^{i-1} \frac{\|U_h\|_2}{\|W_h\|_2}. \tag{37}$$

By the triangle inequality, we have

$$\left\|(W_i+U_i) z_{W+U}^{i-1} - W_i z_W^{i-1}\right\| \le \|W_i+U_i\|_2 \left\|z_{W+U}^{i-1} - z_W^{i-1}\right\| + \|U_i\|_2 \left\|z_W^{i-1}\right\|. \tag{38}$$

From equation (36), (37), and (38), we finally obtain that the Jacobian of z_W^i with respect to W_i is at most

$$\left\|\frac{\partial}{\partial W_i} z_W^i \Big|_{W_k+U_k,\forall 1\le k\le i} - \frac{\partial}{\partial W_i} z_W^i\right\|_F$$
$$\le \left(2\left(1+\frac{1}{L}\right)^{2i-1} \kappa_1 \kappa_0^{2i-1} d_i d_{i-1} \max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{i-1} \|W_h\|_2\right)\left(\max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{i} \|W_h\|_2 \sum_{h=1}^{i} \frac{\|U_h\|_2}{\|W_h\|_2}\right).$$

Hence, we know that equation (30) will be correct when $j = i$. Assuming that equation (30) will be correct for any i up to $j \ge i$, we have

$$\left\|\frac{\partial}{\partial W_i} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} - \frac{\partial}{\partial W_i} z_W^j\right\|_F$$
$$\le (j-i+2)\left(1+\frac{1}{L}\right)^{3j} \kappa_1 \kappa_0^{2j-1} \left(\max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{l=i-1}^{j} d_l \prod_{h=1}^{j} \|W_h\|_2\right)\left(\max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{j} \|W_h\|_2 \sum_{h=1}^{j} \frac{\|U_h\|_2}{\|W_h\|_2}\right).$$

From layer i to layer $j+1$, we calculate the derivative on $z_W^{j+1} = \phi_{j+1}(W_{j+1} z_W^j)$ with respect to $W_i^{p,q}$:

$$\frac{\partial}{\partial W_i^{p,q}} \phi_{j+1}(W_{j+1} z_W^j) = \phi'_{j+1}(W_{j+1} z_W^j) \odot \left(W_{j+1} \frac{\partial}{\partial W_i^{p,q}} z_W^j\right),$$

where $\odot$ is the Hadamard product operator between two vectors. Then we have

$$\left\|\frac{\partial}{\partial W_i^{p,q}} z_W^{j+1} \Big|_{W_k+U_k,\forall 1\le k\le j+1} - \frac{\partial}{\partial W_i^{p,q}} z_W^{j+1}\right\|$$
$$= \left\|\phi'_{j+1}\big((W_{j+1}+U_{j+1}) z_{W+U}^j\big) \odot \left((W_{j+1}+U_{j+1}) \frac{\partial}{\partial W_i^{p,q}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j}\right) - \phi'_{j+1}(W_{j+1} z_W^j) \odot \left(W_{j+1} \frac{\partial}{\partial W_i^{p,q}} z_W^j\right)\right\|$$

By the Cauchy-Schwarz inequality and the triangle inequality, the above equation is at most

$$\left\|\phi'_{j+1}\big((W_{j+1}+U_{j+1}) z_{W+U}^j\big) - \phi'_{j+1}(W_{j+1} z_W^j)\right\| \left\|(W_{j+1}+U_{j+1}) \frac{\partial}{\partial W_i^{p,q}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j}\right\| \tag{39}$$

$$+ \left\|\phi'_{j+1}(W_{j+1} z_W^j)\right\| \left\|(W_{j+1}+U_{j+1}) \frac{\partial}{\partial W_i^{p,q}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} - W_{j+1} \frac{\partial}{\partial W_i^{p,q}} z_W^j\right\|. \tag{40}$$

Using the triangle inequality and the κ_1-Lipschitz of $\phi'_{j+1}(\cdot)$, equation (39) is at most

$$\kappa_1\Big(\|W_{j+1}+U_{j+1}\|_2 \left\| z^j_{W+U} - z^j_W \right\| + \|U_{j+1}\|_2 \left\| z^j_W \right\| \Big) \cdot \|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \Big|_{W_k+U_k, \forall 1\le k\le j} \right\|.$$

Using the triangle inequality and $|\phi'_{j+1}(\cdot)| \le \kappa_0$, equation (40) is at most

$$d_{j+1}\kappa_0 \cdot \left(\|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \Big|_{W_k+U_k, \forall 1\le k\le j} - \frac{\partial}{\partial W_i^{p,q}} z^j_W \right\| + \|U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \right\| \right).$$

From Claim A.6, we have the norm of the derivative of z^j_W with respect to $W_i^{p,q}$ bounded by its Lipschitzness in equation (30):

$$\left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \right\| \le (1+\frac{1}{L})^j \kappa_0^j \max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j} \|W_h\|_2 . \tag{41}$$

$$\left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \Big|_{W_k+U_k, \forall 1\le k\le j} \right\| \le (1+\frac{1}{L})^{2j-1} \kappa_0^j \max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j} \|W_h\|_2 . \tag{42}$$

Then, the Frobenius norm of the Jacobian is at most

$$\begin{aligned}
&\left\| \frac{\partial}{\partial W_i} z^{j+1}_W \Big|_{W_k+U_k, \forall 1\le k\le j+1} - \frac{\partial}{\partial W_i} z^{j+1}_W \right\|_F \\
\le& d_i d_{i-1} \cdot \kappa_1 \left(\|W_{j+1}+U_{j+1}\|_2 \left\| z^j_{W+U} - z^j_W \right\| + \|U_{j+1}\|_2 \left\| z^j_W \right\| \right) \|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \Big|_{W_k+U_k, \forall 1\le k\le j} \right\| & (43)\\
&+ d_{j+1}\kappa_0 \cdot \left(\|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \Big|_{W_k+U_k, \forall 1\le k\le j} - \frac{\partial}{\partial W_i^{p,q}} z^j_W \right\|_F + d_i d_{i-1} \|U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z^j_W \right\| \right). & (44)
\end{aligned}$$

From equation (36), (37), (42), we have equation (43) will be at most

$$\Big(1+\frac{1}{L}\Big)^{3j+3} \kappa_1 \kappa_0^{2j+1} \Big(\max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{l=i-1}^{j+1} d_l \prod_{h=1}^{j+1} \|W_h\|_2 \Big) \Big(\max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{j+1} \|W_h\|_2 \sum_{h=1}^{j+1} \frac{\|U_h\|_2}{\|W_h\|_2} \Big). \tag{45}$$

From the induction and equation (36) and (41), the term in equation (44) will be at most

$$(j-i+2)\Big(1+\frac{1}{L}\Big)^{3j+3} \kappa_1 \kappa_0^{2j+1} \Big(\max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{l=i-1}^{j+1} d_l \prod_{h=1}^{j+1} \|W_h\|_2 \Big) \Big(\max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{j+1} \|W_h\|_2 \sum_{h=1}^{j+1} \frac{\|U_h\|_2}{\|W_h\|_2} \Big). \tag{46}$$

After we combine equation (45) and (46), the proof is complete. □

Based on the above two results, we state the proof of Proposition A.5.

Proof of Proposition A.5. We will prove using induction. Let z^j_W be an j layer neural network with parameters W. The change in the Hessian of output of the j layer network z^j_W with respect to W_i under perturbation U can be bounded as follows:

$$\left\| \mathbf{H}_i[z^j_W]\Big|_{W_k+U_k, \forall 1\le k\le j} - \mathbf{H}_i[z^j_W] \right\|_F \le G_{i,j} \Big(\max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{j} \|W_h\|_2 \sum_{h=1}^{j} \frac{\|U_h\|_2}{\|W_h\|_2} \Big). \tag{47}$$

For any $j = 1, 2, \cdots, L$, $G_{i,j}$ has the following equation:

$$G_{i,j} = \frac{3}{2}(j-i+2)^2 \Big(1+\frac{1}{L}\Big)^{6j} \kappa_2 \kappa_1^2 \kappa_0^{3j-3} \prod_{l=i-1}^{j} d_l^3 \Big(\max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j} \|W_h\|_2 \Big)^2,$$

For any fixed $i = 1, 2, \cdots, L$, we will prove using induction with respect to j. If $j < i$, we have $\frac{\partial^2}{\partial W_i^2} z_W^j = 0$ since z_W^j remains unchanged regardless of any change of W_i. When $j = i$, we have

$$\left\| \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^i \Big|_{W_k + U_k, \forall 1 \le k \le i} - \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^i \right\|$$
$$= \left| \left([z_{W+U}^{i-1}]^q [z_{W+U}^{i-1}]^t \right) \phi_i'' \left(\sum_{r=1}^{d_{i-1}} \left((W_i^{p,r} + U_i^{p,r}) [z_{W+U}^{i-1}]^r \right) \right) - \left([z_W^{i-1}]^q [z_W^{i-1}]^t \right) \phi_i'' \left(\sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right) \right|,$$

where $p = s$. If $p \neq s$, we have $\frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^i = 0$, where $W_i^{p,q}, U_i^{p,q}$ are the (p, q)'th entry of W_i and U_i. Let $[z_W^{i-1}]^q$ be the q'th entry of z_W^{i-1}. Next, by the triangle inequality, the above equation is smaller than

$$\left| [z_{W+U}^{i-1}]^q [z_{W+U}^{i-1}]^t \right| \left| \phi_i'' \left(\sum_{r=1}^{d_{i-1}} \left((W_i^{p,r} + U_i^{p,r}) [z_{W+U}^{i-1}]^r \right) - \phi_i'' \left(\sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right) \right| \tag{48}$$

$$+ \left| [z_{W+U}^{i-1}]^q [z_{W+U}^{i-1}]^t - [z_W^{i-1}]^q [z_W^{i-1}]^t \right| \left| \phi_i'' \left(\sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right) \right|. \tag{49}$$

Since ϕ_i'' is κ_2-Lipschitz by Assumption A.3, equation (48) is at most

$$\left| [z_{W+U}^{i-1}]^q [z_{W+U}^{i-1}]^t \right| \cdot \kappa_2 \left| \sum_{r=1}^{d_{i-1}} (W_i^{p,r} + U_i^{p,r}) [z_{W+U}^{i-1}]^r - \sum_{r=1}^{d_{i-1}} W_i^{p,r} [z_W^{i-1}]^r \right|. \tag{50}$$

Since Assumption A.3 holds, we have $|\phi_i''(\cdot)| \le \kappa_1$ by the Lipschitz condition on the second-order derivatives. Thus, equation (49) is at most

$$\left| [z_{W+U}^{i-1}]^q [z_{W+U}^{i-1}]^t - [z_W^{i-1}]^q [z_W^{i-1}]^t \right| \cdot \kappa_1$$
$$\le \left(\left| [z_{W+U}^{i-1}]^q [z_{W+U}^{i-1}]^t - [z_W^{i-1}]^q [z_{W+U}^{i-1}]^t \right| + \left| [z_W^{i-1}]^q [z_{W+U}^{i-1}]^t - [z_W^{i-1}]^q [z_W^{i-1}]^t \right| \right) \cdot \kappa_1. \tag{51}$$

We use the triangle inequality in the last step. Taking equation (50) and (51) together, the Frobenius norm of the Hessian is

$$\left\| \mathbf{H}_i[z_W^i] \Big|_{W_k + U_k, \forall 1 \le k \le i} - \mathbf{H}_i[z_W^i] \right\|_F$$
$$\le d_i \left\| z_{W+U}^{i-1} \right\|^2 \cdot k_2 d_{i-1} \left\| (W_i + U_i) z_{W+U}^{i-1} - W_i z_W^{i-1} \right\| + \kappa_1 d_i \left\| z_{W+U}^{i-1} - z_W^{i-1} \right\| \left(\left\| z_{W+U}^{i-1} \right\| + \left\| z_W^{i-1} \right\| \right).$$

By the Lipschitz continuity of Z_{W+U}^{i-1} and equation (30) in Claim A.6, we have

$$\left\| z_{W+U}^{i-1} \right\| \le \left(1 + \frac{1}{L} \right)^{i-1} \kappa_0^{i-1} \max_{x \in \mathcal{X}} \|x\| \prod_{h=1}^{i-1} \|W_h\|_2. \tag{52}$$

$$\left\| z_{W+U}^{i-1} - z_W^{i-1} \right\| \le \left(1 + \frac{1}{L} \right)^{i-1} \kappa_0^{i-1} \max_{x \in \mathcal{X}} \|x\| \prod_{h=1}^{i-1} \|W_h\|_2 \sum_{h=1}^{i-1} \frac{\|U_h\|_2}{\|W_h\|_2}. \tag{53}$$

By the triangle inequality, we have

$$\left\| (W_i + U_i) z_{W+U}^{i-1} - W_i z_W^{i-1} \right\| \le \|W_i + U_i\|_2 \left\| z_{W+U}^{i-1} - z_W^{i-1} \right\| + \|U_i\|_2 \left\| z_W^{i-1} \right\|. \tag{54}$$

From equation (52), (53), and (54), we finally obtain that the Hessian of z_W^i with respect to W_i is at most

$$\begin{aligned}&\left\|\mathbf{H}_i[z_W^i]\Big|_{W_k+U_k,\forall 1\le k\le i}-\mathbf{H}_i[z_W^i]\right\|_F\\ \le&3\Big(1+\frac{1}{L}\Big)^{3i}\kappa_2\kappa_1\kappa_0^{3i-3}d_id_{i-1}\left(\max_{x\in\mathcal{X}}\|x\|\prod_{h=1}^{i-1}\|W_h\|_2\right)^2\left(\max_{x\in\mathcal{X}}\|x\|\prod_{h=1}^{i}\|W_h\|_2\sum_{h=1}^{i}\frac{\|U_h\|_2}{\|W_h\|_2}\right).\end{aligned}$$

Thus, we have known that equation (47) is correct when $j=i$ with respect to $\mathbf{H}_i$. Assume that (47) is correct for any i up to $j\ge i$ with respect to $\mathbf{H}_i$. We have the following inequality:

$$\begin{aligned}&\left\|\mathbf{H}_i[z_W^j]\Big|_{W_k+U_k,\forall 1\le k\le j}-\mathbf{H}_i[z_W^j]\right\|_F\\ \le&\frac{3}{2}(j-i+2)^2\Big(1+\frac{1}{L}\Big)^{6j}\kappa_2\kappa_1^2\kappa_0^{3(j-1)}\prod_{l=i-1}^{j}d_l^3\left(\max_{x\in\mathcal{X}}\|x\|\frac{1}{\|W_i\|_2}\prod_{h=1}^{j}\|W_h\|_2\right)^2\left(\max_{x\in\mathcal{X}}\|x\|\prod_{h=1}^{j}\|W_h\|_2\sum_{h=1}^{j}\frac{\|U_h\|_2}{\|W_h\|_2}\right).\end{aligned}$$

Then, we will calculate the twice-derivative on $z_W^{j+1}=\phi_{j+1}(W_{j+1}z_W^j)$ with respect to $W_i^{p,q}$ and $W_i^{s,t}$:

$$\begin{aligned}\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}\phi_{j+1}(W_{j+1}z_W^j)=&\phi''_{j+1}(W_{j+1}z_W^j)\odot\left(W_{j+1}\frac{\partial}{\partial W_i^{p,q}}z_W^j\right)\odot\left(W_{j+1}\frac{\partial}{\partial W_i^{s,t}}z_W^j\right)\\ &+\phi'_{j+1}(W_{j+1}z_W^j)\odot\left(W_{j+1}\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^j\right),\end{aligned}$$

where $\odot$ is the Hadamard product operator between two vectors. Then, we have

$$\begin{aligned}&\left\|\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^{j+1}\Big|_{W_k+U_k,\forall 1\le k\le j+1}-\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^{j+1}\right\|\\ =&\left\|\phi''_{j+1}((W_{j+1}+U_{j+1})z_{W+U}^j)\odot\left((W_{j+1}+U_{j+1})\frac{\partial}{\partial W_i^{p,q}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right)\odot\left((W_{j+1}+U_{j+1})\frac{\partial}{\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right)\right.\\ &+\phi'_{j+1}((W_{j+1}+U_{j+1})z_{W+U}^j)\odot\left((W_{j+1}+U_{j+1})\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right)\\ &\left.-\phi''_{j+1}(W_{j+1}z_W^j)\odot\left(W_{j+1}\frac{\partial}{\partial W_i^{p,q}}z_W^j\right)\odot\left(W_{j+1}\frac{\partial}{\partial W_i^{s,t}}z_W^j\right)-\phi'_{j+1}(W_{j+1}z_W^j)\odot\left(W_{j+1}\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^j\right)\right\|.\end{aligned}$$

By the Cauchy-Schwarz inequality and the triangle inequality, the above equation is at most

$$\left\|\phi''_{j+1}((W_{j+1}+U_{j+1})z_{W+U}^j)-\phi''_{j+1}(W_{j+1}z_W^j)\right\|\left\|(W_{j+1}+U_{j+1})\frac{\partial}{\partial W_i^{p,q}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right\|\left\|(W_{j+1}+U_{j+1})\frac{\partial}{\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right\| \tag{55}$$

$$+\left\|\phi''_{j+1}(W_{j+1}z_W^j)\right\|\left\|(W_{j+1}+U_{j+1})\frac{\partial}{\partial W_i^{p,q}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}-W_{j+1}\frac{\partial}{\partial W_i^{p,q}}z_W^j\right\|\left\|(W_{j+1}+U_{j+1})\frac{\partial}{\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right\| \tag{56}$$

$$+\left\|\phi''_{j+1}(W_{j+1}z_W^j)\right\|\left\|W_{j+1}\frac{\partial}{\partial W_i^{p,q}}z_W^j\right\|\left\|(W_{j+1}+U_{j+1})\frac{\partial}{\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}-W_{j+1}\frac{\partial}{\partial W_i^{s,t}}z_W^j\right\| \tag{57}$$

$$+\left\|\phi'_{j+1}((W_{j+1}+U_{j+1})z_{W+U}^j)-\phi'_{j+1}(W_{j+1}z_W^j)\right\|\left\|(W_{j+1}+U_{j+1})\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right\| \tag{58}$$

$$+\left\|\phi'_{j+1}(W_{j+1}z_W^j)\right\|\left\|(W_{j+1}+U_{j+1})\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}-W_{j+1}\frac{\partial^2}{\partial W_i^{p,q}\partial W_i^{s,t}}z_W^j\right\|. \tag{59}$$

Thus, using the triangle inequality and the fact that $\phi''(\cdot)$ is κ_2-Lipschitz, equation (55) is at most

$$\kappa_2\left(\|W_{j+1}+U_{j+1}\|_2\left\|z_{W+U}^j-z_W^j\right\|+\|U_{j+1}\|_2\left\|z_W^j\right\|\right)\|W_{j+1}+U_{j+1}\|_2^2\left\|\frac{\partial}{\partial W_i^{p,q}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right\|\left\|\frac{\partial}{\partial W_i^{s,t}}z_W^j\Big|_{W_k+U_k,\forall 1\le k\le j}\right\|.$$

Using the triangle inequality and $|\phi''(\cdot)| \le \kappa_1$, equation (56) is at most

$$\kappa_1 d_{j+1}\left(\|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} - \frac{\partial}{\partial W_i^{p,q}} z_W^j \right\| + \|U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z_W^j \right\| \right) \|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{s,t}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} \right\|.$$

Using the triangle inequality and $|\phi''(\cdot)| \le \kappa_1$, equation (57) is at most

$$\kappa_1 d_{j+1} \|W_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{p,q}} z_W^j \right\| \left(\|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{s,t}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} - \frac{\partial}{\partial W_i^{s,t}} z_W^j \right\| + \|U_{j+1}\|_2 \left\| \frac{\partial}{\partial W_i^{s,t}} z_W^j \right\| \right).$$

Using the triangle inequality and the fact that $\phi'(\cdot)$ is κ_1-Lipschitz, equation (58) is at most

$$\kappa_1 \left(\|W_{j+1}+U_{j+1}\|_2 \left\| z_{W+U}^j - z_W^j \right\| + \|U_{j+1}\|_2 \left\| z_W^j \right\| \right) \|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} \right\|.$$

Using the triangle inequality and that $|\phi'(\cdot)| \le \kappa_0$, equation (59) is at most

$$\kappa_0 d_{j+1} \left(\|W_{j+1}+U_{j+1}\|_2 \left\| \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} - \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^j \right\| + \|U_{j+1}\|_2 \left\| \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^j \right\| \right).$$

From Claim A.6, we know that the norm of the derivative of z_W^j with respect to $W_i^{p,q}$ is bounded by its Lipschitzness in equation (30):

$$\left\| \frac{\partial}{\partial W_i^{p,q}} z_W^j \right\| \le (1+\frac{1}{L})^j \kappa_0^j \max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j} \|W_h\|_2 . \tag{60}$$

$$\left\| \frac{\partial}{\partial W_i^{p,q}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} \right\| \le (1+\frac{1}{L})^{2j-1} \kappa_0^j \max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j} \|W_h\|_2 . \tag{61}$$

From Claim A.7, the norm of the twice-derivative of z_W^j with respect to $W_i^{p,q}$ and $W_i^{s,t}$ is bounded by its Lipschitzness in equation (31):

$$\left\| \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^j \right\| \le A_{i,j} \max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j} \|W_h\|_2 . \tag{62}$$

$$\left\| \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^j \Big|_{W_k+U_k,\forall 1\le k\le j} \right\| \le \left(1+\frac{1}{L}\right)^{2j-2} A_{i,j} \max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j} \|W_h\|_2 . \tag{63}$$

Then, the Frobenius norm of the Hessian $\mathbf{H}_i$ is at most

$$\left\| \mathbf{H}_i[z_W^{j+1}] \Big|_{W_k+U_k,\forall 1\le k\le j+1} - \mathbf{H}_i[z_W^{j+1}] \right\|_F \le (d_i d_{i-1})^2 \left\| \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^{j+1} \Big|_{W_k+U_k,\forall 1\le k\le j+1} - \frac{\partial^2}{\partial W_i^{p,q} \partial W_i^{s,t}} z_W^{j+1} \right\|$$

Since we have listed the upper bound of all the terms in the Hessian with respect to W_i from equation (52), (53), (60), (61), (62), (63), and the induction, equation (55) times $(d_i d_{i-1})^2$ is at most

$$\left(1+\frac{1}{L}\right)^{6(j+1)} \kappa_2 \kappa_1^2 \kappa_0^{3j} \prod_{l=i-1}^{j+1} d_l^3 \left(\max_{x\in\mathcal{X}} \|x\| \frac{1}{\|W_i\|_2} \prod_{h=1}^{j+1} \|W_h\|_2 \right)^2 \left(\max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{j+1} \|W_h\|_2 \sum_{h=1}^{j+1} \frac{\|U_h\|_2}{\|W_h\|_2} \right).$$

Equation (56),(57), and (58) times $(d_i d_{i-1})^2$ is at most

$$(j-i+2)\left(1+\frac{1}{L}\right)^{6(j+1)} \kappa_2 \kappa_1^2 \kappa_0^{3j} \prod_{l=i-1}^{j+1} d_l^3 \left(\max_{x\in\mathcal{X}} \|x\|_2 \frac{1}{\|W_i\|_2} \prod_{h=1}^{j+1} \|W_h\|_2 \right)^2 \left(\max_{x\in\mathcal{X}} \|x\| \prod_{h=1}^{j+1} \|W_h\|_2 \sum_{h=1}^{j+1} \frac{\|U_h\|_2}{\|W_h\|_2} \right).$$

Equation (59) times $(d_i d_{i-1})^2$ is at most

$$\frac{3}{2}(j-i+2)^2\Big(1+\frac{1}{L}\Big)^{6(j+1)}\kappa_2\kappa_1^2\kappa_0^{3j}\prod_{l=i-1}^{j+1} d_l^3\left(\max_{x\in\mathcal{X}}\|x\|\frac{1}{\|W_i\|_2}\prod_{h=1}^{j+1}\|W_h\|_2\right)^2\left(\max_{x\in\mathcal{X}}\|x\|\prod_{h=1}^{j+1}\|W_h\|_2\sum_{h=1}^{j+1}\frac{\|U_h\|_2}{\|W_h\|_2}\right).$$

After we combine above five terms, the proof of induction is complete. □

Now we are ready to state the proof of Lemma 2.4.

Proof of Lemma 2.4. Let $B,\epsilon>0$, $p=\sum_{i=1}^L d_i d_{i-1}$, and let $S=\{\mathrm{vec}(W)\in\mathbb{R}^p:\|\mathrm{vec}(W)\|_2\le B\}$. Then there exists an ϵ-cover of S with respect to the ℓ_2-norm at most $\max\left(\left(\frac{3B}{\epsilon}\right)^p,1\right)$ elements from Example 5.8, Wainwright [Wai19]. From Proposition A.5, the Hessian $\mathbf{H}_i[\ell(f_W(x),y)]$ is G-Lipschitz for all $(W+U),W\in S$ and any $i=1,2,\cdots,L$. Then we have

$$\|\mathbf{H}_i[\ell(f_{W+U}(x),y)]-\mathbf{H}_i[\ell(f_W(x),y)]\|_F\le G\|U\|_F,$$

where we can find G from equation (29). Here we follow the proof in Ma [Ma21]. For parameters $\delta,\epsilon>0$, let $\mathcal{N}$ be the ϵ-cover of S with respect to the ℓ_2-norm. Define event

$$E=\left\{\forall W\in\mathcal{N},\left\|\frac{1}{n}\sum_{j=1}^n\mathbf{H}_i[\ell(f_W(x_j),y_j)]-\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{H}_i[\ell(f_W(x),y)]]\right\|_F\le\delta\right\}.$$

By the matrix Bernstein inequality (cf. Theorem 6.1.1 in Tropp [Tro15]), we have

$$\Pr(E)\ge 1-4|\mathcal{N}|d_i d_{i-1}\exp(-n\delta^2/(2\mathcal{H}_i^2)).$$

Next, for any $W\in S$, we can pick some $W+U\in\mathcal{N}$ such that $\|U\|_F\le\epsilon$. We have

$$\left\|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{H}_i[\ell(f_{W+U}(x),y)]]-\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{H}_i[\ell(f_W(x),y)]]\right\|_F\le G\|U\|_F\le G\epsilon,\ \text{and}$$
$$\left\|\frac{1}{n}\sum_{j=1}^n\mathbf{H}_i[\ell(f_{W+U}(x_j),y_j)]-\frac{1}{n}\sum_{j=1}^n\mathbf{H}_i[\ell(f_W(x_j),y_j)]\right\|_F\le G\|U\|_F\le G\epsilon.$$

Therefore, for any $W\in S$, we obtain

$$\left\|\frac{1}{n}\sum_{j=1}^n\mathbf{H}_i[\ell(f_W(x_j),y_j)]-\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{H}_i[\ell(f_W(x),y)]]\right\|_F\le 2G\epsilon+\delta.$$

We also need to choose the suitable parameter δ and ϵ. First, set $\epsilon=\delta/(2G)$ so that conditional on E,

$$\left\|\frac{1}{n}\sum_{j=1}^n\mathbf{H}_i[\ell(f_W(x_j),y_j)]-\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{H}_i[\ell(f_W(x),y)]]\right\|_F\le 2\delta.$$

The event E happens with probability

$$1-4|\mathcal{N}|d_i d_{i-1}\exp(-n\delta^2/(2\mathcal{H}_i^2))=1-4d_i d_{i-1}\exp(\log|\mathcal{N}|-n\delta^2/(2\mathcal{H}_i^2)).$$

We have $\log|\mathcal{N}|\le p\log(3B/\epsilon)=p\log(6BG/\delta)$. If we set

$$\delta=\sqrt{\frac{4p\mathcal{H}_i^2\log(3\tau BGn/\mathcal{H}_i)}{n}}$$

with $\log(3\tau BGn/\mathcal{H}_i) \geq 1$ ($n \geq \frac{e\mathcal{H}_i}{3BG}, \tau \geq 1$). Then we get

$$\begin{aligned}
&p\log(6BG/\delta) - n\delta^2/(2\mathcal{H}_i^2)\\
=&p\log(\frac{6BG\sqrt{n}}{\sqrt{4p\mathcal{H}_i^2\log(3\tau BGn/\mathcal{H}_i)}}) - 2p\log(3\tau BGn/\mathcal{H}_i)\\
=&p\log(\frac{3BG\sqrt{n}}{\mathcal{H}_i\sqrt{p\log(3\tau BGn/\mathcal{H}_i)}}) - 2p\log(3\tau BGn/\mathcal{H}_i)\\
\leq&p\log(3\tau BGn/\mathcal{H}_i) - 2p\log(3\tau BGn/\mathcal{H}_i) && (\tau \geq 1, \log(3\tau BGn/\mathcal{H}_i) \geq 1)\\
=&-p\log(3\tau BGn/\mathcal{H}_i)\\
\leq&-p\log(e\tau). && (3BGn/\mathcal{H}_i \geq e)
\end{aligned}$$

Therefore, with probability greater than

$$1 - 4|\mathcal{N}|d_i d_{i-1}\exp(-n\delta^2/(2\mathcal{H}_i^2)) \geq 1 - 4d_i d_{i-1}(e\tau)^{-p} \geq 1 - 4d_i d_{i-1}/(e\tau),$$

we have

$$\left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{H}_i[\ell(f_W(x_j),y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{H}_i[\ell(f_W(x),y)]]\right\|_F \leq \sqrt{\frac{16p\mathcal{H}_i^2\log(3\tau BGn/\mathcal{H}_i)}{n}}.$$

Denote $\delta' = 4d_i d_{i-1}/(e\tau)$, $C_2 = 4\mathcal{H}_i\sqrt{p}$, and $C_3 = 12d_i d_{i-1}BG/(e\mathcal{H}_i)$. With probability greater than $1-\delta'$, we have the final result:

$$\left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{H}_i[\ell(f_W(x_j),y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{H}_i[\ell(f_W(x),y)]]\right\|_F \leq C_2\sqrt{\frac{\log(C_3 n/\delta')}{n}}.$$

□

A.4. Proof of Lemma 2.3

Claim A.8. *Let f_W be a feedforward neural network with the activation function ϕ_i for any $i = 1, 2, \cdots, L$, parametrized by the layer weight matrices W. Let U be any matrices that have the same dimension as W. For any W and U, the following identity holds*

$$\begin{aligned}
\ell(f_{W+U}(x),y) = \ell(f_W(x),y) &+ \langle U, \nabla_W(\ell(f_W(x),y))\rangle\\
&+ \langle \mathbf{vec}(U), \mathbf{H}[\ell(f_W(x),y)]\mathbf{vec}(U)\rangle + R_2(\ell(f_W(x),y)).
\end{aligned}$$

where $R_2(\ell(f_W(x),y))$ is a second-order error term in the Taylor's expansion.

Proof. The proof of Claim A.8 follows by the fact that the activation $\phi_i(x)$ is twice-differentiable. From the mean value theorem, let η be an matrix that has the same dimension as W and U. There must exist an η between W and $U+W$ to keep the following equality:

$$R_2(\ell(f_W(x),y)) = \langle \mathbf{vec}(U), \big(\mathbf{H}[\ell(f_W(x),y)]\mid_\eta - \mathbf{H}[\ell(f_W(x),y)]\big)\mathbf{vec}(U)\rangle.$$

□

Proof of Lemma 2.3. We take the expectation of U for both sides of the equation in Lemma A.8. The result becomes

$$\begin{aligned}
\mathbb{E}_U[\ell(f_{W+U}(x),y)] =&\mathbb{E}_U\big[\ell(f_W(x),y) + \langle U, \nabla_W(\ell(f_W(x),y))\rangle\\
&+ \langle \mathbf{vec}(U), \mathbf{H}[\ell(f_W(x),y)]\mathbf{vec}(U)\rangle + R_2(\ell(f_W(x),y))\big].
\end{aligned}$$

Then, we use the notation of the distribution $\mathcal{Q}$ on $\mathbb{E}_U[\ell(f_{W+U}(x), y)]$. We have

$$\begin{aligned}\ell_{\mathcal{Q}}(f_W(x), y) =& \mathbb{E}_U[\ell(f_W(x), y)] + \mathbb{E}_U[\langle U, \nabla_W(\ell(f_W(x), y))\rangle] \\ &+ \mathbb{E}_U[\langle \mathbf{vec}(U), \mathbf{H}[\ell(f_W(x), y)]\mathbf{vec}(U)\rangle] + \mathbb{E}_U[R_2(\ell(f_W(x), y))].\end{aligned}$$

Since $\mathbb{E}[U_i] = 0$ for every $i = 1, 2, \cdots, L$, the first-order term will be

$$\mathbb{E}_U[\langle U, \nabla_W(\ell(f_W(x), y))\rangle] = \sum_{i=1}^{L} \mathbb{E}_{U_i}[\langle U_i, \nabla_W(\ell(f_W(x), y))\rangle] = 0. \tag{64}$$

Since we assume that U_i and U_j are independent variables for any $i \neq j$, we have

$$\begin{aligned}&\mathbb{E}_U[\langle \mathbf{vec}(U), \mathbf{H}[\ell(f_W(x), y)]\mathbf{vec}(U)\rangle] \\ =& \sum_{i,j=1}^{L} \mathbb{E}_U[\langle \mathbf{vec}(U_i), \frac{\partial^2}{\partial \mathbf{vec}(W_i) \partial \mathbf{vec}(W_j)} \ell(f_W(x), y) \mathbf{vec}(U_j)\rangle] \\ =& \sum_{i=1}^{L} \mathbb{E}_U[\langle \mathbf{vec}(U_i), \mathbf{H}_i[\ell(f_W(x), y)]\mathbf{vec}(U_i)\rangle].\end{aligned}$$

For each U_i, let Σ_i be the variance of U_i. The second-order term will be

$$\mathbb{E}_{U_i}[\langle \mathbf{vec}(U_i), \mathbf{H}_i[\ell(f_W(x), y)]\mathbf{vec}(U_i)\rangle] = \big\langle \Sigma_i, \mathbf{H}_i[\ell(f_W(x), y)]\big\rangle. \tag{65}$$

The expectation of the error term $R_2(\ell(f_W(x), y))$ be the following

$$\begin{aligned}\mathbb{E}_U[R_2(\ell(f_W(x), y))] &= \mathbb{E}_U[\langle \mathbf{vec}(U)\mathbf{vec}(U)^\top, \big(\mathbf{H}[\ell(f_W(x), y)] \mid_\eta - \mathbf{H}[\ell(f_W(x), y)]\big)\rangle] \\ &= \sum_{i=1}^{L} \mathbb{E}_{U_i}[\langle \mathbf{vec}(U_i)\mathbf{vec}(U_i)^\top, \big(\mathbf{H}_i[\ell(f_W(x), y)] \mid_{\eta_i} - \mathbf{H}_i[\ell(f_W(x), y)]\big)\rangle].\end{aligned}$$

From Proposition A.5, let $\frac{\sqrt{3}}{9}C_1$ equal to the following number

$$\frac{3}{2}(L+1)^2 e^6 \kappa_2 \kappa_1^2 \kappa_0^{3L} \Big(\max_{x\in\mathcal{X}} \|x\|_2 \max_{1\le i\le L} \frac{1}{\|W_i\|_2} \prod_{h=0}^{L} d_h \prod_{h=1}^{L} \|W_h\|_2 \Big)^3.$$

Hence, we have the following inequality:

$$\begin{aligned}&\mathbb{E}_{U_i}[\langle \mathbf{vec}(U_i)\mathbf{vec}(U_i)^\top, \big(\mathbf{H}_i[\ell(f_W(x), y)] \mid_{\eta_i} - \mathbf{H}_i[\ell(f_W(x), y)]\big)\rangle] \\ \le& \mathbb{E}_{U_i}\Big[\|U_i\|_F^2 \, \|\mathbf{H}_i[\ell(f_W(x), y)] \mid_{\eta_i} - \mathbf{H}_i[\ell(f_W(x), y)]\|_F \Big] \\ \le& \mathbb{E}_{U_i}\Big[\|U_i\|_F^2 \, \frac{\sqrt{3}}{9} C_1 \|U_i\|_F \Big] \\ =& \frac{\sqrt{3}}{9} C_1 \mathbb{E}_{U_i}\Big[\|U_i\|_F^3 \Big].\end{aligned}$$

From Lemma 2 in Jin et al. [JNG+19], we have $(\mathbb{E}_{U_i} \|U_i\|_F^3)^{1/3} \le \sqrt{3}\,\|\Sigma_i\|_F$. Then, the error term $R_2(\ell(f_W(x), y))$ is smaller than $\frac{\sqrt{3}}{9} C_1 \sum_{i=1}^{L} (\sqrt{3}\,\|\Sigma_i\|_F)^3 = C_1 \sum_{i=1}^{L} \|\Sigma_i\|_F^{3/2}$. Thus, the proof is complete. □

B. Omitted Details in the Experiments

B.1. Experimental setup for measuring generalization

Figure 1 compares our Hessian distance measure with previous distance measures and generalization errors of fine-tuning. In this illustration, we fine-tune the ViT-Base [DBK+20] model on the Clipart dataset with weakly-supervised label noise.

We control the model distance from the initialization v_i in fine-tuning and sweep the range of distance constraints in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. For the distance measure, we evaluate the generalization bound from Pitas et al. [PDV17]. For the scale of the bound is orders of magnitude larger than the generalization error, we show their values divided by 10^{74} in Figure 1. For Hessian distance measure, we evaluate the quantity $\sqrt{C}\sum_{i=1}^{L}\sqrt{\mathcal{H}_i}/\sqrt{n}$ from Equation 3. The generalization error is measured by the difference between averaged test loss and averaged training loss.

Figure 2a and 2b compare our generalization measure from Theorem 2.1 with the empirical generalization errors in fine-tuning ResNet. We use ResNet-50 models fine-tuned on the Indoor and CalTech-256 datasets. We study the following regularization method: (1) fine-tuning with early stopping; (2) weight decay; (3) label smoothing [MKH19]; (4) Mixup [ZCD+18]; (5) ℓ^2 soft penalty [LGD18]; and (6) ℓ^2 projection gradient methods [GHP21]. Figure 2c compares our generalization measure with the generalization errors in BERT fine-tuning. We consider BERT-Base on the MRPC dataset. We study the following regularization methods: (1) early stopping; (2) weight decay. (3) ℓ^2 soft penalty [LGD18]. (4) ℓ^2 projected gradient descent [GHP21]. Figure 2d numerically measures our Hessian distances $\sum_i^L\sqrt{\mathcal{H}_i}$ on ResNet-50 and BERT-Base models. We measure these quantities for models fine-tuned with early stopping (implicit regularization) and ℓ^2 projection gradient methods [GHP21] (explicit regularization). Hyper-parameters in the fine-tuning algorithms are selected based on the accuracy of the validation dataset.

Figure 3 measures the traces of the loss's Hessian by varying the noisy label. We fine-tune both ResNet-18 and ResNet-50 on CIFAR-10 and CIFAR-100. Each subfigure is a matrix corresponding to ten classes. We use the 10 classes for CIFAR-10 and randomly select 10 classes for CIFAR-100. The (i, j)-th entry of the matrix is the trace of the loss's Hessian calculated from using noisy label $\tilde{y} = j$ on samples with true label i. For presentation, we select samples whose loss value under the true label is less than 1e-4. The trace values are then averaged over 200 samples from each class.

Figure 4 shows our Hessian distance measure of models fine-tuned with different algorithms, including early stopping, label smoothing [MKH19], Mixup [ZCD+18], SupCon [GDC+20], SAM [FKM+21], and our algorithm. We fine-tune ResNet-18 models on the Clipart dataset with weakly-supervised label noise. The Hessian distance measures are evaluated as $\sum_{i=1}^{L}\sqrt{\mathcal{H}_i}$ from Equation 3.

Table 5 compares the our generalization measure from Equation 3 with previous generalization bounds. We report the results of fine-tuning ResNet-50 on CIFAR-10 and CIFAR-100. We also include the results of fine-tuning BERT-Base on MRPC and SST-2. We describe the precise quantity for previous results:

- Theorem 1 from Gouk et al. [GHP21]: $\frac{1}{\sqrt{n}}\prod_{i=1}^{L}(2||W_i||_\infty)\sum_{i=1}^{L}(\frac{||W_i-W_i^{(s)}||_\infty}{||W_i||_\infty})$ where $||\cdot||_\infty$ is the maximum absolute row sum matrix norm.
- Theorem 4.1 from Li and Zhang [LZ21]: $\sqrt{\frac{1}{\epsilon^2 n}\left(\sum_{i=1}^{L}\frac{\prod_{j=1}^{L}(B_j+D_j)}{B_i+D_i}\right)^2\left(\sum_{i=1}^{L}D_i^2\right)}$ where ϵ is a small constant value set as 0.1 of the average training loss, $B_i = ||W_i^{(s)}||_2$, and $D_i = ||W_i - W_i^{(s)}||_F$.
- Theorem 2.1 from Long and Sedghi [LS20]: $\sqrt{\frac{M}{n}\prod_{i=1}^{L}||W_i||_2^2\sum_{i=1}^{L}||W_i - W_i^{(s)}||_2}$ where M is the total number of trainable parameters.
- Theorem 1 from Neyshabur et al. [NBS18]: $\sqrt{\frac{1}{\gamma^2 n}\prod_{i=1}^{L}||W_i||_2^2\sum_{i=1}^{L}\frac{||W_i||_F^2}{||W_i||_2^2}}$
- Theorem 4.2 from Pitas et al. [PDV17]: $\sqrt{\frac{1}{\gamma^2 n}\prod_{i=1}^{L}||W_i||_2^2\sum_{i=1}^{L}\frac{||W_i-W_i^{(s)}||_F^2}{||W_i||_2^2}}$
- Theorem 5.1 from Arora et al. [AGN+18]: $\sqrt{\frac{1}{\gamma^2 n}\max_x||f_w(x)||_2^2\sum_{i=1}^{L}\frac{\beta^2 c_i^2\lceil\kappa/s\rceil}{\mu_{i\rightarrow}^2\mu_i^2}}$

The margin bounds require specifying the desired margin γ. In Figure 5, we show the classification error rates, i.e., the margin loss values, of fine-tuned ResNet-50 on CIFAR-10 and CIFAR-100 with regard to the margin. The results show that the classification error depends heavily on γ. For a fair comparison, we select the largest margin so that the difference between margin loss and zero-one loss is less than 1%. This leads to a margin of 0.4 on CIFAR-10 and 0.01 on CIFAR-100.

Optimization procedure. In the experiments on image classification datasets, we fine-tune the model for 30 epochs. We use Adam optimizer with learning rate 1e-4 and decay the learning rate by 10 every 10 epochs. In the experiments on text

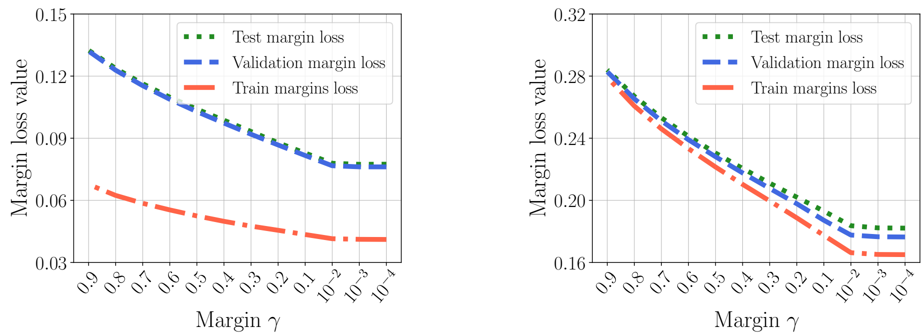


Figure 5: Evaluating the margin loss with various margins for a fine-tuned network on CIFAR-10 and CIFAR-100 datasets.

classification datasets, we fine-tune the BERT-Base model for 5 epochs. We use Adam optimizer with an initial learning rate of 5e-4 and then linearly decay the learning rate.

B.2. Experimental setup for fine-tuning with noisy labels

Datasets. We evaluate our algorithm for both image and text classification tasks. We use six image classification tasks derived from the DomainNet dataset [PBX+19]. The DomainNet dataset contains images from 345 different classes in 6 different domains, including Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. We use the same data processing as in Mazzetto et al. [MCS+21] and use 5 classes randomly selected from the 25 classes with the largest number of instances. We also use MRPC dataset from the GLUE benchmark [WSM+18].

We consider independent and correlated label noise in our experiments. Independent label noise is generated by flipping the labels of a given proportion of training samples to other classes uniformly. Correlated label noise is generated following the Mazzetto et al. [MCS+21] for annotating images from the DomainNet dataset. For completeness, we include a description of the protocol. For each domain, we learn a multi-class predictor for the 5 classes. The predictor is trained by fine-tuning a pretrained ResNet-18 network [HZR+16], using 60% of the labeled data for that domain. For each domain, we consider the classifiers trained in other domains as weak supervision sources. We generate noisy labels by taking majority votes from the weak supervision sources. Table 6 shows the statistics of the dataset.

Models. For the experiments on the six tasks from DomainNet, we fine-tune ResNet-18 and ResNet-101 with the Adam optimizer. We use an initial learning rate of 0.0001 for 30 epochs and decay the learning rate by 0.1 every 10 epochs. For the experiments on MRPC, we fine-tune the RoBERTa-Base model for 10 epochs. We use the Adam optimizer with an initial learning rate of 5e-4 and then linearly decay the learning rate. We report the Top-1 accuracy on the test set. We average the result over 10 random seeds.

Implementation. In our algorithm, we use the confusion matrix F obtained by the estimation method from Yao et al. [YLH+20]. For applying regularization constraints, we use the layerwise regularization method in Li and Zhang [LZ21]. The distance constraint parameter $D_i = \gamma^{i-1} D$ is set for layer i in equation (4). γ is a hyperparameter controlling the scaling of the constraints, and D is the constraint set for layer 1. We search the distance constraint parameter D in $[0.05, 10]$ and the scaling parameter γ in $[1, 5]$. For the results reported in Table 3, we searched the hyper-parameters for 30 times via cross-validation. We use the Optuna [ASY+19] package to search for the two hyper-parameters.

For the baselines, we report the results from running their open-sourced implementations. The hyper-parameters for each baseline are as follows.

Table 6: Basic statistics for six datasets with noisy labels [MCS+21].

	DomainNet					
	Clipart	Infograph	Painting	Quickdraw	Real	Sketch
Number of training Samples	750	858	872	1500	1943	732
Number of validation Samples	250	286	291	500	648	245
Number of test Samples	250	287	291	500	648	245
Noise rate	0.4147	0.6329	0.4450	0.6054	0.3464	0.4768

Table 7: Removing either component in Algorithm 1 degrades the test accuracy. Results are averaged over 10 random seeds.

Correlated label noise	DomainNet, ResNet-18					
	Clipart	Infograph	Painting	Quickdraw	Real	Sketch
Using only loss reweighting	77.96±1.57	41.78±2.86	72.13±1.66	49.10±2.42	85.34±1.51	64.78±1.60
Using only regularization	82.84±1.11	42.34±1.92	75.81±2.07	49.76±2.86	91.67±0.85	65.96±2.73
Ours	**83.28±1.64**	**43.38±2.45**	**76.32±1.08**	50.32±2.74	**92.36±0.78**	**66.86±3.29**

- For label smoothing [MKH19], we search the hyper-parameter α in $\{0.2, 0.4, 0.6, 0.8\}$.
- For Mixup [ZCD+18], we search the hyper-parameter α in $\{0.2, 0.4, 0.6, 0.8, 1.0, 2.0\}$.
- For FixMatch [SBL+20], we adopt it on fine-tuning from noisy labels. We set a starting epoch to apply FixMatch algorithm. We search its pseudo-labeling threshold in $\{0.7, 0.8, 0.9\}$. The starting epoch is searched in $\{2, 4, 6, 8\}$.
- For APL [MHW+20], we choose the active loss as normalized cross-entropy loss and the passive loss as reversed cross-entropy loss. We search the loss factor α in $\{1, 10, 100\}$ and β in $\{0.1, 1.0, 10\}$.
- For ELR [LNR+20], we search the momentum factor β in $\{0.5, 0.7, 0.9, 0.99\}$ and the weight factor λ in $\{0.05, 0.3, 0.5, 0.7\}$.
- For SAT [HZZ20], the start epoch is searched in $\{3, 5, 8, 10, 13\}$, and the momentum is searched in $\{0.6, 0.8, 0.9, 0.99\}$.
- For GJS [EA21], we search the weight in generalized Jensen-Shannon Divergence in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$.
- For DualT[YLH+20], we search the epoch to start reweighting in $\{1, 3, 5, 7, 9\}$.
- For SupCon[GDC+20], we search the contrastive loss weighting parameter λ in $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ and the temperature parameter τ in $\{0.1, 0.3, 0.5, 0.7\}$.
- For SAM [FKM+21], we search the neighborhood size parameter ρ in $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$.