Normalized Flat Minima: Exploring Scale Invariant Definition of Flat Minima for Neural Networks Using PAC-Bayesian Analysis

Yusuke Tsuzuku [1 2 *] **Issei Sato** [1 2] **Masashi Sugiyama** [2 1]

Abstract

The notion of flat minima has gained attention as a key metric of the generalization ability of deep learning models. However, current definitions of flatness are known to be sensitive to parameter rescaling. While some previous studies have proposed to rescale flatness metrics using parameter scales to avoid the scale dependence, the normalized metrics lose the direct theoretical connections between flat minima and generalization. We first provide generalization error bounds using existing normalized flatness measures for smooth and stochastic networks using second-order approximation. Using the analysis, we then propose a novel normalized flatness metric. The proposed metric enjoys both direct theoretical connections and better empirical correlation to generalization error.

1. Introduction

Deep learning methods have achieved significant performance improvement in many domains, such as computer vision, language processing, and speech processing (Krizhevsky et al., 2012; Devlin et al., 2019; van den Oord et al., 2018). However, we are still on the way to understand when they perform well on unseen data. Better understanding of the generalization performance of deep learning methods would help improve their performance. Deep learning community has made tremendous effort to understand generalization of neural networks, both theoretically and empirically (Zhang et al., 2017; Neyshabur et al., 2017; Arora et al., 2018).

The notion of "flat minima" has gained attention as a possible explanation for the generalization ability of deep learning methods (Hochreiter & Schmidhuber, 1997). Many empirical studies have supported the usefulness of this notion. For example, the notion shed light on the reason for larger generalization gaps in large-batch training (Keskar et al., 2017; Yao et al., 2018), and also inspired various training methods (Hochreiter & Schmidhuber, 1997; Chaudhari et al., 2017; Hoffer et al., 2017). Notably, Jiang et al. (2020) suggested that PAC-Bayes based generalization metrics, which have deep connections with flatness, are effective and promising among many other generalization metrics. As measures of flatness, previous studies proposed the volume of the region in which a network keeps roughly the same loss value (Hochreiter & Schmidhuber, 1997), the maximum loss value around minima (Keskar et al., 2017), and the spectral norm of the Hessian (Yao et al., 2018).

Despite the empirical connections of "flatness" to generalization, current definitions of flatness suffer from unnecessary scale dependence. Dinh et al. (2017) showed that we can arbitrarily change the flatness of the loss-landscape for some networks without changing the functions represented by the networks. Such scale dependence appears in networks with ReLU activation functions or normalization layers such as batch-normalization (Ioffe & Szegedy, 2015) and weight-normalization (Salimans & Kingma, 2016). A major counterargument to the scale dependence is that flatness provides valid generalization error bounds when parameter scales are taken into account (Dziugaite & Roy, 2017; Neyshabur et al., 2017). However, both the flatness and parameter scales have weak correlations with the generalization error in practical settings (Sec. 8).

Previous studies have tried to make flatness metrics invariant with respect to parameter scaling. Li et al. (2018) has successfully visualized connections between empirical performances and flatness hypotheses by scaling the loss-landscape using parameter scales. The success suggests that the normalized loss-landscape better captures networks' generalization and motivates us to provide their theoretical interpretations. Neyshabur et al. (2017), Achille & Soatto (2018), and Liang et al. (2019) also suggested scaling flatness metrics by parameter scales. While these metrics are

[*] Author is currently working at Preferred Networks, Inc. [1] The University of Tokyo, Tokyo, Japan [2] RIKEN AIP, Tokyo, Japan. Correspondence to: Yusuke Tsuzuku <tsuzuku@ms.k.u-tokyo.ac.jp>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

invariant with respect to known scaling issues, they did not provide direct connections between the metrics and generalization error. From the PAC-Bayesian perspective, providing scale-invariant bounds is not straightforward because we need to use the log-uniform distribution as a prior so that the Kullback-Leibler divergence term in PAC-Bayes bounds becomes scale-invariant (Kingma et al., 2015), while the use of the prior invalidates the PAC-Bayesian analysis.

On the way to understand generalization of deep learning and its connections to a normalized loss-landscape, this paper makes two contributions.

- We provide PAC-Bayesian generalization error bounds using scale-invariant loss curvature metrics.
- We propose a novel normalized loss curvature metric that has a close connection to the bounds.

This paper is organized as follows. We start our discussion with the PAC-Bayesian generalization error bound and its connection to flat minima (Sec. 3). Next, we connect a flatness metric normalized by parameter scales per parameter to the PAC-Bayesian analysis (Sec. 4). Unfortunately, per-parameter normalization yields a constant term proportional to the number of parameters in the generalization error bounds, which makes them vacuous for deep learning methods. Thus, we re-analyze the known scale dependence issue (Sec. 5). We identify that we do not need to scale the loss-landscape per-parameter, but we only need to scale it per node. Then, applying the analysis, we improve the method of loss-landscape normalization (Sec. 6). Using the novel normalized flatness definition, we provide generalization error bounds that do not have constant terms that scale with the number of parameters. Empirically, the proposed metric can predict the generalization performance of models more accurately than current unnormalized metrics (Sec. 8).

Remark: We provide the generalization bounds for stochastic networks. Most of them rely on the quadratic approximation of the loss function, but not all of them[1].

2. Related work

The notion of flat minima has its origin in the minimum description length principle (MDL) (Hinton & van Camp, 1993; Rissanen, 1986; Honkela & Valpola, 2004). Dziugaite & Roy (2017) connected flat minima to PAC-Bayesian arguments, which is a generalization of the MDL. They pointed out that the sharpness of local minima is not sufficient for measuring generalization, and that parameter scales need to be taken into account. PAC-Bayesian analysis has also been used for generalization analysis outside the context of flat minima (Neyshabur et al., 2018; Letarte et al., 2019). Neyshabur et al. (2018) analyzed the propagation of parameters' perturbations to bound the generalization error. Given the existence of adversarial examples (Szegedy et al., 2014), however, this approach inevitably provides vacuous bounds. Instead, in this paper, we stick to flat-minima arguments, which will better capture the effect of parameter perturbations.

Since Dinh et al. (2017) pointed out the scale dependences of flat minima, many studies have attempted to resolve the issue. A common approach is scaling sharpness metrics by parameter scales. Neyshabur et al. (2017) pointed out that the scale dependence can be removed by balancing parameter scales and sharpness metrics. However, their proposal requires using data-dependent priors (Guedj, 2019; Parrado-Hernández et al., 2012; Dziugaite & Roy, 2018) or equivalent alternatives, which adds non-trivial costs to the generalization bounds. In this paper, we make the costs explicit in Sec. 3 and later reduce them in Sec. 6. Prior to Neyshabur et al. (2017), Dziugaite & Roy (2017) tried to achieve balance using numerical optimization. While they partially mitigated the scale dependence, they could not completely overcome scale dependence because the prior variance was tied into whole network. Wang et al. (2018) explored a better choice of posterior variance. However, not only their analysis could not remove scale dependence, it was based on a parameter-wise argument, which involved a factor that scales with the number of parameters, making their bound as vacant as naive parameter counting[2]. Our analysis overcomes both problems.

Sharpness metrics normalization has also appeared outside the PAC-Bayesian context. Li et al. (2018) proposed rescaling the loss-landscape by filters' scales of convolutional layers. Even though they provided useful visualizations and insights, it is unclear how the scaling relates to generalization. Our work bridges the gap between the emprical success and theoretical understandings. Achille & Soatto (2018) proposed a notion of information in weights, which has connections to flat minima and PAC-Bayes. However, their arguments were based on an improper log-uniform prior and did not provide valid generalization error bounds. Even if we avoid the log-uniform prior to use a method described in Neklyudov et al. (2017), the notion does not provide non-vacuous bounds, as we show in Sec. 4. Liang et al. (2019) proposed the Fisher-Rao norm, which is a metric invariant with respect to known scaling issues. While the Fisher-Rao norm has interesting functional equivalence, in addition to invariance, it has been hard to directly connect the metric to generalization.

[1] Sec. 6 briefly explains that the stochastic network becomes close to a deterministic network as sample size increases.

[2] More comparison with Wang et al. (2018) is available in Appendix L in the supplementary material.

We propose a novel scale-invariant sharpness metric for loss-landscapes. Our definition distinguishes itself by its direct connection to PAC-Bayesian generalization error bounds. While previous methods for normalization have been based on per-parameter normalization, ours is based on per-node normalization. This difference is critical from the viewpoint of our PAC-Bayesian analysis in terms of the tightness of the bounds.

3. Flat minima from PAC-Bayesian perspective

This section reviews a fundamental PAC-Bayesian generalization error bound and its connection to flat minima provided by prior work. Table 3 in Appendix A in the supplementary material summarizes the notations used in this paper.

3.1. PAC-Bayesian generalization error bound

The following is one of the most basic PAC-Bayesian generalization error bounds (Germain et al., 2016; Alquier et al., 2016). Let $\mathcal{D}$ be a distribution over input-output space $\mathcal{Z}$, S be a set of m samples drawn *i.i.d.* from $\mathcal{D}$, $\mathcal{H}$ be a set of parameters $\boldsymbol{\theta}$, $\ell : X \times \mathcal{H} \rightarrow [0, 1]$ be a loss function, $\mathbb{P}$ be a distribution over $\mathcal{H}$ independent of S, and $\mathbb{Q}$ be a distribution over $\mathcal{H}$. For any $\delta \in (0, 1]$, any distribution $\mathbb{Q}$, and any nonnegative real number λ, with probability at least $1 - \delta$,

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) + \frac{\lambda}{2m} + \frac{1}{\lambda}\ln\frac{1}{\delta} + \frac{1}{\lambda}\mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right) \parallel \mathbb{P}\left(\boldsymbol{\theta}\right)], \tag{1}$$

where

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) := \mathop{\mathbb{E}}_{z\sim\mathcal{D},\theta\sim\mathbb{Q}\left(\boldsymbol{\theta}\mid\bar{\boldsymbol{\theta}}\right)}\left[\ell(z,\theta)\right], \tag{2}$$

$$\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) := \mathop{\mathbb{E}}_{z\sim\mathcal{S},\theta\sim\mathbb{Q}\left(\boldsymbol{\theta}\mid\bar{\boldsymbol{\theta}}\right)}\left[\ell(z,\theta)\right]. \tag{3}$$

We reorganize the PAC-Bayesian bound (1) for later use as

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\lambda}{2m} + \frac{1}{\lambda}\ln\frac{1}{\delta} + \underbrace{\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) - \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)}_{\text{(A)}} + \underbrace{\frac{1}{\lambda}\mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right) \parallel \mathbb{P}\left(\boldsymbol{\theta}\right)]}_{\text{(B)}}. \tag{4}$$

Similar decompositions can be found in prior work (Dziugaite & Roy, 2017; Neyshabur et al., 2017; 2018). We use a different PAC-Bayes bound (1) for later analysis, but they are essentially the same.[3] Flat minima, which are the noise-stable solution with respect to parameters, naturally correspond to term (A) in Eq. (4). Following prior work (Langford & Caruana, 2002; Hochreiter & Schmidhuber, 1997; Dziugaite & Roy, 2017; Arora et al., 2018), we analyze the true error of the stochastic classifier $\mathbb{Q}$ in the following sections. Nagarajan & Kolter (2019) presented a method to generalize the PAC-Bayes bounds to deterministic classifiers. We will leave combining our work and theirs as future work.

3.2. Effect of noise under second-order approximation

To connect PAC-Bayesian analysis with the Hessian of the loss-landscape as in prior work (Keskar et al., 2017; Dinh et al., 2017; Yao et al., 2018), we consider the second-order approximation of some surrogate loss functions. We use a Gaussian with a covariance matrix $\sigma^2 I$ as the posterior of parameters[4]. Then term (A) in the PAC-Bayesian bound (4) can be calculated as

$$\begin{aligned}
&\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) - \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma I)}\left[\mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}} + \boldsymbol{\epsilon}\right)\right] - \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right) \qquad (5) \\
&\approx \frac{1}{2}\mathrm{Tr}\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)\sigma^2. \qquad (6)
\end{aligned}$$

Thus, we can approximate term (A) by the trace of the Hessian and use it as a generalization metric. The effect of the approximation is further discussed in Appendix H in the supplementary material. Note, connecting PAC-Bayes with Hessian appears in the literature repeatedly (Dziugaite & Roy, 2017; Wang et al., 2018). By tuning σ with an appropriate prior, we can balance terms (A) and (B) (Dziugaite & Roy, 2017; Neyshabur et al., 2017). However, appropriate methods to balance them have not been extensively studied. Moreover, while some prior work proposed scaling σ by parameter scales, the operation has not been justified from the PAC-Bayesian viewpoint. We address the issues in the next section.

4. PAC-Bayes and parameter-wise normalization

In this section, we propose a framework that connects parameter-wisely normalized flatness metrics to PAC-Bayesian generalization error analysis. As a first step, we virtually decompose the parameters:

$$\boldsymbol{\theta} = \eta\boldsymbol{\mu} \odot \mathrm{e}^{[\boldsymbol{\sigma}]}, \tag{7}$$

[3] To apply our analysis, we can also use other PAC-Bayesian bounds such as Theorem 1.2.6 in Catoni (2007), which is known to be relatively tight in some cases and successfully provided empirically nontrivial bounds for ImageNet scale networks in Zhou et al. (2019).

[4] Remark: Eq. (1) holds with not only Gaussian posteriors but also arbitrary posteriors.

where $\odot$ is the Hadamard product, and $\mathrm{e}^{[\cdot]}$ is an element-wise exponential. The codomain of random variables $\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ is $\mathbb{R}^n$, and a hyperparameter η is a positive real number $\eta \in \mathbb{R}_{>0}$. Intuitively, this is a scale-sign decomposition of parameters with continuous relaxation. We extend the decomposition to a scale-direction decomposition in Sec. 6. We design a specific prior and posterior on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ to mitigate the scale dependence. We first define our prior design.

$$\mathbb{P}\left(\boldsymbol{\mu}\right) = \mathcal{N}\left(\boldsymbol{\mu} \mid \mathbf{0}, \boldsymbol{I}\right), \tag{8}$$

$$\mathbb{P}\left(\boldsymbol{\sigma}\right) = \mathcal{N}\left(\boldsymbol{\sigma} \mid \boldsymbol{\beta}, \eta' \boldsymbol{I}\right), \tag{9}$$

where $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\eta' \in \mathbb{R}_{>0}$ are hyperparameters. Next, we define our posterior design.

$$\mathbb{Q}\left(\boldsymbol{\mu} \mid \bar{\boldsymbol{\mu}}\right) = \mathcal{N}\left(\boldsymbol{\mu} \mid \bar{\boldsymbol{\mu}}, \boldsymbol{I}\right), \tag{10}$$

$$\mathbb{Q}\left(\boldsymbol{\sigma} \mid \bar{\boldsymbol{\sigma}}\right) = \mathcal{N}\left(\boldsymbol{\sigma} \mid \bar{\boldsymbol{\sigma}}, \eta' \boldsymbol{I}\right), \tag{11}$$

where $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\sigma}}$ are real vectors in $\mathbb{R}^n$ satisfying

$$\bar{\boldsymbol{\theta}} = \eta \bar{\boldsymbol{\mu}} \odot \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]}, \tag{12}$$

where a real vector $\bar{\boldsymbol{\theta}}$ is a parameter assignment. Note that the choice of $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\sigma}}$ is not necessarily unique, and we can balance them arbitrarily. Furthermore, the parameter η was introduced to address a technical issue concerning PAC-Bayesian analysis.

Under the second-order approximation, terms (A) and (B) in the PAC-Bayesian bound (1) can now be approximated as follows[5].

$$\begin{aligned}\text{(A)}: \quad & \frac{\eta^2}{2} \sum_{i=1}^{n} \left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \mathrm{e}^{2\bar{\boldsymbol{\sigma}}_i} \\ & + \frac{(\eta')^2}{2} \sum_{i=1}^{n} \left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \left(\bar{\boldsymbol{\theta}}_i\right)^2 + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right),\end{aligned} \tag{13}$$

$$\text{(B)}: \quad \frac{1}{(2\lambda)} \left(\frac{1}{\eta^2} \left\| \bar{\boldsymbol{\theta}} \odot \mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}]} \right\|_2^2 + \frac{1}{(\eta')^2} \left\| \bar{\boldsymbol{\sigma}} - \boldsymbol{\beta} \right\|_2^2 \right), \tag{14}$$

where $\mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right)$ is an error term introduced by the second-order approximation. We obtain the following bound by combining (13) and (14), optimizing $\boldsymbol{\sigma}$, and taking the union bound with respect to η, η', and $\boldsymbol{\beta}$.

Proposition 4.1. *For any networks, any $\lambda \in \Lambda$ and $\lambda' \in \Lambda'$ where $\Lambda \subset \mathbb{R}_{>0}$ and $\Lambda' \subset \mathbb{R}_{>0}$ are a finite set of real numbers independent of S, any 0–1 bounded twice continuously differentiable loss function with respect to parameters, any finite set of real numbers $\mathcal{B} \subset \mathbb{R}$ independent of S, and any parameter assignment $\bar{\boldsymbol{\theta}}$, there exists a stochastic classifier $\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)$ such that with probability $1-\delta$ over the training set and $\mathbb{Q}$, the expected error $\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\right)$ is bounded by*

$$\begin{aligned}\mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right) + \frac{\lambda'}{2\sqrt{\lambda}} \mathrm{FR} + \frac{1}{2\sqrt{\lambda}} \mathrm{NS}_0\left(\lambda'\right) \\ + \frac{1}{\lambda} \Xi + \frac{\lambda}{2m} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right),\end{aligned} \tag{15}$$

where

$$\mathrm{FR} = \sum_{i=1}^{n} \left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \bar{\boldsymbol{\theta}}_i^2, \tag{16}$$

$$\begin{aligned}&\mathrm{NS}_0 = \\ &\sum_{i=1}^{n} \min_{\substack{\bar{\boldsymbol{\sigma}}_i \in \mathbb{R} \\ \boldsymbol{\beta}_i \in \mathcal{B}}} \Big[\quad \left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \mathrm{e}^{2\bar{\boldsymbol{\sigma}}_i} + \frac{\bar{\boldsymbol{\theta}}_i^2}{\mathrm{e}^{2\bar{\boldsymbol{\sigma}}_i}} + \frac{\left(\bar{\boldsymbol{\sigma}}_i - \boldsymbol{\beta}_i\right)^2}{\lambda'} \Big],\end{aligned} \tag{17}$$

and

$$\Xi = \ln \frac{1}{\delta} + n \ln |\mathcal{B}| + 2 \ln |\Lambda| + \ln |\Lambda'|. \tag{18}$$

The optimal choice of λ is $O(m^{-\frac{1}{2}})$.

The FR and NS in the theorem stands for Fisher-Rao and normalized sharpness respectively, where the latter will be elaborated later and defined in Sec. 7. The convergence rate of the whole bound is $O(m^{-1/4})$ with the optimal choice of λ. This is not worse than existing bounds such as those of Bartlett et al. (2017), Neyshabur et al. (2018), and Arora et al. (2018). Appendix I in the supplementary material further discusses the convergence rate. The second term, containing FR, coincides with the second-order approximation of the scaled expected sharpness suggested by Neyshabur et al. (2017), the information in weight (Achille & Soatto, 2018), Gauss-Newton norm (Zhang et al., 2019), and Fisher-Rao norm under appropriate regularity conditions (Liang et al., 2019). The third term, containing NS_0, also has a connection to normalized loss-landscapes given by

$$\mathrm{NS}_1 := \frac{1}{2} \lim_{\lambda' \to \infty} \mathrm{NS}_0\left(\lambda'\right) = \sum_{i=1}^{n} \sqrt{\left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \left(\bar{\boldsymbol{\theta}}_i\right)^2}. \tag{19}$$

While this equation only holds at the limit, we can match the sides of the equation by modifying the prior and posterior design. We also can remove the second term at the same time. First, we replace $\mathcal{B}$ to be a uniform distribution an a finite set. The set is carefully designed so that it does not become too large but sufficiently dense so that the scale dependence can be ignored. Next, we replace both of the prior and the posterior of $\boldsymbol{\sigma}$ with a distribution which takes 1 on a real vector and 0 anywhere else. The replacement

[5] See Appendix D.1 in the supplementary material for the derivation.

corresponds to setting η' very small in their original prior and posterior. A more concrete explanation is provided in the proof of the next proposition, provided in Appendix D.3 in the supplementary material.

Proposition 4.2. *For any networks, any $\lambda \in \Lambda$ where $\Lambda \subset \mathbb{R}_{>0}$ is a finite set of real number independent of S, any positive number $\epsilon \in \mathbb{R}_{>0}$, any 0–1 bounded twice continuously differentiable loss function with respect to parameters, and any parameter assignment $\bar{\boldsymbol{\theta}}$ such that $\max_i \bar{\boldsymbol{\theta}}_i \leq b$, if the diagonal elements of the Hessian of the training loss $\mathcal{L}_{\mathcal{S}}$ is bounded by $M \in \mathbb{R}_{>0}$ and nonnegative, there exists a stochastic classifier $\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)$ such that with probability $1-\delta$ over the training set and $\mathbb{Q}$, the expected error $\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\right)$ is bounded by*[6]

$$\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1+\epsilon+\epsilon^2}{\sqrt{\lambda}}\mathrm{NS}_1 + \frac{1}{\lambda}\Xi_1 + \frac{\lambda}{2m} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right), \tag{20}$$

where

$$\Xi_1 = O(n) \cdot C_{\epsilon,\delta,|\Lambda|,M,b,\lambda}. \tag{21}$$

The assumptions that we can bound all elements of $\bar{\boldsymbol{\theta}}$ and the diagonal of the Hessian are reasonable because we typically train neural networks on computers with finite precision. Even though ReLU networks do not satisfy the smoothness assumption, which is necessary because we use the Hessian, the proposition at least addresses the scale dependence due to normalization layers. Section 6 discusses a possible method to address the smoothness and nonnegativity assumptions for the Hessian's diagonal elements. The method is also applicable to the above propositions. The ϵ that appears in the proposition comes from an approximation error introduced by the discretization trick. Setting $\epsilon = 1$ suggests that the parameter's existence does not worsen the order of the bound.

5. Finer-grained scale dependence

The generalization error bounds (15) and (20) have constant terms proportional to the number of parameters. This is because specifying scaling parameters per parameter is almost as difficult as directly specifying all parameters. Since deep learning methods typically use overparametrized models, the constant terms make the bounds vacuous. Prior parameter-wisely normalized sharpness metrics also suffer from the problem explictly or inexplicitly (Wang et al., 2018; Achille & Soatto, 2018). Fortunately, it turns out that we can control parameter scales by using fairly small number of scaling parameters. For example, we can guess that parameters belonging to the same weight matrix have similar scales. In this section, we discuss what controls the parameter scales in deep learning models. This discussion is critical for our improved PAC-Bayesian analysis in Sec. 6. It also reveals the scale dependence in existing matrix-norm-based generalization bounds[7].

Here, we analyze the scale dependences appearing in networks with the ReLU activation function. The similar dependences appear in networks with normalization layers, which we defer the explanation to Appendix C.1 in the supplementary material. We point out that the matrix-wise scale dependence introduced by Dinh et al. (2017) does not fully cover scale dependences[8]. To illustrate the hidden scale dependence, we consider a simple network with a single hidden layer and ReLU activation:

$$f_{\boldsymbol{\theta}}(z) = W^{(2)}(\mathrm{ReLU}(W^{(1)}(z))), \tag{22}$$

where weight matrices $W^{(1)}$ and $W^{(2)}$ are subsets of the parameters $\boldsymbol{\theta}$, and z is an input of the network $f_{\boldsymbol{\theta}}$. We can scale the i-th column of $W^{(2)}$ by $\alpha > 0$ and i-th row of $W^{(1)}$ by $1/\alpha$ without modifying the function that the network represents.[9] Since we are using the ReLU activation function, which has positive homogeneity, the transformation does not change the represented function. By the transformation, the diagonal elements of the Hessian corresponding to the i-th row of $W^{(1)}$ are scaled by α^2. The transformation has essentially the same effect as the one proposed by Dinh et al. (2017). The difference is that the above transformation runs node-wise instead of matrix-wise. In terms of weight matrices, the node-wise scale dependence can be translated into a row- and column-wise scale dependence.

6. Improved PAC-Bayesian analysis

In this section, we tighten our PAC-Bayesian bounds in Sec. 4 based on Sec. 5. In Sec. 5, we pointed out the row- and column-wise scale dependencies in modern network architectures. Thus, we at least need to absorb the row and column scales of weight matrices. It seems, however, that we do not need to control all parameter scales separately. Thus, we modify the decomposition per weight matrix in Sec. 4 in the following way.

$$W^{(l)} = \eta \mathrm{Diag}\left(\mathrm{e}^{\left[\boldsymbol{\gamma}^{(l)}\right]}\right) V^{(l)} \mathrm{Diag}\left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}\right]}\right), \eta \in \mathbb{R}, \tag{23}$$

where $\mathrm{Diag}\left(\mathrm{e}^{\left[\boldsymbol{\gamma}^{(l)}\right]}\right)$ and $\mathrm{Diag}\left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}\right]}\right)$ are diagonal matrices whose diagonal elements are $\mathrm{e}^{\left[\boldsymbol{\gamma}^{(l)}\right]}$ and $\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}\right]}$, re-

[6] The exact form of Ξ_1 can be found in the Appendix D.3 in the supplementary material.

[7] See Appendix C.2 in the supplementary material for the explanation.

[8] The same scale dependences in ReLU networks were introduced by Neyshabur et al. (2015).

[9] Running examples of the transformation can be found in Appendix B.1 in the supplementary material.

spectively. The codomains of random variables $\boldsymbol{\gamma}^{(l)}$, $\boldsymbol{\gamma}'^{(l)}$, and $V^{(l)}$ are $\mathbb{R}^{h_1^{(l)}}$, $\mathbb{R}^{h_2^{(l)}}$, and $\mathbb{R}^{h_1^{(l)} \times h_2^{(l)}}$, respectively. Similar decomposition with (23) has been explored for the unit-invariant SVD (Uhlmann, 2018). In the method, weight matrices are normalized using the solution of *Program II* (Rothblum & Zenios, 1992). Even though the same method is available for removing the scale dependence, we consider the Hessian jointly and use different scaling control that has some optimality from the PAC-Bayesian perspective. We first define some notations for convenience. Let $\bar{U}^{(l)}$ be a matrix defined as $\bar{U}^{(l)} = \bar{W}^{(l)} \odot \bar{W}^{(l)}$, where $\bar{W}^{(l)}$ is a subset of a parameter assignment $\bar{\boldsymbol{\theta}}$. Let $\bar{H}^{(l)}$ be a matrix such that

$$\bar{H}^{(l)}_{i,j} = \left(\left. \frac{\partial^2 \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})}{\partial {W^{(l)}}_{i,j} \partial {W^{(l)}}_{i,j}} \right|_{\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}} \right). \tag{24}$$

By extending Prop. 4.2 to decomposition (23), we have the following proposition.

Proposition 6.1. *For any networks with d weight matrices, any $\lambda \in \Lambda$ where $\Lambda \subset \mathbb{R}_{>0}$ is a finite set of real number independent of S, any positive number $\epsilon \in \mathbb{R}_{>0}$, any positive number $R \in \mathbb{R}_{>0}$, any 0–1 bounded twice continuously differentiable loss function with respect to parameters, and a parameter assignment $\bar{\boldsymbol{\theta}}$ such that $\max_i \bar{\boldsymbol{\theta}}_i \leq b$, if the diagonal elements of the Hessian of the training loss $\mathcal{L}_{\mathcal{S}}$ are bounded by $M \in \mathbb{R}_{>0}$ and nonnegative, then there exists a stochastic classifier $\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)$ such that with probability $1-\delta$ over the training set and $\mathbb{Q}$, the expected error $\mathcal{L}_{\mathcal{D}}(\mathbb{Q})$ is bounded by*[10]

$$\mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right) + \frac{(1+\epsilon)^2}{\sqrt{\lambda}} \mathrm{NS}_2(R) + \frac{1}{\lambda}\Xi_2 + \frac{\lambda}{2m} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right), \tag{25}$$

where $\mathrm{NS}_2(R) =$

$$\sum_{l=1}^{d} \min_{\bar{\boldsymbol{\gamma}}^{(l)} \in [-R,R]^{h_1^{(l)}}} \sqrt{\mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{H}^{(l)} \left(\bar{U}^{(l)}\right)^\top \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]}}, \tag{26}$$

$$\Xi_2 = \tilde{O}(hd) \cdot C_{\epsilon,\delta,M,b,R,|\Lambda|}, \tag{27}$$

and $h = \frac{1}{d} \max\left(\sum_{l=1}^{d} h_1^{(l)}, \sum_{l=1}^{d} h_2^{(l)}\right)$.

The constant term, Ξ_2, scales by the number of nodes, instead of parameters. The reduction of the constant term from Prop. 4.2 makes the bound meaningful in practical settings. For example, ResNet50 has tens of millions of parameters, while it only has tens of thousands of nodes. In classification on ImageNet, which has millions of images, this reduction is critical.

Note that the proof of Prop. 6.1 is constructive. In the construction, the posterior $\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)$ is a Gaussian centered at $\bar{\boldsymbol{\theta}}$ with a diagonal covariance matrix Σ such that $\|\Sigma\|_{\mathrm{F}} = \boldsymbol{O}(\lambda^{-\frac{1}{2}})$. Since the $\mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right)$ term comes from a second-order approximation of the loss function, as λ increases, which means a training set size m increases, the second-order approximation will hold better, and the $\mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right)$ term will decrease.

For the sake of theoretical completeness, we address the $\mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right)$ term, smoothness assumption, and nonnegative assumption. In Props. 4.2 and 6.1, the $\mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right)$ term was introduced as a term satisfying the following equation.

$$\begin{aligned} &\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \\ &= \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \\ &\quad + \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim \mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)} \left[\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right)^\top \left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right) \left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right)\right]. \end{aligned} \tag{28}$$

As long as the equation is satisfied, we can use any vectors which mimick the Hessian's diagonal elements. For example, we can use the following as an estimation of the Hessian diagonal.

$$\begin{aligned} \boldsymbol{f} = \mathbb{E}_{\boldsymbol{s} \in \{-1,1\}^n} \Big[\mathrm{Div}\big[&\boldsymbol{s} \odot \left(g_{\mathcal{S},\bar{\boldsymbol{\theta}},\epsilon}(r, \boldsymbol{s}) - g_{\mathcal{S},\bar{\boldsymbol{\theta}},\epsilon}(r, -\boldsymbol{s}) \right), \\ &2r(\mathrm{Abs}[\boldsymbol{\theta}] + \epsilon \mathbf{1}) \big] \Big], \end{aligned} \tag{29}$$

where

$$g_{\mathcal{S},\bar{\boldsymbol{\theta}},\epsilon}(r, \boldsymbol{s}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}} + r(\mathrm{Abs}(\bar{\boldsymbol{\theta}})\boldsymbol{s} + \epsilon \mathbf{1})\right), \tag{30}$$

$\mathrm{Div}[\cdot, \cdot]$ is an elementwise division, and $\mathrm{Abs}[\cdot, \cdot]$ is an elementwise absolute. We added a parameter ϵ for numerical stability. When a network is smooth, $\boldsymbol{f}$ is an approximation of the diagonal elements of the Hessian. Note, the estimation is scale-invariant when $\epsilon = 0$. The estimation is defined even when networks are not smooth. Furthermore, we can carry out the following modification to $\boldsymbol{f}$ to enforce the nonnegative condition.

$$\hat{\boldsymbol{f}} = \mathrm{Max}\left[\boldsymbol{f}, \mathbf{0}\right], \tag{31}$$

where $\mathrm{Max}[\cdot, \cdot]$ is an elementwise maximum. Let $\bar{F}^{(l)}$ be a matrix such that we substitute the diagonal elements of the Hessian by $\hat{\boldsymbol{f}}$ in the definition of $\bar{H}^{(l)}$ (24). Using $\bar{F}^{(l)}$ instead of $\bar{H}^{(l)}$, we have the following theorem.

Theorem 6.1. *For any networks with d weight matrices, for any $\lambda \in \Lambda$ where $\Lambda \subset \mathbb{R}_{>0}$ is a finite set of real number independent of S, a positive number $\epsilon \in (0, 0.1]$, any positive number $R \in \mathbb{R}_{>0}$, any 0–1 bounded loss function, and a parameter assignment $\bar{\boldsymbol{\theta}}$ such that $\max_i \bar{\boldsymbol{\theta}}_i \leq b$, if all elements of $\hat{\boldsymbol{f}}$ are bounded by $M \in \mathbb{R}_{>0}$, then there exists a Gaussian distribution $\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)$ with a mean $\bar{\boldsymbol{\theta}}$ and a covariance matrix Σ such that $\|\Sigma\|_{\mathrm{F}} = \boldsymbol{O}(\lambda^{-1/2})$ in terms of*

[10] The exact form of Ξ_2 can be found in the Appendix D.4 in the supplementary material.

λ, and with probability $1-\delta$ over the training set $\mathcal{S}$ and $\mathbb{Q}$, the expected error $\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\right)$ is bounded by[11]

$$\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta}\mid\bar{\boldsymbol{\theta}}\right)\right)+\frac{1}{\sqrt{\lambda}}\mathrm{NS}_3(R)+\frac{1}{\lambda}\Xi_3+\frac{\lambda}{2m}. \tag{32}$$

where $\mathrm{NS}_2(R) =$

$$\sum_{l=1}^{d}\min_{\bar{\boldsymbol{\gamma}}^{(l)}\in[-R,R]^{h_1^{(l)}}}\sqrt{\mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]}{}^{\top}\bar{F}^{(l)}\left(\bar{U}^{(l)}\right)^{\top}\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]}} \tag{33}$$

and

$$\Xi_3=\tilde{O}(hd)\cdot C_{\epsilon,\delta,M,b,R,|\Lambda|}. \tag{34}$$

Note that we can statistically bound the term $\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta}\mid\bar{\boldsymbol{\theta}}\right)\right)$ using the Monte-Carlo algorithm and Hoeffding's inequality. In Theorem 6.1, we use the second-order approximation of the loss to estimate the best choice of posterior variance to minimize both terms (A) and (B) in Eq. (4).

7. Normalized sharpness definition and calculation

In this section, based on the analysis in Sec. 6, we define normalized sharpness, a sharpness and generalization metric invariant with respect to node-wise rescaling. We also describe a practical calculation technique for the metric.

We define normalized sharpness as

$$\mathrm{NS}=\sum_{l=1}^{d}\inf_{\bar{\boldsymbol{\gamma}}^{(l)}\in\mathbb{R}^{h_1^{(l)}}}\sqrt{\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]}{}^{\top}\bar{F}^{(l)}\left(\bar{U}^{(l)}\right)^{\top}\mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]}}, \tag{35}$$

where the notations $\bar{U}^{(l)}$ and $\bar{F}^{(l)}$ were introduced in Sec. 6. Normalized sharpness appeared in Prop. 6.1 and Theorem 6.1. The reason for the scale-invariance of normalized sharpness is explained in Appendix J in the supplementary material.

In convolutional layers, since the same filter has the same scale, we can tie the scaling parameter per channel. We defer a more detailed description of normalized sharpness for convolutional layers to Appendix F in the supplementary material. Fortunately, the optimization problem (35) is convex with respect to $\bar{\boldsymbol{\gamma}}^{(l)}$.[12] Thus, we can estimate a near-optimal solution to the optimization problem by gradient descent. It is straightforward to see that the convexity also holds with convolutional layers. When calculating the normalized sharpness, we need to ensure that the choice of surrogate loss function does not introduce other scale dependences. We discuss this choice in Appendix G in the supplementary material

8. Numerical evaluation

We numerically evaluated the effectiveness of normalized sharpness (35). We mainly validated the following two points.

- Unnormalized sharpness metrics can fail to predict generalization performance.
- Normalized sharpness (35) better predicts generalization than unnormalized metrics.

For these purposes, we checked the metrics' ability to distinguish models trained on random labels (Sec. 8.1) and predict the generalization performance of models trained with different hyperparameters (Sec. 8.2). As existing current sharpness metrics, we used the trace of the Hessian without normalization (6) and the sum of the squared Frobenius norm of the weight matrices (Neyshabur et al., 2017). We also investigated the empirical dependence of normalized sharpness on the width and depth of networks (Sec. 8.3). In all experiments, we trained multilayer perceptrons with three hidden layers, LeNet (Lecun et al., 1998), and Wide ResNet (Zagoruyko & Komodakis, 2016) with 16 layers and width factor 4 on MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009). More detailed experimental setups are described in Appendix M in the supplementary material.

8.1. Random labels

We first investigated whether normalized sharpness (35) and current unnormalized sharpness metrics can distinguish models trained on random labels. In this experiment, sharper minima are expected to indicate larger generalization gaps.

Results: Figure 1 shows plots of the mean sharpness metrics for models trained on datasets with different random label ratios. The results show that networks trained on random labels had larger normalized sharpness to fit the random labels. Thus, we can say that normalized sharpness provides a fairly good hierarchy in the hypothesis class. Even though sharpness without normalization and the squared Frobenius norm of weight matrices could also distinguish models trained on random labels to some extent, the signal was much weaker than that of normalized sharpness. Note that normalized sharpness has an advantage in that it does not require a biplot for both the unnormalized sharpness and the squared Frobenius norm.

[11] The exact form of Ξ_3 can be found in the Appendix D.5 in the supplementary material.

[12] See Appendix E in the supplementary material.

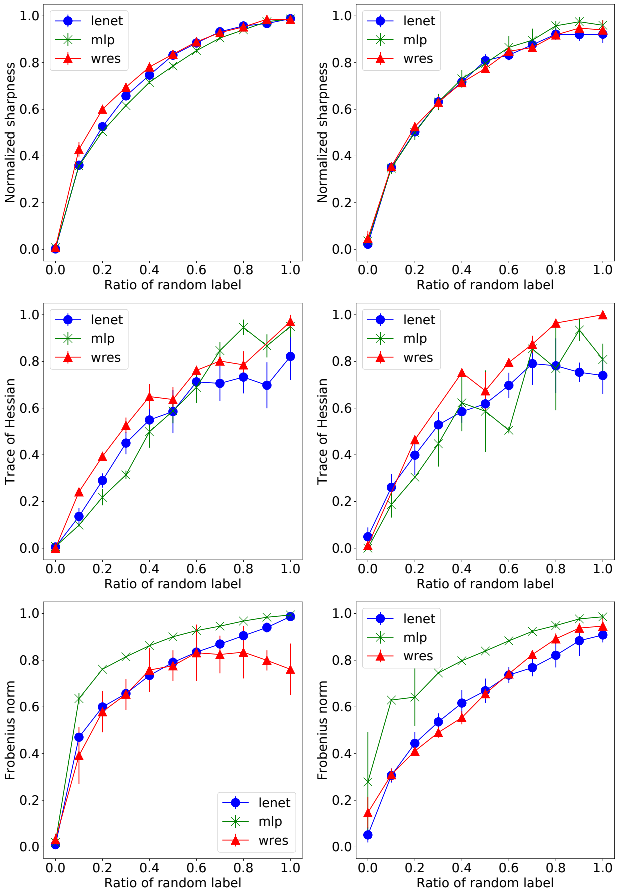

Figure 1. The figure shows the values of normalized sharpness (35), trace of the Hessian (6), and the sum of the squared Frobenius norms of weight matrices of models trained on datasets with different random label ratios. Each dot represents the mean of the generalization metric among trained networks and the error bars represent the standard deviation. The left column is the results on MNIST and the right is on CIFAR10. All generalization metrics were rescaled to $[0, 1]$ by their maximum and minimum per network architecture.

8.2. Different hyperparameters

We tested whether normalized sharpness and the existing unnormalized sharpness metrics can predict the generalization performance of models trained with different hyperparameters. We then checked the correlations between the generalization metrics and their generalization gaps, defined by $(\text{test misclassification ratio}) - (\text{train misclassification ratio})$. Models are trained with different strengths of l2-regularization, weight decay (Loshchilov & Hutter, 2019), and dropout. This is an adversarial setting for existing unnormalized sharpness metrics because both regularizations on weights and dropout directly affects the balance between parameter scales and flatness of minima.

Table 1. Correlation coefficients between generalization gap and the generalization metrics for models trained on MNIST.

	MLP	Lenet	WResNet
Normalized sharpness	**0.73**	**0.73**	**0.59**
Trace of Hessian	-0.62	-0.79	-0.59
Frobenius norm	0.58	0.58	0.49

Table 2. Correlation coefficients between generalization gap and the generalization metrics for models trained on CIFAR10.

	MLP	Lenet	WResNet
Normalized sharpness	**0.92**	**0.98**	**0.92**
Trace of Hessian	-0.42	-0.51	-0.53
Frobenius norm	0.72	0.76	0.43

Results: We summarized the correlation coefficients between curvature metrics and the accuracy gaps in Table 1 and Table 2. Scatter plots can be found in Appendix N.1 in the supplementary material with more detailed results. On CIFAR10, especially for LeNet and Wide ResNet, we observed almost linear correlations between normalized sharpness and accuracy gap. Thus, we can confirm the usefulness of our generalization error bounds and normalized sharpness. On MNIST, even though there were weak correlations between normalized sharpness and the accuracy gap, the correlation was weaker than the results on CIFAR10. A possible explanation of this phenomenon is that the scale of the accuracy gap was too small on MNIST, which was at most 0.02. Since we could not create models with various accuracy gaps by merely changing the regularization parameters on MNIST, the effect of noise would have become larger. In all settings, both the trace of the Hessian and the squared Frobenius norm of the weight matrices had a weaker ability to predict the generalization. We can confirm that normalized sharpness had consistently stronger correlations with generalization. Notably, we observed a negative correlation with the trace of the Hessian and positive correlation with the squared Frobenius norm. If we use the tuples of the two as a generalization metric (Neyshabur et al., 2017), we cannot determine which models will generalize better. We explain the negative correlations as follows. In our expeirments, we used three regularizers: l2-regularization, weight-decay, and dropout. When we use l2-regularization or weight-decay, stronger regularization decreases the generalization error and parameter scales. Since there are certain trade-offs between the sharpness and parameter scales, as shown in Eq.(4), the stronger regularization makes the Hessian larger.

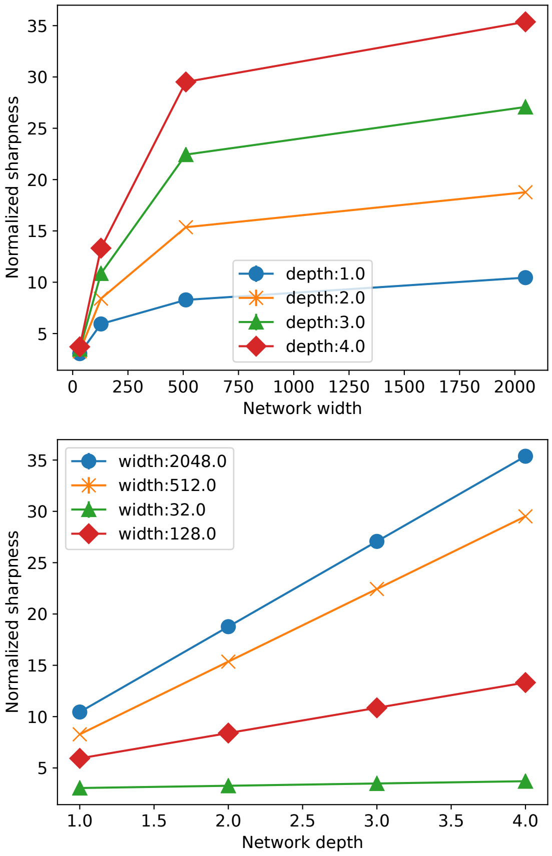

Figure 2. Plot of the dependence of normalized sharpness on width and depth. The top figure plots the dependence on width of the normalized shapness of multilayer-perceptron with different width. The bottom figure plots those on depth. The normalized sharpness showed almost linear dependence on network depth and sublinear dependence on network width.

8.3. Dependence on width and depth

To verify that the reduction of the constant term presented in Sec. 6 improves the overall bounds, we empirically investigated the dependency of normalized sharpness on network width and depth. We trained multilayer perceptron with depth-$\{1, 2, 3, 4\}$ and width-$\{32, 128, 512, 2048\}$ on MNIST three times for each and calculated normalized sharpness for every trained model. Figure 2 shows the results. It indicates that the normalized sharpness has an almost linear dependence on the network depth and sub-linear dependence on the network width, at least under this setting. This suggests that the constant term reduction presented in Sec. 6 contributes to tightening the overall generalization bounds.

9. Comparison with other generalization bounds:

We compare the tightness of our bounds and a prior PAC-Bayesian approach[13] (Neyshabur et al., 2018). First, we remark that Neyshabur et al. (2018) relies on a margin parameter. The margin parameter controls the term (A) in Eq. (4) to some extent. Importantly, the margin loss they rely on does not decrease when the sample size increases. On the other hand the term (A) is solely controlled by the posterior choice in our bounds. As a result, the term decrease as the sample size increases. However, the difference makes the rate of convergence different and their direct comparison harder. The convergence rate is further discussed in Appendix I in the supplementary material. Nevertheless, we can still compare the effect of our constant term, which is $O(hd)$, with existing bounds. The main bound in Neyshabur et al. (2018) scales at least $O(d^2h)$. Compared to that, our constant term will not be critical to the tightness of our bound.

An $O(hd)$ generalization bound corresponds to $O(1/h)$ compression per layer. This is fairly competitive with the reported result by Arora et al. (2018). Even though our work is not directly comparable to Arora et al. (2018), this suggests that our bounds are comparably tight compared to those derived by Arora et al. (2018).

Observations in Sec. 8.3 suggest an advantage of normalized sharpness over the Fisher-Rao norm. Since the Fisher-Rao norm scales $O(d^2)$ with respect to depth, when the represented function is identical (Liang et al., 2019), normalized sharpness might be more robust against architecture changes, especially concerning depth of networks.

10. Conclusion

We have formally connected normalized loss curvatures with generalization through PAC-Bayesian analysis. The analysis bridged the known gap between theoretical understandings and empirical connections between normalized loss-landscape and generalization. The proof consists of two steps: scale-direction decompositions of parameters and discretization trick. In the analysis, we found that using a smaller number of scaling parameters is critical for meaningful generalization bounds, at least within our framework. Applying this discovery, we proposed *normalized sharpness* as a novel generalization metric. Experimental results suggest that this metric is more powerful than unnormalized loss sharpness metrics as a measure of generalization.

[13] Remark: Our primal goal is not providing the state-of-the-art tightest bound, but connecting scale-invariant flatness metric and generalization.

ACKNOWLEDGEMENTS

YT was supported by Toyota/Dwango AI scholarship. IS was supported by KAKENHI 20H04239. MS was supported by the International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study.

References

Achille, A. and Soatto, S. Emergence of Invariance and Disentanglement in Deep Representations. *Journal of Machine Learning Research*, 19, 2018.

Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger Generalization Bounds for Deep Nets via a Compression Approach. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 254–263, 2018.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer Normalization. *ArXiv e-prints*, abs/1607.06450, 2016.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30*, pp. 6240–6249. 2017.

Catoni, O. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-SGD: Biasing Gradient Descent Into Wide Valleys. In *International Conference on Learning Representations*, 2017.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp Minima Can Generalize For Deep Nets. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1019–1028, 2017.

Dziugaite, G. K. and Roy, D. M. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Dziugaite, G. K. and Roy, D. M. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems 31*, pp. 8430–8441. 2018.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. PAC-Bayesian Theory Meets Bayesian Inference. In *Advances in Neural Information Processing Systems 29*, pp. 1884–1892. 2016.

Guedj, B. A Primer on PAC-Bayesian Learning. *ArXiv e-prints*, abs/1901.05353, 2019.

Hinton, G. E. and van Camp, D. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pp. 5–13, 1993.

Hochreiter, S. and Schmidhuber, J. Flat Minima. *Neural Computation*, 9(1):1–42, 1997.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems 30*, pp. 1731–1741. 2017.

Honkela, A. and Valpola, H. Variational Learning and Bits-back Coding: An Information-theoretic View to Bayesian Learning. *Trans. Neur. Netw.*, 15(4):800–810, 2004. ISSN 1045-9227.

Ioffe, S. and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

Jiang, Y., Neyshabur, B., Krishnan, D., Mobahi, H., and Bengio, S. Fantastic Generalization Measures and Where to Find Them. In *International Conference on Learning Representations*, 2020.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*, 2017.

Kingma, D. P., Salimans, T., and Welling, M. Variational Dropout and the Local Reparameterization Trick. In *Advances in Neural Information Processing Systems 28*, pp. 2575–2583. 2015.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. 2012.

Langford, J. and Caruana, R. (Not) Bounding the True Error. In *Advances in Neural Information Processing Systems 14*, pp. 809–816. 2002.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

LeCun, Y., Cortes, C., and Burges, C. J. C. The MNIST Database of Handwritten Digits. 1998.

Letarte, G., Germain, P., Guedj, B., and Laviolette, F. Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks. In *Advances in Neural Information Processing Systems 32*, pp. 6869–6879. 2019.

Li, H., Xu, Z., Taylor, G., and Goldstein, T. Visualizing the Loss Landscape of Neural Nets. In *Advances in Neural Information Processing Systems 31*, pp. 6391–6401. 2018.

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. Fisher-Rao Metric, Geometry, and Complexity of Neural Networks. In *Proceedings of Machine Learning Research*, volume 89, pp. 888–896, 2019.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Nagarajan, V. and Kolter, Z. Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2019.

Neklyudov, K., Molchanov, D., Ashukha, A., and Vetrov, D. P. Structured Bayesian Pruning via Log-Normal Multiplicative Noise. In *Advances in Neural Information Processing Systems 30*, pp. 6775–6784. 2017.

Neyshabur, B., Salakhutdinov, R. R., and Srebro, N. Path-SGD: Path-Normalized Optimization in Deep Neural Networks. In *Advances in Neural Information Processing Systems 28*, pp. 2422–2430. 2015.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems 30*, pp. 5947–5956. 2017.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. In *International Conference on Learning Representations*, 2018.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. PAC-Bayes Bounds with Data Dependent Priors. *Journal of Machine Learning Research*, 13:3507–3531, 2012.

Rissanen, J. Stochastic Complexity and Modeling. *Ann. Statist.*, 14(3):1080–1100, 09 1986.

Rothblum, U. G. and Zenios, S. A. Scalings of Matrices Satisfying Line-Product Constraints and Generalizations. *Linear Algebra and Its Applications*, pp. 159–175, 1992.

Salimans, T. and Kingma, D. P. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Advances in Neural Information Processing Systems 29*, pp. 901–909. 2016.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Uhlmann, J. A Generalized Matrix Inverse that is Consistent with Respect to Diagonal Transformations. *SIAM Journal on Matrix Analysis and Applications*, pp. 781–800, 2018.

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., and Hassabis, D. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3918–3926, 2018.

Wang, H., Shirish Keskar, N., Xiong, C., and Socher, R. Identifying Generalization Properties in Neural Networks. *ArXiv e-prints*, abs/1809.07402, 2018.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5987–5995, 2017.

Yao, Z., Gholami, A., Lei, Q., Keutzer, K., and Mahoney, M. W. Hessian-based Analysis of Large Batch Training and Robustness to Adversaries. In *Advances in Neural Information Processing Systems 31*, pp. 4954–4964. 2018.

Zagoruyko, S. and Komodakis, N. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference*, pp. 87.1–87.12, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding Deep Learning Requires Rethinking Generalization. In *International Conference on Learning Representations*, 2017.

Zhang, G., Wang, C., Xu, B., and Grosse, R. Three Mechanisms of Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Zhou, W., Veitch, V., Austern, M., Adams, R. P., and Orbanz, P. Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach. In *International Conference on Learning Representations*, 2019.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018.

A. Notation

Our notation is summarized in Table 3.

Table 3. Notation table.

$\mathbb{P}$: prior distribution

$\mathbb{Q}$: posterior distribution

$\mathbb{P}\left(\boldsymbol{\theta}\right)$: prior distribution of random variable $\boldsymbol{\theta}$

$\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)$: prior distribution of random variable $\boldsymbol{\theta}$, $\bar{\boldsymbol{\theta}}$ is a hyperparameter of $\mathbb{Q}$

$\mathcal{D}$: underlying (true) data distribution

$\mathcal{S}$: training set with size m, *i.i.d.* sample from $\mathcal{D}$

z_i: i-th sample in $\mathcal{S}$, $z_i \in \mathcal{S}$

m: number of data points in a training set

K: number of classes

d: number of layers in NN (depth)

$W^{(l)}$: l-th weight matrix

$h_1^{(l)}$: input dimension of the l-th weight matrix, random variable

$h_2^{(l)}$: output dimension of the l-th weight matrix

h: average width of a network, defined as $\mathrm{Max}\left(\sum_l h_1^{(l)}, \sum_l h_2^{(l)}\right)/d$

$\boldsymbol{\theta}$: parameters of a network, a random variable

$\bar{U}^{(l)}$: a real matrix whose i,j-th element is a squared i,j-th element of the l-th weight matrix

$\bar{H}^{(l)}$: a real matrix whose i,j-th element is $\partial \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)/\partial W^{(l)}{}_{i,j} \partial W^{(l)}{}_{i,j}$ at $\bar{\boldsymbol{\theta}}$

$\bar{\boldsymbol{\theta}}$: parameters of a network, a real vector

$\mathcal{L}_{\mathcal{D}}\left(\bar{\boldsymbol{\theta}}\right)$: expected loss over a distribution $\mathcal{D}$ at $\bar{\boldsymbol{\theta}}$

$\mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right)$: expected loss over a training set $\mathcal{S}$ at $\bar{\boldsymbol{\theta}}$

$\mathcal{L}_{z}\left(\bar{\boldsymbol{\theta}}\right)$: loss on a data point z at $\bar{\boldsymbol{\theta}}$

$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\right)$: $\mathbb{E}_{\bar{\boldsymbol{\theta}}' \sim \mathbb{Q}}\left[\mathcal{L}_{\mathcal{D}}\left(\bar{\boldsymbol{\theta}}'\right)\right]$

$\mathrm{KL}[\mathbb{Q} \parallel \mathbb{P}]$: KL divergence

$\nabla_{\boldsymbol{\theta}}$: derivative concerning parameter $\boldsymbol{\theta}$

$\nabla^2_{\boldsymbol{\theta}}$: Hessian concerning parameter $\boldsymbol{\theta}$

$\left\|\bar{W}\right\|_{\mathrm{F}}$: Frobenius norm of a matrix $\bar{W}$

$\left\|\bar{W}\right\|_{2}$: Spectral norm of a matrix $\bar{W}$

$f(z)$: output of a network f at a data point z

$\bar{\boldsymbol{x}}_i$: the i-th element of a vector $\bar{\boldsymbol{x}}$

$\bar{A}_{i,j}$: the (i,j)-th element of a matrix $\bar{A}$

$\bar{A}_{i,:}$: the i-th row of a matrix $\bar{A}$

$\bar{A}_{:,j}$: the j-th column of a matrix $\bar{A}$

y_z: label of data point z

I: identity matrix

$\mathrm{Diag}\left(\bar{\boldsymbol{x}}\right)$: a diagonal matrix which a vector consists of its diagonal elements is a vector $\bar{\boldsymbol{x}}$

B. Running examples

B.1. Row and column scaling

We show running examples of the transformation proposed in Sec. 5. We consider the following network.

$$f(X) = W^{(2)}(\mathrm{ReLU}(W^{(1)}(X))), \tag{36}$$

$$W^{(1)} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \tag{37}$$

$$W^{(2)} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}. \tag{38}$$

Now, matrix norms are as follows.

$$\left\| W^{(1)} \right\|_{\mathrm{F}} = \sqrt{30}, \tag{39}$$

$$\left\| W^{(1)} \right\|_{2} \approx 5.48, \tag{40}$$

$$\left\| W^{(2)} \right\|_{\mathrm{F}} = \sqrt{174}, \tag{41}$$

$$\left\| W^{(2)} \right\|_{2} \approx 13.19. \tag{42}$$

We apply the transformation to the first row of $W^{(1)}$ and the first column of $W^{(2)}$ with $\alpha = 10$. Then, the parameters change as follows.

$$W^{(1)} = \begin{pmatrix} 0.1 & 0.2 \\ 3 & 4 \end{pmatrix}, \tag{43}$$

$$W^{(2)} = \begin{pmatrix} 50 & 6 \\ 70 & 8 \end{pmatrix}. \tag{44}$$

Next, we apply the transformation to the second row of $W^{(1)}$ and the second column of $W^{(2)}$ with $\alpha = 0.1$. The parameters change as follows.

$$W^{(1)} = \begin{pmatrix} 0.1 & 0.2 \\ 30 & 40 \end{pmatrix}, \tag{45}$$

$$W^{(2)} = \begin{pmatrix} 50 & 0.6 \\ 70 & 0.8 \end{pmatrix}. \tag{46}$$

Now, matrix norms have changed as follows.

$$\left\| W^{(1)} \right\|_{\mathrm{F}} = \sqrt{2500.05}, \tag{47}$$

$$\left\| W^{(1)} \right\|_{2} \approx 50.00, \tag{48}$$

$$\left\| W^{(2)} \right\|_{\mathrm{F}} = \sqrt{6101.13}, \tag{49}$$

$$\left\| W^{(2)} \right\|_{2} \approx 78.11. \tag{50}$$

Using the same method, we can make matrix norms of both $W^{(1)}$ and $W^{(2)}$ arbitrarily large.

C. Scale Dependences

In this section, we first review scale dependencies appear in normalization layers. Next, we point out how the scale dependencies can affect to known matrix-norm based generalization error bounds.

C.1. Scale dependence in normalization layers

Normalization layers such as layer-normalization (Ba et al., 2016), batch-normalization, and weight-normalization are critical components of modern network architectures. These normalization layers also introduce scale dependences to flatness metrics (Dinh et al., 2017). Let us take weight-normalization as an example. We consider the following simple network.

$$f_{\boldsymbol{\theta}}(x) = \mathrm{WeightNorm}(W)(x). \tag{51}$$

By the definition of weight-normalization, scaling the i-th row of W by $\alpha > 0$ does not change the function represented by the network. However, its gradients are scaled by α^{-1} (Salimans & Kingma, 2016), and the corresponding diagonal elements of the Hessian are scaled by α^{-2}. Again, we have the node-wise scale dependence.

C.2. Scale dependence in known generalization metrics

This subsection is not critical for this paper, but the row- and column-wise scale dependence reveals scale dependence in known capacity controls with matrix-norms, which will take other interests. We reconsider a network (22). Assume that $\bar{W}^{(1)}$ has at least two non-zero rows, and $\bar{W}^{(2)}$ has at least two non-zero columns. Using the row- and column-wise rescaling, we can make both $\bar{W}^{(1)}$, and $\bar{W}^{(2)}$ have at least one arbitrarily large element. In other words, both weight matrices have arbitrarily large spectral norms and Frobenius norms. Thus, matrix-norm based capacity controls (Bartlett et al., 2017; Neyshabur et al., 2018) also suffer from the same scale dependencies as flatness metrics.

D. Proofs

D.1. Derivation of (A) and (B) term

This section derives the (A) and (B) term using a Taylor expansion in the new parameter space.

D.1.1. (A) TERM

$$\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) - \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) \tag{52}$$

$$= \mathop{\mathbb{E}}_{\substack{\boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, I) \\ \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, \eta' I)}} \left[\mathcal{L}_{\mathcal{S}}\left(\eta\left(\bar{\boldsymbol{\mu}} + \boldsymbol{\epsilon}_0\right) \odot \mathrm{e}^{[\bar{\boldsymbol{\sigma}} + \boldsymbol{\epsilon}_1]}\right) - \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right)\right] \tag{53}$$

$$\tag{54}$$

$$= \frac{1}{2} \sum_i \left(\left.\frac{\partial^2 \mathcal{L}_{\mathcal{S}}}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}}\right|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} + \frac{1}{2} \sum_i \left(\left.\frac{\partial^2 \mathcal{L}_{\mathcal{S}}}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}}\right|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \left(\eta'\right)^2 + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \tag{55}$$

$$= \frac{1}{2} \eta^2 \sum_i \left(\left.\frac{\partial^2 \mathcal{L}_{\mathcal{S}}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}}\right|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + \frac{1}{2} \left(\eta'\right)^2 \sum_i \left(\left.\frac{\partial^2 \mathcal{L}_{\mathcal{S}}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}}\right|_{\bar{\boldsymbol{\theta}}}\right)_{i,i} \bar{\boldsymbol{\theta}}_i^2 + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right). \tag{56}$$

D.1.2. (B) TERM

$$\frac{1}{\lambda} \mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right) \parallel \mathbb{P}\left(\boldsymbol{\theta}\right)] \leq \frac{1}{\lambda}\left(\mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\mu} \mid \bar{\boldsymbol{\mu}}\right) \parallel \mathbb{P}\left(\boldsymbol{\mu}\right)] + \mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\sigma} \mid \bar{\boldsymbol{\sigma}}\right) \parallel \mathbb{P}\left(\boldsymbol{\sigma}\right)]\right) \tag{57}$$

$$= \frac{1}{\lambda}\left(\frac{\|\bar{\boldsymbol{\mu}}\|_2^2}{2} + \frac{\|\bar{\boldsymbol{\sigma}} - \boldsymbol{\beta}\|_2^2}{2\left(\eta'\right)^2}\right) \tag{58}$$

$$= \frac{1}{\lambda}\left(\frac{\|\bar{\boldsymbol{\theta}} \odot \mathrm{e}^{[-2\boldsymbol{\sigma}]}\|_2^2}{2\eta^2} + \frac{\|\bar{\boldsymbol{\sigma}} - \boldsymbol{\beta}\|_2^2}{2\left(\eta'\right)^2}\right) \tag{59}$$

D.2. Proposition 4.1

Let $H = \{x^{-\frac{1}{4}} | x \in \Lambda\}$ and $H' = \{x^{-\frac{1}{4}} y^{-\frac{1}{4}} | x \in \Lambda', y \in \Lambda\}$. Taking union bound for η over H, η' over H', and $\boldsymbol{\beta}$ over $\mathcal{B}^n$, with probability at least $1 - \delta$,

$$
\begin{aligned}
&\forall \eta \in H, \forall \eta' \in H', \forall \boldsymbol{\beta} \in \mathcal{B}^n, \forall \boldsymbol{\sigma} \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}_{>0}, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1}{2}\eta^2 \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]^\top} (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]} + \frac{1}{2}{\eta'}^2 \bar{\boldsymbol{\theta}}^\top (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \bar{\boldsymbol{\theta}} \\
&\qquad + \frac{1}{2\lambda}\left(\frac{1}{\eta^2}\left\|\bar{\boldsymbol{\theta}} \odot \mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}]}\right\|_2^2 + \frac{1}{{\eta'}^2}\left\|\bar{\boldsymbol{\sigma}} - \boldsymbol{\beta}\right\|_2^2\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \\
&\qquad + \frac{1}{\lambda}\left(\ln|H| + \ln|H'| + n \ln|\mathcal{B}| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m}.
\end{aligned} \tag{60}
$$

Limiting η' to $\eta\Lambda'$, we have

$$
\begin{aligned}
&\forall \eta \in H, \forall \eta' \in \eta\Lambda', \forall \boldsymbol{\beta} \in \mathcal{B}^n, \forall \boldsymbol{\sigma} \in \mathbb{R}^n, \forall \lambda \in \Lambda, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1}{2}\eta^2 \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]^\top} (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]} + \frac{1}{2}{\eta'}^2 \bar{\boldsymbol{\theta}}^\top (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \bar{\boldsymbol{\theta}} \\
&\qquad + \frac{1}{2\lambda}\left(\frac{1}{\eta^2}\left\|\bar{\boldsymbol{\theta}} \odot \mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}]}\right\|_2^2 + \frac{1}{{\eta'}^2}\left\|\bar{\boldsymbol{\sigma}} - \boldsymbol{\beta}\right\|_2^2\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \\
&\qquad + \frac{1}{\lambda}\left(\ln|H| + \ln|H'| + n \ln|\mathcal{B}| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m}.
\end{aligned} \tag{61}
$$

By setting $\eta = \lambda^{-\frac{1}{4}} \in H$ and using $|H| = |\Lambda|, |H'| = |\Lambda||\Lambda'|$,

$$
\begin{aligned}
&\forall \lambda' \in \Lambda', \forall \boldsymbol{\beta} \in \mathcal{B}^n, \bar{\boldsymbol{\sigma}} \in \mathbb{R}^n, \forall \lambda \in \Lambda, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\lambda'}{2\sqrt{\lambda}} \bar{\boldsymbol{\theta}}^\top (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \bar{\boldsymbol{\theta}} + \frac{1}{2\sqrt{\lambda}} \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]^\top} (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]} \\
&\qquad + \frac{1}{2\sqrt{\lambda}}\left(\left\|\bar{\boldsymbol{\theta}} \odot \mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}]}\right\|_2^2 + \frac{1}{{\lambda'}^2}\left\|\bar{\boldsymbol{\sigma}} - \boldsymbol{\beta}\right\|_2^2\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \\
&\qquad + \frac{1}{\lambda}\left(2\ln|\Lambda| + \ln|\Lambda'| + n \ln|\mathcal{B}| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m}.
\end{aligned} \tag{62}
$$

Thus,

$$
\begin{aligned}
&\forall \lambda' \in \Lambda', \forall \lambda \in \Lambda, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\lambda'}{2\sqrt{\lambda}} \bar{\boldsymbol{\theta}}^\top (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \bar{\boldsymbol{\theta}} + \frac{1}{2\sqrt{\lambda}} \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]^\top} (\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}) \mathrm{e}^{[\bar{\boldsymbol{\sigma}}]} \\
&\qquad + \frac{1}{2\sqrt{\lambda}} \min_{\substack{\boldsymbol{\beta} \in \mathcal{B}^n \\ \bar{\boldsymbol{\sigma}} \in \mathbb{R}^n}} \left(\left\|\bar{\boldsymbol{\theta}} \odot \mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}]}\right\|_2^2 + \frac{1}{{\lambda'}^2}\left\|\bar{\boldsymbol{\sigma}} - \boldsymbol{\beta}\right\|_2^2\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \\
&\qquad + \frac{1}{\lambda}\left(2\ln|\Lambda| + \ln|\Lambda'| + n \ln|\mathcal{B}| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m}.
\end{aligned} \tag{63}
$$

D.3. Proposition 4.2

First, we change our design of the prior and the posterior. Let $\mathcal{B}$ be a set of positive real numbers such that $\mathcal{B} \subset \mathbb{R}_{>0}$. We describe our choice of $\mathcal{B}$ later.

$$\mathbb{P}\left(\boldsymbol{\mu}\right) = \mathcal{N}\left(\boldsymbol{\mu} \mid \mathbf{0}, \boldsymbol{I}\right), \tag{64}$$

$$\mathbb{P}\left(\boldsymbol{\sigma}\right) = \begin{cases} \frac{1}{|\mathcal{B}|^n}, & \text{if } \boldsymbol{\sigma} \in \mathcal{B}^n \\ 0, & \text{otherwise.} \end{cases} \tag{65}$$

$$\mathbb{Q}\left(\boldsymbol{\mu} \mid \bar{\boldsymbol{\mu}}\right) = \mathcal{N}\left(\boldsymbol{\mu} \mid \bar{\boldsymbol{\mu}}, \boldsymbol{I}\right), \tag{66}$$

$$\mathbb{Q}\left(\boldsymbol{\sigma} \mid \bar{\boldsymbol{\sigma}}\right) = \begin{cases} 1, & \text{if } \boldsymbol{\sigma} = \bar{\boldsymbol{\sigma}} \\ 0, & \text{otherwise,} \end{cases} \tag{67}$$

where $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\sigma}}$ are vectors which satisfies $\boldsymbol{\theta} = \eta\bar{\boldsymbol{\mu}} \odot \bar{\boldsymbol{\sigma}}$. Now, (A) term and (B) term in the PAC-Bayes bound (4) can be written as follows.

$$\text{(A)} : \frac{\eta^2}{2}\sum_{i=1}^{n}\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right), \tag{68}$$

$$\text{(B)} : \begin{cases} \frac{1}{\lambda}\left(\frac{1}{2\eta^2}\sum_{i=1}^{n}\boldsymbol{\theta}_i{}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]} + n\ln|\mathcal{B}|\right), & \text{if } \bar{\boldsymbol{\sigma}} \in \mathcal{B}^n \\ \infty, & \text{otherwise,} \end{cases} \tag{69}$$

Let $H = \{x^{-\frac{1}{4}} | x \in \Lambda\}$. Taking union bound for η over H, with probability at least $1-\delta$,

$$\begin{aligned} &\forall \eta \in H, \forall \bar{\boldsymbol{\sigma}} \in \mathcal{B}^n, \forall \lambda \in \Lambda, \\ &\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\eta^2}{2}\sum_{i=1}^{n}\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + \frac{1}{2\lambda\eta^2}\sum_{i=1}^{n}\boldsymbol{\theta}_i{}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]} \\ &\qquad + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) + \frac{1}{\lambda}\left(\ln|H| + n\ln|\mathcal{B}| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m}. \end{aligned} \tag{70}$$

By setting $\eta = \lambda^{-\frac{1}{4}}$, we have

$$\begin{aligned} \forall \lambda \in \Lambda, \mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq &\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1}{2\sqrt{\lambda}}\sum_{i=1}^{n}\min_{\bar{\boldsymbol{\sigma}}_i \in \mathcal{B}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + \boldsymbol{\theta}_i{}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right) \\ &+ \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) + \frac{1}{\lambda}\left(\ln|\Lambda| + n\ln|\mathcal{B}| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m}. \end{aligned} \tag{71}$$

Now, we specify the set $\mathcal{B}$. Let

$$\mathcal{B} = \left\{c + \frac{d-c}{p}\left(i + \frac{1}{2}\right)\middle| i \in \{0, 1, \ldots, p-1\}\right\}$$

for some $c, d \in \mathbb{R}$ and $p \in \mathbb{N}$. We specify c, d, and p later so that $\mathcal{B}$ become sufficiently dense.

In the rest of the proof, we bound the error caused by the discretization of $\mathbb{R}$ by $\mathcal{B}$. First, we bound

$$D_i = \min_{\bar{\boldsymbol{\sigma}}_i \in [c,d]}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + \boldsymbol{\theta}_i{}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right) - \inf_{\bar{\boldsymbol{\sigma}}_i \in \mathbb{R}_{>0}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + \boldsymbol{\theta}_i{}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right). \tag{72}$$

Due to

$$\frac{\partial}{\partial\bar{\boldsymbol{\sigma}}_i}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + \boldsymbol{\theta}_i{}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right) = 2\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} - \boldsymbol{\theta}_i{}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right) \tag{73}$$

and convexity of D_i, when

$$\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2c]} \leq {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2c]} \tag{74}$$

and

$$\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2d]} \geq {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2d]}, \tag{75}$$

the minimizer of the second term of D_i lies in $[c, d]$. Thus $D_i = 0$. Otherwise, when

$$\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2c]} > {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2c]}, \tag{76}$$

$$\begin{aligned}
D_i &\leq \min_{\bar{\boldsymbol{\sigma}}_i\in[c,d]}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right)\\
&\leq \left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2c]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2c]}\\
&\leq 2\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2c]}\\
&\leq 2M\mathrm{e}^{[2c]}.
\end{aligned} \tag{77}$$

When

$$\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2d]} \geq {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2d]}, \tag{78}$$

$$\begin{aligned}
D_i &\leq \min_{\bar{\boldsymbol{\sigma}}_i\in[c,d]}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right)\\
&\leq \left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2d]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2d]}\\
&\leq 2{\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2d]}\\
&\leq 2b^2\mathrm{e}^{[-2d]}.
\end{aligned} \tag{79}$$

Thus,

$$D \leq \max\left(2M\mathrm{e}^{[2c]} + 2b^2\mathrm{e}^{[-2d]}\right). \tag{80}$$

Next, we bound

$$\begin{aligned}
E_i = &\min_{\bar{\boldsymbol{\sigma}}_i\in\left\{\left(c+\frac{d-c}{p}i\right)\middle|i\in\{0,1,\ldots,p\}\right\}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right)\\
&- \min_{\bar{\boldsymbol{\sigma}}_i\in[c,d]}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right).
\end{aligned} \tag{81}$$

Let

$$\bar{\boldsymbol{\sigma}}_i^* = \operatorname*{argmin}_{\bar{\boldsymbol{\sigma}}_i\in[c,d]}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right), \tag{82}$$

$$j^* = \operatorname*{argmin}_{j}\left\{\left|\left(c + \frac{d-c}{p}\left(j+\frac{1}{2}\right)\right) - \bar{\boldsymbol{\sigma}}_i^*\right|\middle| j\in\{0,1,\ldots,p-1\}\right\}, \tag{83}$$

and

$$y^* = c + \frac{d-c}{p}\left(j^* + \frac{1}{2}\right). \tag{84}$$

Now,

$$\begin{aligned}
E_i &\leq \left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2y^*]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2y^*]}\right) - \left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}^*_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}^*_i]}\right) \\
&\leq \left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}^*_i]}\left|\mathrm{e}^{[2y^*-2\bar{\boldsymbol{\sigma}}^*_i]} - 1\right| + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}^*_i]}\left|\mathrm{e}^{[-2y^*+2\bar{\boldsymbol{\sigma}}^*_i]} - 1\right| \\
&\leq \left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}^*_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}^*_i]}\right)\left(\mathrm{e}^{[2|y^*-\bar{\boldsymbol{\sigma}}^*_i|]} - 1\right) \\
&\leq \left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}^*_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}^*_i]}\right)\left(\mathrm{e}^{\left[\frac{d-c}{p}\right]} - 1\right) \\
&= \left(\min_{\bar{\boldsymbol{\sigma}}_i\in[c,d]}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\bar{\boldsymbol{\sigma}}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right)\right)\left(\mathrm{e}^{\left[\frac{d-c}{p}\right]} - 1\right).
\end{aligned} \tag{85}$$

Thus,

$$\begin{aligned}
&\min_{\boldsymbol{\sigma}_i\in\mathcal{B}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + \bar{\boldsymbol{\theta}}_i\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right) - \inf_{\boldsymbol{\sigma}_i\in\mathbb{R}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\boldsymbol{\sigma}_i]}\right) \\
=&D_i + E_i \\
\leq&D_i + \left(\inf_{\boldsymbol{\sigma}_i\in\mathbb{R}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\boldsymbol{\sigma}_i]}\right) + D_i\right)\left(\mathrm{e}^{\left[\frac{d-c}{p}\right]} - 1\right).
\end{aligned} \tag{86}$$

We set c and d so that $\mathrm{e}^{[2c]} = \frac{\epsilon}{M\sqrt{\lambda}}$ and $\mathrm{e}^{[-2d]} = \frac{\epsilon}{b^2\sqrt{\lambda}}$. Then,

$$\begin{aligned}
&\min_{\boldsymbol{\sigma}_i\in\mathcal{B}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + \bar{\boldsymbol{\theta}}_i^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right) - \inf_{\boldsymbol{\sigma}_i\in\mathbb{R}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\boldsymbol{\sigma}_i]}\right) \\
\leq&\left(\inf_{\boldsymbol{\sigma}_i\in\mathbb{R}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\boldsymbol{\sigma}_i]}\right) + \frac{2}{\sqrt{\lambda}}\epsilon\right)\left(\mathrm{e}^{\left[\frac{d-c}{p}\right]} - 1\right) + \frac{2}{\sqrt{\lambda}}\epsilon.
\end{aligned} \tag{87}$$

For any $\epsilon \in \mathbb{R}_{>0}$, if we choose $\frac{\ln(M)+2\ln(b)+\ln(\lambda)+2\ln\left(\frac{1}{\epsilon}\right)}{\ln(1+\epsilon)} \leq p < \frac{\ln(M)+2\ln(b)+\ln(\lambda)+2\ln\left(\frac{1}{\epsilon}\right)}{\ln(1+\epsilon)}$ + 1,

$$\begin{aligned}
&\min_{\boldsymbol{\sigma}_i\in\mathcal{B}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + (\bar{\boldsymbol{\theta}}_i^2\mathrm{e}^{[-2\bar{\boldsymbol{\sigma}}_i]}\right) - \inf_{\boldsymbol{\sigma}_i\in\mathbb{R}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\boldsymbol{\sigma}_i]}\right) \\
\leq&\inf_{\boldsymbol{\sigma}_i\in\mathbb{R}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\boldsymbol{\sigma}_i]}\right)\epsilon + \frac{2}{\sqrt{\lambda}}(\epsilon + \epsilon^2).
\end{aligned} \tag{88}$$

Thus,

$$\begin{aligned}
&\forall\lambda\in\Lambda, \forall\epsilon\in\mathbb{R}_{>0}, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta}\mid\bar{\boldsymbol{\theta}}\right)\right) \\
\leq&\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1+\epsilon+\epsilon^2}{2\sqrt{\lambda}}\sum_{i=1}^{n}\inf_{\boldsymbol{\sigma}_i\in\mathbb{R}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\mathrm{e}^{[2\boldsymbol{\sigma}_i]} + {\boldsymbol{\theta}_i}^2\mathrm{e}^{[-2\boldsymbol{\sigma}_i]}\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \\
+&\frac{1}{\lambda}\left(n(\epsilon+\epsilon^2) + \ln|\Lambda| + n\ln\left(\frac{\ln(M)+2\ln(b)+\ln(\lambda)+2\ln\left(\frac{1}{\epsilon}\right)}{\ln\left(1+\epsilon\right)} + 1\right) + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m} \\
\leq&\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1+\epsilon+\epsilon^2}{\sqrt{\lambda}}\sum_{i=1}^{n}\sqrt{\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}{\boldsymbol{\theta}_i}^2} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \\
+&\frac{1}{\lambda}\left(n(\epsilon+\epsilon^2) + \ln|\Lambda| + n\ln\left(\frac{\ln(M)+2\ln(b)+\ln(\lambda)+2\ln\left(\frac{1}{\epsilon}\right)}{\ln\left(1+\epsilon\right)} + 1\right) + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m}.
\end{aligned} \tag{89}$$

D.4. Proposition 6.1

First, we describe our design of the prior and the posterior. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be a set of positive real numbers such that $\mathcal{B}_1 \subset \mathbb{R}_{\geq 0}$ and $\mathcal{B}_2 \subset \mathbb{R}_{\geq 0}$. We describe our choice of $\mathcal{B}_1$ and $\mathcal{B}_2$ later.

$$\mathbb{P}\left(V^{(l)}\right) = \mathcal{N}\left(V^{(l)} \mid \mathbf{0}, \boldsymbol{I}\right), \tag{90}$$

$$\mathbb{P}\left(\boldsymbol{\gamma}^{(l)}\right) = \begin{cases} \frac{1}{|\mathcal{B}_1|^{h_1^{(l)}}}, & \text{if } \boldsymbol{\gamma}^{(l)} \in \mathcal{B}_1^{h_1^{(l)}} \\ 0, & \text{otherwise.} \end{cases}, \tag{91}$$

$$\mathbb{P}\left(\boldsymbol{\gamma}'^{(l)}\right) = \begin{cases} \frac{1}{|\mathcal{B}_2|^{h_2^{(l)}}}, & \text{if } \boldsymbol{\gamma}'^{(l)} \in \mathcal{B}_2^{h_2^{(l)}} \\ 0, & \text{otherwise.} \end{cases} \tag{92}$$

$$\mathbb{Q}\left(V^{(l)} \mid \bar{V}^{(l)}\right) = \mathcal{N}\left(V^{(l)} \mid \bar{V}^{(l)}, \boldsymbol{I}\right), \tag{93}$$

$$\mathbb{Q}\left(\boldsymbol{\gamma}^{(l)} \mid \bar{\boldsymbol{\gamma}}^{(l)}\right) = \begin{cases} 1, & \text{if } \boldsymbol{\gamma}^{(l)} = \bar{\boldsymbol{\gamma}}^{(l)} \\ 0, & \text{otherwise,} \end{cases} \tag{94}$$

$$\mathbb{Q}\left(\boldsymbol{\gamma}'^{(l)} \mid \bar{\boldsymbol{\gamma}}'^{(l)}\right) = \begin{cases} 1, & \text{if } \boldsymbol{\gamma}'^{(l)} = \bar{\boldsymbol{\gamma}}'^{(l)} \\ 0, & \text{otherwise.} \end{cases} \tag{95}$$

Now, (A) term and (B) term in the PAC-Bayes bound (4) can be written as follows.

$$\text{(A)}: \frac{\eta^2}{2} \sum_{l=1}^{d} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{H}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right), \tag{96}$$

$$\text{(B)}: \begin{cases} \frac{1}{\lambda}\left(\frac{1}{2\eta^2} \sum_{l=1}^{d} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{U}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + \sum_{l=1}^{d}\left(h_1^{(l)} \ln|\mathcal{B}_1| + h_2^{(l)} \ln|\mathcal{B}_2|\right)\right), & \text{if } \bar{\boldsymbol{\gamma}}^{(l)} \in \mathcal{B}_1^{h_1^{(l)}} \text{ and } \bar{\boldsymbol{\gamma}}'^{(l)} \in \mathcal{B}_2^{h_2^{(l)}} \\ \infty, & \text{otherwise.} \end{cases} \tag{97}$$

Let $H = \{x^{-\frac{1}{4}} | x \in \Lambda\}$. Taking union bound for η over H, with probability at least $1 - \delta$,

$$\begin{aligned}
&\forall \eta \in H, \left(\forall l \in \{1, \ldots, l\}, \left(\forall \bar{\boldsymbol{\gamma}}^{(l)} \in \mathcal{B}_1^{h_1^{(l)}}, \forall \bar{\boldsymbol{\gamma}}'^{(l)} \in \mathcal{B}_2^{h_2^{(l)}}\right)\right), \forall \lambda \in \Lambda, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \\
&\leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\eta^2}{2} \sum_{l=1}^{d} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{H}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + \frac{1}{2\lambda\eta^2} \sum_{l=1}^{d} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{U}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]} \\
&+ \frac{1}{\lambda}\left(\ln|\Lambda| + h_1 d \ln|\mathcal{B}_1| + h_2 d \ln|\mathcal{B}_2| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right),
\end{aligned} \tag{98}$$

where

$$h_1 := \frac{1}{d} \sum_{i=1}^{d} h_1^{(l)}, \tag{99}$$

$$h_2 := \frac{1}{d} \sum_{i=1}^{d} h_2^{(l)}. \tag{100}$$

By setting $\eta = \lambda^{-\frac{1}{4}}$, we have

$$\begin{aligned}&\left(\forall l \in \{1, \dots, l\}, \left(\forall \bar{\boldsymbol{\gamma}}^{(l)} \in \mathcal{B}_1^{h_1^{(l)}}, \forall \bar{\boldsymbol{\gamma}}'^{(l)} \in \mathcal{B}_2^{h_2^{(l)}}\right)\right), \forall \lambda \in \Lambda, \\ &\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \\ &\leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1}{2\sqrt{\lambda}} \sum_{l=1}^{d} \left({\mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]}}^\top \bar{H}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + {\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]}}^\top \bar{U}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]} \right) \\ &+ \frac{1}{\lambda}\left(\ln|\Lambda| + h_1 d \ln|\mathcal{B}_1| + h_2 d \ln|\mathcal{B}_2| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right).\end{aligned} \tag{101}$$

Thus,

$$\begin{aligned}\forall \lambda \in \Lambda, \mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq &\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1}{2\sqrt{\lambda}} \sum_{l=1}^{d} \min_{\bar{\boldsymbol{\gamma}}^{(l)} \in \mathcal{B}_1^{h_1^{(l)}}, \bar{\boldsymbol{\gamma}}'^{(l)} \in \mathcal{B}_2^{h_2^{(l)}}} \left({\mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]}}^\top \bar{H}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + {\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]}}^\top \bar{U}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]} \right) \\ &+ \frac{1}{\lambda}\left(\ln|\Lambda| + h_1 d \ln|\mathcal{B}_1| + h_2 d \ln|\mathcal{B}_2| + \ln\frac{1}{\delta}\right) + \frac{\lambda}{2m} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right).\end{aligned} \tag{102}$$

In the rest of the proof, we bound

$$\min_{\boldsymbol{x} \in \mathcal{B}_1^{h_1^{(l)}}, \boldsymbol{y} \in \mathcal{B}_2^{h_2^{(l)}}} \left({\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]} \right) - \inf_{\boldsymbol{x} \in [-R,R]^{h_1^{(l)}}, \boldsymbol{y} \in \mathbb{R}^{h_2^{(l)}}} \left({\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]} \right) \tag{103}$$

by choosing $\mathcal{B}_1$ and $\mathcal{B}_2$ appropriately.

First, we bound

$$\begin{aligned}D_l := &\min_{\boldsymbol{x} \in [-R,R]^{h_1^{(l)}}, \boldsymbol{y} \in \mathcal{B}_2^{h_2^{(l)}}} \left({\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]} \right) \\ &- \inf_{\boldsymbol{x} \in [-R,R]^{h_1^{(l)}}, \boldsymbol{y} \in \mathbb{R}^{h_2^{(l)}}} \left({\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]} \right)\end{aligned} \tag{104}$$

by choosing $\mathcal{B}_2$. Let $\mathcal{B}_2$ be a set of real numbers defined as follows.

$$\mathcal{B}_2 = \left\{ c + \frac{d-c}{p} i \middle| i \in \{0, 1, \dots, p\} \right\} \tag{105}$$

for some $c, d \in \mathbb{R}$ such that $c < d$ and $p \in \mathbb{N}$. We specify c, d, and p later. It is straightforward to bound D_l if we can bound the following quantity for arbitral $j \in \{1, \dots, h_1^{(l)}\}$ and $\boldsymbol{x} \in [-R, R]^{h_1^{(l)}}$:

$$D'_{l,j} := \min_{\boldsymbol{y}_j \in \mathcal{B}_2} \left({\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}_j]} + {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}_j]} \right) - \inf_{\boldsymbol{y}_j \in \mathbb{R}} \left({\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}_j]} + {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}_j]} \right). \tag{106}$$

Since

$$\frac{\partial}{\partial \boldsymbol{y}_i} \left({\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}_j]} + {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}_j]} \right) = {\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}_j]} - {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}_j]}, \tag{107}$$

when

$${\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[c]} \leq {\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-c]} \tag{108}$$

and

$${\mathrm{e}^{[-\boldsymbol{x}]}}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-d]} \leq {\mathrm{e}^{[\boldsymbol{x}]}}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[d]} \tag{109}$$

there exists a minimizer of the second term of $D'_{l,j}$ in $[c, d]$. We bound $D'_{l,j}$ by division into cases.

Case 1:

$$\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[c]} \le \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-c]} \quad \text{and} \quad \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-d]} \le \mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[d]} \tag{110}$$

Let

$$\boldsymbol{y}_j{}^* = \operatorname*{argmin}_{\boldsymbol{y}_j \in [-R,R]} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j]}\right), \tag{111}$$

$$k^* = \operatorname*{argmin}_{k} \left\{ \left| \left(c + \frac{d-c}{p}k\right) - \boldsymbol{y}_j{}^* \right| \middle| k \in \{0, 1, \dots, p\} \right\}, \tag{112}$$

$$z^* = c + \frac{d-c}{p}k^*. \tag{113}$$

Note, $c \le \boldsymbol{y}_j{}^* \le d$ by the assumption. Then,

$$D'_{l,j} \le \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[z^*]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-z^*]}\right) - \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j{}^*]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j{}^*]}\right) \tag{114}$$

$$\le \left|\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[z^*]} - \mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j{}^*]}\right| + \left|\mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-z^*]} - \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j{}^*]}\right| \tag{115}$$

$$\le \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j{}^*]}\right)\left(\mathrm{e}^{[|\boldsymbol{y}_j{}^* - z^*|]} - 1\right) + \left(\mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j{}^*]}\right)\left(\mathrm{e}^{[|\boldsymbol{y}_j{}^* - z^*|]} - 1\right) \tag{116}$$

$$\le \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j{}^*]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j{}^*]}\right)\left(\mathrm{e}^{\left[\frac{d-c}{2p}\right]} - 1\right). \tag{117}$$

Case 2:

$$\mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-c]} \le \mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[c]} \tag{118}$$

Since

$$\min_{\boldsymbol{y}_j \in \mathcal{B}_2} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j]}\right) = \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[c]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-c]}\right), \tag{119}$$

$$D'_{l,j} \le \mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[c]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-c]} \tag{120}$$

$$\le 2\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[c]} \tag{121}$$

$$\le 2Mh_1^{(l)}e^R\mathrm{e}^{[c]}. \tag{122}$$

Case 3:

$$\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[d]} \le \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-d]} \tag{123}$$

Since

$$\min_{\boldsymbol{y}_j \in \mathcal{B}_2} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j]}\right) = \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[d]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-d]}\right), \tag{124}$$

$$D'_{l,j} \le \mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[c]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-c]} \tag{125}$$

$$\le 2\mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-d]} \tag{126}$$

$$\le 2b^2h_1^{(l)}e^R\mathrm{e}^{[-d]}. \tag{127}$$

Combining case 1–3, we have

$$D'_{l,j} \le \left(\inf_{\boldsymbol{y}_j \in \mathbb{R}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}_j]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}_j]}\right)\right)\left(\mathrm{e}^{\left[\frac{d-c}{2p}\right]} - 1\right)$$

$$+ \max\left(2Mh_1^{(l)}e^R\mathrm{e}^{[c]} + 2b^2h_1^{(l)}e^R\mathrm{e}^{[-d]}\right). \tag{128}$$

We set c, d, and p so that

$$\mathrm{e}^{[c]} = \frac{hd\epsilon}{nM\sqrt{\lambda}e^{R}}, \tag{129}$$

$$\mathrm{e}^{[-d]} = \frac{hd\epsilon}{nb^2\sqrt{\lambda}e^{R}}, \tag{130}$$

$$\frac{d-c}{2\ln(1+\epsilon)} \leq p < \frac{d-c}{2\ln(1+\epsilon)} + 1. \tag{131}$$

Then,

$$D_l \leq \inf_{\boldsymbol{y}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j}\mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j}\mathrm{e}^{[-\boldsymbol{y}]}\right)\epsilon + 2\frac{h_1^{(l)}h_2^{(l)}}{n\sqrt{\lambda}}hd\epsilon. \tag{132}$$

Next, we bound

$$\begin{aligned} E_l := &\min_{\boldsymbol{x}\in\mathcal{B}_1^{h_1^{(l)}},\boldsymbol{y}\in\mathcal{B}_2^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}]}\right) \\ &- \min_{\boldsymbol{x}\in[-R,R]^{h_1^{(l)}},\boldsymbol{y}\in\mathcal{B}_2^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}]}\right). \end{aligned} \tag{133}$$

by choosing $\mathcal{B}_1$. Let $\mathcal{B}_1$ be a set of real numbers defined as follows.

$$\mathcal{B}_1 = \left\{-R + \frac{2R}{p'}\left(i+\frac{1}{2}\right)\middle| i\in\{0,1,\ldots,p'-1\}\right\}. \tag{134}$$

Let

$$\boldsymbol{x}^*, \boldsymbol{y}^* = \operatorname*{argmin}_{\boldsymbol{x}\in[-R,R]^{h_1^{(l)}},\boldsymbol{y}\in\mathcal{B}_2^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}]}\right), \tag{135}$$

$$\boldsymbol{j}^* = \operatorname*{argmin}_{\boldsymbol{j}} \left\{\left|\left(-R\mathbf{1} + \frac{2R}{p'}\odot\left(\boldsymbol{j}+\frac{1}{2}\mathbf{1}\right)\right) - \boldsymbol{x}^*\right|\middle| \boldsymbol{j}\in\{0,1,\ldots,p'-1\}^{h_1^{(l)}}\right\}, \tag{136}$$

and

$$\boldsymbol{z}^* = -R\mathbf{1} + \frac{2R}{p'}\odot\left(\boldsymbol{j}^*+\frac{1}{2}\mathbf{1}\right). \tag{137}$$

Then,

$$\begin{aligned} E_l &\leq \left(\mathrm{e}^{[\boldsymbol{z}^*]}{}^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}^*]} + \mathrm{e}^{[-\boldsymbol{z}^*]}{}^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}^*]}\right) - \left(\mathrm{e}^{[\boldsymbol{x}^*]}{}^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}^*]} + \mathrm{e}^{[-\boldsymbol{x}^*]}{}^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}^*]}\right) && (138) \\ &\leq \left(\mathrm{e}^{[\boldsymbol{x}^*]}\odot\left(\mathrm{e}^{[|\boldsymbol{z}^*-\boldsymbol{x}^*|]}-\mathbf{1}\right)\right)^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}^*]} + \left(\mathrm{e}^{[-\boldsymbol{x}^*]}\odot\left(\mathrm{e}^{[|\boldsymbol{z}^*-\boldsymbol{x}^*|]}-\mathbf{1}\right)\right)^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}^*]} && (139) \\ &\leq \left(\mathrm{e}^{[\boldsymbol{x}^*]}\odot\left(e^{\frac{R}{p'}\mathbf{1}}-\mathbf{1}\right)\right)^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}^*]} + \left(\mathrm{e}^{[-\boldsymbol{x}^*]}\odot\left(e^{\frac{R}{p'}\mathbf{1}}-\mathbf{1}\right)\right)^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}^*]} && (140) \\ &= \left(\mathrm{e}^{[\boldsymbol{x}^*]}{}^\top \bar{H}^{(l)}\mathrm{e}^{[\boldsymbol{y}^*]} + \mathrm{e}^{[-\boldsymbol{x}^*]}{}^\top \bar{U}^{(l)}\mathrm{e}^{[-\boldsymbol{y}^*]}\right)\left(e^{\frac{R}{p'}}-1\right). && (141) \end{aligned}$$

We set p' so that

$$\frac{R}{\ln(1+\epsilon)} \leq p' < \frac{R}{\ln(1+\epsilon)} + 1. \tag{142}$$

Then,

$$E_l \leq \min_{\boldsymbol{x}\in[-R,R]^{h_1^{(l)}}, \boldsymbol{y}\in\mathcal{B}_2^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]}\right) \epsilon. \tag{143}$$

Combining Eq. (132) and Eq. (143),

$$\begin{aligned}
&\sum_{l=1}^{d} \Bigg(\min_{\boldsymbol{x}\in\mathcal{B}_1^{h_1^{(l)}}, \boldsymbol{y}\in\mathcal{B}_2^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]}\right) \\
&\qquad - \inf_{\boldsymbol{x}\in[-R,R]^{h_1^{(l)}}, \boldsymbol{y}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]}\right) \Bigg) && (144) \\
&= \sum_{l=1}^{d} \left(D_l + E_l\right) && (145) \\
&\leq \sum_{l=1}^{d} \Bigg(\inf_{\boldsymbol{y}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}]}\right) \epsilon + 2\frac{h_1^{(l)} h_2^{(l)}}{n\sqrt{\lambda}} hd\epsilon && (146) \\
&\quad + \min_{\boldsymbol{x}\in[-R,R]^{h_1^{(l)}}, \boldsymbol{y}\in\mathcal{B}_2^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]}\right) \epsilon \Bigg) && (147) \\
&\leq \sum_{l=1}^{d} \Bigg(\inf_{\boldsymbol{y}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}]}\right) \epsilon + 2\frac{h_1^{(l)} h_2^{(l)}}{n\sqrt{\lambda}} hd\epsilon && (148) \\
&\quad + \left(\inf_{\boldsymbol{y}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]}\right) (1+\epsilon) + 2\frac{h_1^{(l)} h_2^{(l)}}{n\sqrt{\lambda}} hd\epsilon \right) \epsilon \Bigg) && (149) \\
&= \sum_{l=1}^{d} \left(\inf_{\boldsymbol{y}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}]}\right) \left(2\epsilon + \epsilon^2\right) + 2\frac{h_1^{(l)} h_2^{(l)}}{n\sqrt{\lambda}} hd \left(\epsilon + \epsilon^2\right) \right) && (150) \\
&= \sum_{l=1}^{d} \inf_{\boldsymbol{y}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)}_{:,j} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)}_{:,j} \mathrm{e}^{[-\boldsymbol{y}]}\right) \left(2\epsilon + \epsilon^2\right) + 2\frac{hd}{\sqrt{\lambda}} \left(\epsilon + \epsilon^2\right) && (151)
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\forall \lambda \in \Lambda, \forall \epsilon \in \mathbb{R}_{>0}, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \\
\leq &\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\left(1+\epsilon\right)^2}{2\sqrt{\lambda}} \sum_{l=1}^{d} \inf_{\boldsymbol{x} \in [-R,R]^{h_1^{(l)}}, \boldsymbol{y} \in \mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]}\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) + \frac{\lambda}{2m} \\
&+ \frac{1}{\lambda}\left(h_2\left(1+\epsilon\right)\epsilon + \ln|\Lambda| + h_2 d \ln\left(\frac{\frac{1}{2}\ln(M) + \ln(b) + \frac{1}{2}\ln(\lambda) + \ln(\frac{n}{hd}) + R + \ln\left(\frac{1}{\epsilon}\right)}{\ln\left(1+\epsilon\right)} + 2\right)\right. \\
&\left. + h_1 d \ln\left(\frac{\ln\left(R\right)}{\ln\left(1+\epsilon\right)} + 1\right) + \ln\frac{1}{\delta}\right) \qquad (152) \\
\leq &\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\left(1+\epsilon\right)^2}{2\sqrt{\lambda}} \sum_{l=1}^{d} \inf_{\boldsymbol{x} \in [-R,R]^{h_1^{(l)}}, \boldsymbol{y} \in \mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \mathrm{e}^{[\boldsymbol{y}]} + \mathrm{e}^{[-\boldsymbol{x}]}{}^\top \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{y}]}\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) + \frac{\lambda}{2m} \\
&+ \frac{1}{\lambda}\left(hd\left(1+\epsilon\right)\epsilon + \ln|\Lambda| + 2hd \ln\left(\frac{\frac{1}{2}\ln(M) + \ln(b) + \frac{1}{2}\ln(\lambda) + \ln(\frac{n}{hd}) + R + \ln\left(\frac{1}{\epsilon}\right)}{\ln\left(1+\epsilon\right)} + 2\right) + \ln\frac{1}{\delta}\right) \qquad (153) \\
= &\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{\left(1+\epsilon\right)^2}{\sqrt{\lambda}} \sum_{l=1}^{d} \inf_{\boldsymbol{x} \in [-R,R]^{h_1^{(l)}}} \left(\sqrt{\mathrm{e}^{[\boldsymbol{x}]}{}^\top \bar{H}^{(l)} \bar{U}^{(l)} \mathrm{e}^{[-\boldsymbol{x}]}}\right) + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) + \frac{\lambda}{2m} \\
&+ \frac{1}{\lambda}\left(hd\left(1+\epsilon\right)\epsilon + \ln|\Lambda| + 2hd \ln\left(\frac{\frac{1}{2}\ln(M) + \ln(b) + \frac{1}{2}\ln(\lambda) + \ln(\frac{n}{hd}) + R + \ln\left(\frac{1}{\epsilon}\right)}{\ln\left(1+\epsilon\right)} + 2\right) + \ln\frac{1}{\delta}\right). \qquad (154)
\end{aligned}
$$

D.5. Theorem 6.1

The theorem immediately follows if we bound the following quantity.

$$G_l := \frac{1}{2}\mathrm{NS}^{(l)} - \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*'^{(l)}\right]}, \tag{155}$$

where

$$\bar{\boldsymbol{\gamma}}_*^{(l)}, \bar{\boldsymbol{\gamma}}_*'^{(l)} = \underset{\bar{\boldsymbol{\gamma}}^{(l)} \in \mathcal{B}_1^{h_1^{(l)}}, \bar{\boldsymbol{\gamma}}'^{(l)} \in \mathcal{B}_2^{h_2^{(l)}}}{\operatorname{argmin}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{U}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} \right), \tag{156}$$

$$\mathrm{NS}^{(l)} = \inf_{\bar{\boldsymbol{\gamma}}^{(l)} \in [-R,R]^{h_1^{(l)}}, \bar{\boldsymbol{\gamma}}'^{(l)} \in \mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]^\top} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{U}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} \right). \tag{157}$$

The sets $\mathcal{B}_1$ and $\mathcal{B}_2$ are specified in D.4.

$$G_l \leq \frac{1}{2} \inf_{\bar{\boldsymbol{\gamma}}'^{(l)} \in \mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} \right) - \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*'^{(l)}\right]}. \tag{158}$$

Let

$$G_{l,j} = \frac{1}{2} \inf_{\bar{\boldsymbol{\gamma}}_j'^{(l)} \in \mathbb{R}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} \right) - \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]} \bar{F}^{(l)} \mathrm{e}^{\left[\left(\bar{\boldsymbol{\gamma}}_*'^{(l)}\right)_j\right]}, \tag{159}$$

and c, d are real numbers defined in Sec. D.4. We bound $G_{l,j}$ by division into cases.

Case 1:

$$\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{H}_{:,j}^{(l)} \mathrm{e}^{[c]} \leq \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{[-c]} \quad \text{and} \quad \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{[-d]} \leq \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{H}_{:,j}^{(l)} \mathrm{e}^{[d]} \tag{160}$$

Since

$$\begin{aligned} &\inf_{\bar{\boldsymbol{\gamma}}_j'^{(l)} \in \mathbb{R}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} \right) \\ &= \min_{\bar{\boldsymbol{\gamma}}_j'^{(l)} \in [c,d]} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} \right), \end{aligned} \tag{161}$$

there exists $z \in [c, d]$ such that

$$z = \underset{\bar{\boldsymbol{\gamma}}_j'^{(l)} \in \mathbb{R}^{h_2^{(l)}}}{\operatorname{argmin}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} \right). \tag{162}$$

Note, z satisfies

$$\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{[z]} = \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{[-z]}. \tag{163}$$

Thus,

$$G_{l,j} = \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \left(\mathrm{e}^{[z]} - \mathrm{e}^{\left[\left(\bar{\boldsymbol{\gamma}}_*'^{(l)}\right)_j\right]} \right). \tag{164}$$

Due to the convexity of

$$\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} \bar{U}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]} \tag{165}$$

with respect to $\bar{\boldsymbol{\gamma}}'^{(l)}$ and the choice of $\mathcal{B}_2$,

$$\begin{aligned} G_{l,j} &\leq \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\left(\bar{\boldsymbol{\gamma}}_*'^{(l)}\right)_j\right]} \left(\mathrm{e}^{\left[\frac{d-c}{2p}\right]} - 1 \right) &(166) \\ &\leq \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\left(\bar{\boldsymbol{\gamma}}_*'^{(l)}\right)_j\right]} \epsilon. &(167) \end{aligned}$$

Case 2:

$$\mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{[-c]} \leq \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{[c]} \tag{168}$$

By the choice of c,

$$\begin{align}
G_{l,j} &\leq \frac{1}{2} \inf_{\bar{\boldsymbol{\gamma}}_j'^{(l)} \in \mathbb{R}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} \right) \tag{169}\\
&\leq \frac{1}{2} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{[c]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{[-c]} \right) \tag{170}\\
&\leq \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{[c]} \tag{171}\\
&\leq \frac{hdh_2^{(l)}}{n\sqrt{\lambda}} \epsilon. \tag{172}
\end{align}$$

Case 3:

$$\mathrm{e}^{[\boldsymbol{x}]^\top} \bar{H}_{:,j}^{(l)} \mathrm{e}^{[d]} \leq \mathrm{e}^{[-\boldsymbol{x}]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{[-d]} \tag{173}$$

By the choice of d,

$$\begin{align}
G_{l,j} &\leq \frac{1}{2} \inf_{\bar{\boldsymbol{\gamma}}_j'^{(l)} \in \mathbb{R}} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} \right) \tag{174}\\
&\leq \frac{1}{2} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{[d]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{[-d]} \right) \tag{175}\\
&\leq \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}_{:,j}^{(l)} \mathrm{e}^{[-d]} \tag{176}\\
&\leq \frac{hdh_2^{(l)}}{n\sqrt{\lambda}} \epsilon. \tag{177}
\end{align}$$

Combining case 1–3,

$$G_{l,j} \leq \frac{1}{2} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}_{:,j}^{(l)} \mathrm{e}^{\left[\left(\bar{\boldsymbol{\gamma}}_*'^{(l)}\right)_j\right]} \epsilon + \frac{hdh_2^{(l)}}{n\sqrt{\lambda}} \epsilon. \tag{178}$$

Thus,

$$\begin{align}
G_l &\leq \frac{1}{2} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*'^{(l)}\right]} \epsilon + \frac{hdh_1^{(l)} h_2^{(l)}}{n\sqrt{\lambda}} \epsilon \tag{179}\\
&\leq \frac{1}{2} \left(\mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{F}^{(l)} \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}_*'^{(l)}\right]} + \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*^{(l)}\right]^\top} \bar{U}^{(l)} \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}_*'^{(l)}\right]} \right) \epsilon + \frac{hdh_1^{(l)} h_2^{(l)}}{n\sqrt{\lambda}} \epsilon. \tag{180}
\end{align}$$

Using Eq. (132) and Eq. (143),

$$\begin{align}
\sum_{l=1}^{d} G_l &\leq \frac{1}{2} \left(\sum_{l=1}^{d} \mathrm{NS}^{(l)} \left(1 + 2\epsilon + \epsilon^2\right) + 2 \frac{hd}{\sqrt{\lambda}} \left(\epsilon + \epsilon^2\right) \right) \epsilon + \frac{hd}{\sqrt{\lambda}} \epsilon \tag{181}\\
&= \frac{\epsilon(1+\epsilon)^2}{2} \sum_{l=1}^{d} \mathrm{NS}^{(l)} + \frac{hd}{\sqrt{\lambda}} \left(\epsilon + \epsilon^2 + \epsilon^3\right). \tag{182}
\end{align}$$

Since

$$\mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right) \leq \mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) - \left(\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \frac{1}{2\sqrt{\lambda}} \sum_{l=1}^{d} \mathrm{NS}^{(l)} \right) + \frac{1}{2\sqrt{\lambda}} \sum_{l=1}^{d} G_l, \tag{183}$$

we can now substitute $\mathcal{E}\left(\mathcal{L}_{\mathcal{S}}, \mathbb{Q}\right)$ in Prop. 6.1 using $\bar{F}^{(l)}$ and Eq. (182) as follows.

$$
\begin{aligned}
&\forall \lambda \in \Lambda, \forall \epsilon \in \mathbb{R}_{>0}, \\
&\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \\
\leq & \mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right)+\frac{(1+\epsilon)^2}{2 \sqrt{\lambda}} \sum_{l=1}^d \mathrm{NS}^{(l)}+\frac{\lambda}{2 m} \\
& +\frac{1}{\lambda}\left(h d\left(1+\epsilon\right) \epsilon+\ln |\Lambda|+2 h d \ln \left(\frac{\frac{1}{2} \ln (M)+\ln (b)+\frac{1}{2} \ln (\lambda)+\ln (\frac{n}{h d})+R+\ln \left(\frac{1}{\epsilon}\right)}{\ln \left(1+\epsilon\right)}+2\right)+\ln \frac{1}{\delta}\right) \\
& +\frac{1}{2 \sqrt{\lambda}}\left(\frac{\epsilon(1+\epsilon)^2}{2} \sum_{l=1}^d \mathrm{NS}^{(l)}+\frac{h d}{\sqrt{\lambda}}\left(\epsilon+\epsilon^2+\epsilon^3\right)\right)-\frac{1}{4 \sqrt{\lambda}} \sum_{l=1}^d \mathrm{NS}^{(l)} \qquad (184)\\
= & \mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right)+\frac{1+5 \epsilon+4 \epsilon^2+\epsilon^3}{4 \sqrt{\lambda}} \sum_{l=1}^d \mathrm{NS}^{(l)}+\frac{\lambda}{2 m} \\
& +\frac{1}{\lambda}\left(\frac{3 \epsilon+3 \epsilon^2+\epsilon^3}{2} h+\ln |\Lambda|+2 h d \ln \left(\frac{\frac{1}{2} \ln (M)+\ln (b)+\frac{1}{2} \ln (\lambda)+\ln (\frac{n}{h d})+R+\ln \left(\frac{1}{\epsilon}\right)}{\ln \left(1+\epsilon\right)}+2\right)+\ln \frac{1}{\delta}\right) \qquad (185)\\
= & \mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right)+\frac{1+5 \epsilon+4 \epsilon^2+\epsilon^3}{2 \sqrt{\lambda}} \mathrm{NS}_2(R)+\frac{\lambda}{2 m} \\
& +\frac{1}{\lambda}\left(\frac{3 \epsilon+3 \epsilon^2+\epsilon^3}{2} h d+\ln |\Lambda|+2 h d \ln \left(\frac{\frac{1}{2} \ln (M)+\ln (b)+\frac{1}{2} \ln (\lambda)+\ln (\frac{n}{h d})+R+\ln \left(\frac{1}{\epsilon}\right)}{\ln \left(1+\epsilon\right)}+2\right)+\ln \frac{1}{\delta}\right) . \qquad (186)
\end{aligned}
$$

Using $\epsilon \leq 0.1$, we recover the theorem.

E. Convexity of variance optimization

We show that the optimization problem (35) is convex with respect to the scaling parameters $\bar{\boldsymbol{\gamma}}^{(l)}$. It suffices to show the convexity of the following minimization problem.

$$\underset{\bar{\boldsymbol{\gamma}}^{(l)} \in \mathbb{R}^{h_1^{(l)}}}{\text{minimize}} \quad S := \mathrm{e}^{\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top} A \mathrm{e}^{\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]}, \tag{187}$$

where A is a nonnegative matrix.

First, we calculate the elements of the Hessian of S.

$$\frac{\partial^2 S}{\partial \bar{\gamma}_i^{(l)} \partial \bar{\gamma}_j^{(l)}} = \frac{\partial S}{\partial \bar{\gamma}_i^{(l)}} \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_j\right]} A_{j,:} \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}\right]} - \mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}\right]} A_{:,j} \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_j\right]} \right) \tag{188}$$

$$= \delta_{i,j} \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_j\right]} A_{j,:} \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}\right]} + \mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}\right]} A_{:,j} \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_j\right]} \right) - \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_j\right]} \left(A_{j,i} + A_{i,j}\right) \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_i\right]} \right) \tag{189}$$

Now, it suffices to show that $\forall v, v^\top \left(\nabla^2_{\bar{\boldsymbol{\gamma}}^{(l)}} S \right) v \geq 0$.

$$v^\top \left(\nabla^2_{\bar{\boldsymbol{\gamma}}^{(l)}} S \right) v \tag{190}$$

$$= \sum_i \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_j\right]} A_{j,:} \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}\right]} + \mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}\right]} A_{:,j} \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_j\right]} \right) v_i^2 - \sum_{i \neq j} \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_j\right]} \left(A_{j,i} + A_{i,j}\right) \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_i\right]} \right) v_i v_j \tag{191}$$

$$= \sum_{i,j} \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_i\right]} \left(A_{i,j} + A_{j,i}\right) \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_j\right]} \right) v_i^2 - \sum_{i,j} \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_j\right]} \left(A_{j,i} + A_{i,j}\right) \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_i\right]} \right) v_i v_j \tag{192}$$

$$= \frac{1}{2} \sum_{i,j} \left(\mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_i\right]} \left(A_{i,j} + A_{j,i}\right) \mathrm{e}^{\left[-\boldsymbol{\gamma}'^{(l)}{}_j\right]} \right) \left(v_i - v_j\right)^2 \tag{193}$$

$$\geq 0 \quad \square \tag{194}$$

F. Normalized sharpness for convolutional layers

Let $K^{(l)} \in \mathbb{R}^{c_\text{out} \times c_\text{in} \times h \times w}$ be a kernel of a convolutional layer, where $c_\text{in}, c_\text{out}, h, w$ are input channel size, output channel size, kernel height, and kernel width, respectively. Since the scale dependence only appears in the input and output channels, we can use the following reparametrization of K.

$$K^{(l)}{}_{i,j,a,b} = \eta \mathrm{e}^{\left[\boldsymbol{\gamma}^{(l)}{}_i\right]} V^{(l)}{}_{i,j,a,b} \mathrm{e}^{\left[\boldsymbol{\gamma}'^{(l)}{}_j\right]}, \tag{195}$$

where $\boldsymbol{\gamma}^{(l)} \in \mathbb{R}^{c_\text{out}}, \boldsymbol{\gamma}'^{(l)} \in \mathbb{R}^{c_\text{in}}$ are scaling parameters and $V^{(l)} \in \mathbb{R}^{c_\text{out} \times c_\text{in} \times h \times w}$ is a normalized kernel. Using the same discussion with the noramalized sharpness for fully-connected layers, we have

$$\begin{aligned}
&\mathrm{KL}[\mathbb{Q}\left(V^{(l)}, \boldsymbol{\gamma}, \boldsymbol{\gamma}' \mid \bar{V}^{(l)}, \bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\gamma}}'\right) \parallel \mathbb{P}\left(V^{(l)}, \boldsymbol{\gamma}, \boldsymbol{\gamma}'\right)] \\
=&\frac{1}{\eta^2} \sum_{i=1}^{c_\text{out}} \sum_{j=1}^{c_\text{in}} \sum_{a=1}^{h} \sum_{b=1}^{w} \left(\bar{K}^{(l)}_{i,j,a,b}\right)^2 \mathrm{e}^{\left[-2\boldsymbol{\gamma}^{(l)}{}_i - 2\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \text{const.} \\
=&\frac{1}{\eta^2} \sum_{i=1}^{c_\text{out}} \sum_{j=1}^{c_\text{in}} \sum_{a=1}^{h} \sum_{b=1}^{w} \left(\bar{K}^{(l)}_{i,j,a,b}\right)^2 \mathrm{e}^{\left[-2\boldsymbol{\gamma}^{(l)}{}_i - 2\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \text{const.}
\end{aligned} \tag{196}$$

Also, let

$$\left(\nabla^2_{K^{(l)}} \mathcal{L}_\mathcal{S}\big|_{\bar{K}^{(l)}}\right)_{i,j,a,b} = \frac{\partial^2 \mathcal{L}_\mathcal{S}}{\partial K^{(l)}{}_{i,j,a,b} \partial K^{(l)}{}_{i,j,a,b}}\bigg|_{\bar{K}^{(l)}} \tag{197}$$

and

$$\left(\nabla^2_{K^{(l)}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{K}^{(l)}}\right)_{i,j} = \sum_{a=1}^{h}\sum_{b=1}^{w}\left(\frac{\partial^2\mathcal{L}_{\mathcal{S}}}{\partial {K^{(l)}}_{i,j,a,b}\partial {K^{(l)}}_{i,j,a,b}}\bigg|_{\bar{K}^{(l)}}\right). \tag{198}$$

Then, we have

$$\begin{aligned}
&\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta}\mid\bar{\boldsymbol{\theta}}\right)\right) - \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right)\\
=&\eta^2\sum_{l}^{d}\sum_{i=1}^{c_{\mathrm{out}}}\sum_{j=1}^{c_{\mathrm{in}}}\sum_{a=1}^{h}\sum_{b=1}^{w}\left(\nabla^2_{K^{(l)}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{K}^{(l)}}\right)_{i,j,a,b}\mathrm{e}^{\left[2{\boldsymbol{\gamma}^{(l)}}_i+2{\boldsymbol{\gamma}'^{(l)}}_j\right]} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}},\mathbb{Q}\right)\\
=&\eta^2\sum_{l}^{d}\sum_{i=1}^{c_{\mathrm{out}}}\sum_{j=1}^{c_{\mathrm{in}}}\left(\nabla^2_{K^{(l)}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{K}^{(l)}}\right)_{i,j}\mathrm{e}^{\left[2{\boldsymbol{\gamma}^{(l)}}_i+2{\boldsymbol{\gamma}'^{(l)}}_j\right]} + \mathcal{E}\left(\mathcal{L}_{\mathcal{S}},\mathbb{Q}\right).
\end{aligned} \tag{199}$$

Thus, a natural extension of the normalized sharpness to convolutional networks is as follows.

$$\sum_{l=1}^{d}\min_{\bar{\boldsymbol{\gamma}}^{(l)},\bar{\boldsymbol{\gamma}}'^{(l)}}\sum_{i=1}^{c_{\mathrm{out}}}\sum_{j=1}^{c_{\mathrm{in}}}\left(\left\|\bar{K}^{(l)}_{i,j,:,:}\right\|_{\mathrm{F}}^2\mathrm{e}^{\left[-2\bar{\boldsymbol{\gamma}}_i^{(l)}-2\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]} + \left(\nabla^2_{K^{(l)}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{K}^{(l)}}\right)_{i,j}\mathrm{e}^{\left[2\bar{\boldsymbol{\gamma}}_i^{(l)}+2\bar{\boldsymbol{\gamma}}_j'^{(l)}\right]}\right). \tag{200}$$

G. Considerations on surrogate loss

When we measure the generalization gap using the 0-1 loss, which is not differentiable with respect to parameters, we need to use surrogate loss functions. The choice of the surrogate loss function needs special care when we use flatness for model comparison. The loss is preferable if the value of sharpness metrics do not change when the accuracy of the models does not change. Thus, the surrogate loss function is better to be invariant to a set of transformations that do not change accuracy, such as scaling and shifting of the networks' outputs. For example, the cross-entropy loss after softmax does not satisfy the first condition. Thus, using the loss function makes model comparison less meaningful. While the above conditions do not make the choices of the surrogate loss function unique, we heuristically use the following loss.

$$-\ln\left(\frac{\exp\left(f'(z)_{y_z}\right)}{\sum_i\exp\left(f'(z)_i\right)}\right), \tag{201}$$

where

$$\mu := \frac{1}{K}\sum_i f(z)_i, f'(z) := \frac{f(z)}{\sqrt{\frac{1}{K}\sum_i\left(f(z)_i-\mu\right)^2}}, \tag{202}$$

$f(z)$ is an output of a network, y_z is a label of z, and K is the number of classes. We refer to the loss function as a normalized-softmax-cross-entropy loss. We use the loss function in later experiments (Sec. 8). Note that we do not need to train networks using the loss function, but we use it only when we calculate normalized sharpness for model comparison.

H. Error in second-order approximation

While the Hessian has been a popular measure of flatness, we discuss when the approximation is reasonable.

First, we review Eq.(6). When the loss is twice-continuously differentiable, we have

$$\begin{aligned}
\exists\alpha\in[0,1], \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}'\right) =&\mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right) + \left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)|_{\bar{\boldsymbol{\theta}}}\right)\left(\bar{\boldsymbol{\theta}}'-\bar{\boldsymbol{\theta}}\right)\\
&+\frac{1}{2}(\bar{\boldsymbol{\theta}}'-\bar{\boldsymbol{\theta}})^\top\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)\Big|_{\bar{\boldsymbol{\theta}}+\alpha(\bar{\boldsymbol{\theta}}'-\bar{\boldsymbol{\theta}})}\right)(\bar{\boldsymbol{\theta}}'-\bar{\boldsymbol{\theta}}).
\end{aligned} \tag{203}$$

We further assume M-Lipschitz condition on the Hessian:

$$\forall\bar{\boldsymbol{\theta}},\forall\bar{\boldsymbol{\theta}}', \frac{\left\|\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)|_{\bar{\boldsymbol{\theta}}'}\right)-\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)|_{\bar{\boldsymbol{\theta}}}\right)\right\|_2}{\left\|\bar{\boldsymbol{\theta}}'-\bar{\boldsymbol{\theta}}\right\|_2}\le M. \tag{204}$$

Combining Eq. (203) and Eq. (204), we have

$$\begin{aligned}\mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}'\right) \leq & \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}\right) + \left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)|_{\bar{\boldsymbol{\theta}}}\right)\left(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}}\right) \\ & + \frac{1}{2}(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}})^\top \left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)\big|_{\bar{\boldsymbol{\theta}}}\right)\left(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}}\right) + \frac{1}{2}M\|\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}}\|_2^3. \end{aligned} \tag{205}$$

Thus,

$$\begin{aligned}\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq & \mathop{\mathbb{E}}_{\bar{\boldsymbol{\theta}}' \sim \mathbb{Q}\left(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}\right)}\left[\mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}'\right)\right] + \mathop{\mathbb{E}}_{\bar{\boldsymbol{\theta}}' \sim \mathbb{Q}\left(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}\right)}\left[\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)|_{\bar{\boldsymbol{\theta}}}\right)\left(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}}\right)\right] \\ & + \frac{1}{2}\mathop{\mathbb{E}}_{\bar{\boldsymbol{\theta}}' \sim \mathbb{Q}\left(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}\right)}\left[(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}})^\top \left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)\big|_{\bar{\boldsymbol{\theta}}}\right)\left(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}}\right)\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\bar{\boldsymbol{\theta}}' \sim \mathbb{Q}\left(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}\right)}\left[M\|\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}}\|_2^3\right]. \end{aligned} \tag{206}$$

When the posterior is a Guassian $\mathcal{N}\left(\boldsymbol{\theta}\middle|\bar{\boldsymbol{\theta}}, \sigma^2 I\right)$,

$$\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}'\right) + \frac{1}{2}\mathrm{Tr}(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)\big|_{\bar{\boldsymbol{\theta}}})\sigma^2 + M\frac{\sqrt{2}\Gamma(\frac{n+3}{2})}{\Gamma(\frac{n}{2})}\sigma^3. \tag{207}$$

Next, we extend the above discussion to more general Gaussian, i.e., $\mathcal{N}\left(\boldsymbol{\theta}\middle|\bar{\boldsymbol{\theta}}, \Sigma\right)$, where Σ is a positive-definite diagonal matrix. We define a noramlized M-Hessian Lipschitz condition with a covariance matrix Σ as follows.

$$\forall\bar{\boldsymbol{\theta}}, \forall\bar{\boldsymbol{\theta}}', \frac{\left\|\Sigma^{\frac{1}{2}}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}'}\right) - \left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}|_{\bar{\boldsymbol{\theta}}}\right)\right)\Sigma^{\frac{1}{2}}\right\|_2}{\sqrt{(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}})^\top\Sigma^{-1}(\bar{\boldsymbol{\theta}}' - \bar{\boldsymbol{\theta}})}} \leq M. \tag{208}$$

Using the same argument, we recover a similar bound with Eq. (207).

$$\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\bar{\boldsymbol{\theta}}'\right) + \frac{1}{2}\mathrm{Tr}\left(\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)\big|_{\bar{\boldsymbol{\theta}}}\right)\Sigma\right) + M\frac{\sqrt{2}\Gamma(\frac{n+3}{2})}{\Gamma(\frac{n}{2})}\sigma^3. \tag{209}$$

I. Rate of convergence concerning m

The NS terms in Prop. 4.1, Prop. 4.2, Prop. 6.1, and Theorem 6.1 are $O(\lambda^{-\frac{1}{2}})$, which are $O(m^{-\frac{1}{4}})$ with the optimal choice of λ. This convergence rate is apparently bad compared to usual $O(m^{-\frac{1}{2}})$ convergence rate. However, by examining exisitng bounds, we can confirm that the convergence rate of our bounds are not looser than existing bounds.

First, we review Eq. (4).

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) + \frac{1}{\lambda}\mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right) \parallel \mathbb{P}\left(\boldsymbol{\theta}\right)] + \frac{\lambda}{2m} + \frac{1}{\lambda}\ln\frac{1}{\delta}. \tag{210}$$

To achieve a bound smaller than $c \in \mathbb{R}_{>0}$, we will set $\lambda < cm$. Thus, apparent rate of convergence is $O(m^{-1})$. However, if we tune λ and use the optimal value, the inequality changes as follows.

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) + \sqrt{\frac{2\mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right) \parallel \mathbb{P}\left(\boldsymbol{\theta}\right)] + 2\ln\frac{1}{\delta}}{m}}. \tag{211}$$

Now, the convergence rate turned out to be $O(m^{-\frac{1}{2}})$. We further decompose the bound as follows.

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \underbrace{\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) - \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)}_{\text{(C)}} + \underbrace{\sqrt{\frac{2\mathrm{KL}[\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right) \parallel \mathbb{P}\left(\boldsymbol{\theta}\right)] + 2\ln\frac{1}{\delta}}{m}}}_{\text{(D)}}. \tag{212}$$

Let the prior be a zero-mean Gaussian with a covariance matrix whose i-th diagonal elements are $\boldsymbol{\sigma}_i$, and the posterior be a Gaussian with a mean $\boldsymbol{\theta}$ and the same covariance matrix. Then,

$$\mathcal{L}_{\mathcal{D}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) \leq \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \underbrace{\mathcal{L}_{\mathcal{S}}\left(\mathbb{Q}\left(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}\right)\right) - \mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right)}_{\text{(C)}} + \underbrace{\sqrt{\frac{\sum_i \frac{\boldsymbol{\theta}_i^2}{\boldsymbol{\sigma}_i^2} + 2\ln\frac{1}{\delta}}{m}}}_{\text{(D)}}. \tag{213}$$

The second-order approximation of (C)-term yields the following.

$$\mathcal{L}_{\mathcal{S}}\left(\boldsymbol{\theta}\right) + \underbrace{\frac{1}{2}\sum_{i}\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{S}}\big|_{\bar{\boldsymbol{\theta}}}\right)_{i,i}\boldsymbol{\sigma}_i^2}_{\text{(C)}} + \underbrace{\sqrt{\frac{\sum_i \frac{\boldsymbol{\theta}_i^2}{\boldsymbol{\sigma}_i^2} + 2\ln\frac{1}{\delta}}{m}}}_{\text{(D)}}. \tag{214}$$

If we tune $\boldsymbol{\sigma}_i$ with some appropriate procedures, it tightens the bound. The point is that $\boldsymbol{\sigma}_i$s depend on m. This makes the rate of convergence concerning m different. Now, we return to Prop. 4.2. In the bound, we first tuned $\boldsymbol{\sigma}$ and then tune λ. Thus, this yields the same convergence rate with Eq.(214).

In Theorem 6.1 in (Neyshabur et al., 2018), we have a trade-off between the first and the second term through a margin parameter, which also depends on m. Even though we cannot make sure that the order of the bound exactly matches with ours, the apparent rate of their bound changes if we tune the parameter. The same applies to Bartlett et al. (2017) and Arora et al. (2018). Note that these two papers are based on Rademacher complexity and compression-based framework, respectively. This suggests that similar changes in convergence rate appear in other frameworks than PAC-Bayes.

J. Scale invariance of normalized sharpness

In this section, we briefly explain why normalized sharpness is invariant to parameter scale changes presented in Sec. 5. The definition of normalized sharpness (35) can be written as follows.

$$\sum_{l=1}^{d} \inf_{\bar{\boldsymbol{\gamma}}^{(l)}\in\mathbb{R}^{h_1^{(l)}},\bar{\boldsymbol{\gamma}}'^{(l)}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}\left[\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top \bar{F}^{(l)}\mathrm{e}\left[\bar{\boldsymbol{\gamma}}_j'^{(l)}\right] + \mathrm{e}\left[-\bar{\boldsymbol{\gamma}}^{(l)}\right]^\top \bar{U}^{(l)}\mathrm{e}\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]\right) \tag{215}$$

It suffices to show that the metric does not change when a row or column of l-th weight matrix is multiplied by α as described in Sec. 5. Assume that an i-th row of the l-th weight matrix is multiplied by $\exp(\alpha)$, and the successive matrix is multiplied by $\exp(-\alpha)$. It suffices to show the invariance of the following metrics.

$$\inf_{\bar{\boldsymbol{\gamma}}_i^{(l)}\in\mathbb{R},\bar{\boldsymbol{\gamma}}'^{(l)}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}\left[\bar{\boldsymbol{\gamma}}_i^{(l)}\right]^\top \bar{F}_{i,:}^{(l)}\mathrm{e}\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right] + \mathrm{e}\left[-\bar{\boldsymbol{\gamma}}_i^{(l)}\right]^\top \bar{U}_{i,:}^{(l)}\mathrm{e}\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]\right), \tag{216}$$

$$\inf_{\bar{\boldsymbol{\gamma}}^{(l+1)}\in\mathbb{R}^{h_1^{(l+1)}},\bar{\boldsymbol{\gamma}}_i'^{(l+1)}\in\mathbb{R}} \left(\mathrm{e}\left[\bar{\boldsymbol{\gamma}}^{(l+1)}\right]^\top \bar{F}_{:,i}^{(l+1)}\mathrm{e}\left[\bar{\boldsymbol{\gamma}}_i'^{(l+1)}\right] + \mathrm{e}\left[-\bar{\boldsymbol{\gamma}}^{(l+1)}\right]^\top \bar{U}_{:,i}^{(l+1)}\mathrm{e}\left[-\bar{\boldsymbol{\gamma}}_i'^{(l+1)}\right]\right). \tag{217}$$

We show the invariance of (216). Applying the same discussion to the other is straightforward. By the transformation, the i-th row of the Hessian with respect to $W^{(l)}$, or $\bar{F}^{(l)}$, is scaled by $\exp(-2\alpha)$. Thus, the metric changes as follows.

$$\inf_{\bar{\boldsymbol{\gamma}}_i^{(l)}\in\mathbb{R},\bar{\boldsymbol{\gamma}}'^{(l)}\in\mathbb{R}^{h_2^{(l)}}} \left(e^{-2\alpha}\mathrm{e}\left[\bar{\boldsymbol{\gamma}}_i^{(l)}\right]^\top \bar{F}_{i,:}^{(l)}\mathrm{e}\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right] + e^{2\alpha}\mathrm{e}\left[-\bar{\boldsymbol{\gamma}}_i^{(l)}\right]^\top \bar{U}_{i,:}^{(l)}\mathrm{e}\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]\right) \tag{218}$$

$$= \inf_{(\bar{\boldsymbol{\gamma}}_i^{(l)}-2\alpha)\in\mathbb{R},\bar{\boldsymbol{\gamma}}'^{(l)}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}\left[\bar{\boldsymbol{\gamma}}_i^{(l)}-2\alpha\right]^\top \bar{F}_{i,:}^{(l)}\mathrm{e}\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right] + \mathrm{e}\left[-(\bar{\boldsymbol{\gamma}}_i^{(l)}-2\alpha)\right]^\top \bar{U}_{i,:}^{(l)}\mathrm{e}\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]\right) \tag{219}$$

$$= \inf_{\bar{\boldsymbol{\gamma}}_i^{(l)}\in\mathbb{R},\bar{\boldsymbol{\gamma}}'^{(l)}\in\mathbb{R}^{h_2^{(l)}}} \left(\mathrm{e}\left[\bar{\boldsymbol{\gamma}}_i^{(l)}\right]^\top \bar{F}_{i,:}^{(l)}\mathrm{e}\left[\bar{\boldsymbol{\gamma}}'^{(l)}\right] + \mathrm{e}\left[-\bar{\boldsymbol{\gamma}}_i^{(l)}\right]^\top \bar{U}_{i,:}^{(l)}\mathrm{e}\left[-\bar{\boldsymbol{\gamma}}'^{(l)}\right]\right) \tag{220}$$

K. Comparison with existing normalized curvatures

In this section, we discuss the connections between normalized sharpness and prior normalized curvature studies such as Fisher-Rao norm (Liang et al., 2019), Gauss-Newton norm (Zhang et al., 2019), information in wegits (Achille & Soatto, 2018), and expected sharpness (Neyshabur et al., 2017).

Advantage of Fisher-Rao norm and Gauss-Newton norm: In some simple architectures, the Fisher-Rao norm and Gauss-Newton norm have connections with input-output Jacobians, parameter-output Jacobians, and function space norms (Liang et al., 2019; Zhang et al., 2019). However, normalized sharpness does not have the interpretations. This is a clear advantage of Fisher-Rao norm and Gauss-Newton norm. Note that the interpretations are lost when networks have skip connections or branches, which are common in modern networks. In our analysis, we did not rely on special structures of network architectures. Thus, normalized sharpness does not lose its connection to generalization even when they exist.

Additional invariance in normalized sharpness: Normalized sharpness has additional invariance compared to current normalized curvature metrics. Let us consider the following network.

$$f(z) = \mathrm{ReLU}(W^{(1)}z) + \mathrm{ReLU}(W^{(2)}z). \tag{221}$$

This type of connection is often found in modern networks (Xie et al., 2017; Zoph et al., 2018). We assume that the following condition is satisfied.

$$W^{(1)} = W^{(2)} = W/2. \tag{222}$$

Next, we rescale the parameters as follows.

$$f(z) = \mathrm{ReLU}(W'^{(1)}z) + \mathrm{ReLU}(W'^{(2)}z), \tag{223}$$

$$W'^{(1)} = W, \quad W'^{(2)} = \boldsymbol{O}. \tag{224}$$

By this rescaling, the existing normalized curvature metrics become double, while normalized sharpness (35) is kept the same. Note, the considered network does not satisfy the assumpsions of the invariance theorem in Liang et al. (2019). This additional invariance suggests that our definition better captures generalization in some cases.

L. Further comparison with existing work

Comparison with Wang et al. (2018): Our paper explores scale-independent generalization error bounds and flatness definition, while Wang et al. (2018) explored to limit the perturbation in a small region to make the quadratic-approximation holds better. Thus, the two papers make orthogonal contributions. We also point out that the generalization error bounds presented by Wang et al. (2018) depend on the sum of the log of parameter scales. It is much larger than the log of the sum of parameter scales and often not negligible. Additionally, the scale-direction decomposition in Sec. 4 is a crucial step to removing parameter-count-dependence in Sec. 6, and the technique in Sec. 6 is not directly applicable to Wang et al. (2018).

M. Experimental setups

We used cross-entropy loss during the training. We used a normalized-softmax-cross-entropy loss (201) to calculate normalized sharpness and the trace of the Hessian. We trained all models using Adam optimizer with its default parameters ($\mathrm{lr} = 0.001$) for 200 epochs with batchsize 128.

Experiment 8.1: We set the ratio of random labels from 0 to 1 at 0.1 intervals and trained models six times for each ratio. We used the Adam optimizer and did not apply regularization or data augmentation so that the training accuracy would reach near 1.0 even with random labels.

Experiment 8.2: We chose the hyperparameter of L2 regularization from $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$, the hyperparameter of weight decay from $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$, and the hyperparameter of dropout from $[0.1, 0.2, 0.3, 0.4, 0.5]$. We did not use more than one of the regularization techniques at the same time. We trained models for five times for each hyperparameter setting, and then removed models with lower training accuracy than 0.5. We used the Adam optimizer and applied no other regularizations.

Computing resource: We used Intel Xeon E5-2695 v4 and NVIDIA Tesla P100 (Pascal) for all experiments.

In the calculation of the trace of the Hessian, Frobenius norms, and normalized sharpness, we excluded the bias terms and scaling parameters of normalization layers from the calculation. This can be justified by applying union bound over numbers representable by floating point numbers. This increase terms Ξ_2 in Prop 6.1 and Ξ_3 in Theorem 6.1 by $O(h)$, which does not conflict with our discussion.

N. Experimental results

N.1. Scatter plots of experiment 8.2

Figure 3 shows the scatter plots of normalized sharpness, the trace of the Hessian, and the squared Frobenius norm of weight matrices v.s. the accuracy gap between train and test data.

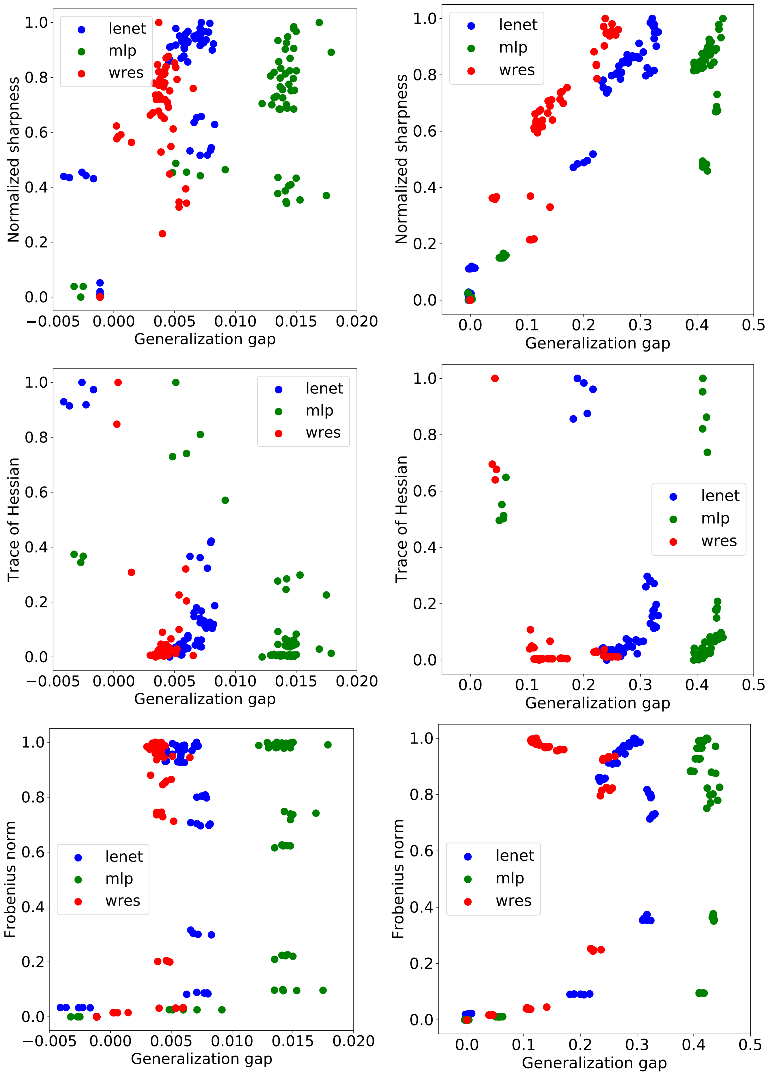

Figure 3. Scatter plot between normalized sharpness (35), trace of the Hessian (6), the sum of the squared Frobenius norms of weight matrices, and the accuracy gap. The left column is the results on MNIST and the right is on CIFAR10. All generalization metrics were rescaled to $[0, 1]$ by their maximum and minimum per network architecture. Thus, direct comparisons between architectures are meaningless.