A. Proofs

A.1. Proof of Theorem 2

We first introduce the following concentration inequality.

Lemma 1. *Let $\{\epsilon_i, i = 1, \dots, k\}$ be independent normal variables with mean 0 and variance σ_i^2. Then,*

$$P\left(\sum_{i=1}^{k} \epsilon_i^2 \geq k\sigma_{\max}^2 \left(1 + \sqrt{\frac{\log n}{k}}\right)^2\right) \leq \frac{1}{\sqrt{n}}$$

where $\sigma_{\max} = \max\{\sigma_i\}$.

Proof. From Lemma 1 in Laurent & Massart (2000), for any $x > 0$,

$$P\left(\sum_{i=1}^{k} \epsilon_i^2 \geq \sum_{i=1}^{k} \sigma_i^2 + 2\sqrt{\sum_{i=1}^{k} \sigma_i^4 x} + 2\sigma_{\max}^2 x\right) \leq \exp(-x).$$

Since

$$\begin{aligned}\sum_{i=1}^{k} \sigma_i^2 + 2\sqrt{\sum_{i=1}^{k} \sigma_i^4 x} + 2\sigma_{\max}^2 x &\leq \sigma_{\max}^2 (k + 2\sqrt{kx} + 2x) \\ &\leq \sigma_{\max}^2 (\sqrt{k} + \sqrt{2x})^2,\end{aligned}$$

plugging $x = \frac{1}{2}\log n$ proves the lemma.

□

Theorem 3. *Let $T_{\mathbf{w}}^{-1}$ be a normalization operator of $\mathbf{w}$ on $\mathbb{R}^k$. If $L_D(\mathbf{w}) \leq E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_D(\mathbf{w} + \boldsymbol{\epsilon})]$ for some $\sigma > 0$, then with probability $1 - \delta$,*

$$L_D(\mathbf{w}) \leq \max_{\|T_{\mathbf{w}}^{-1}\epsilon\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \sqrt{\frac{1}{n-1}\left(k \log\left(1 + \frac{\|\mathbf{w}\|_2^2}{\eta^2\rho^2}\left(1 + \sqrt{\frac{\log n}{k}}\right)^2\right) + 4\log\frac{n}{\delta} + O(1)\right)}$$

where $n = |S|$ and $\rho = \sqrt{k}\sigma(1 + \sqrt{\log n/k})/\eta$.

Proof. The idea of the proof is given in Foret et al. (2021). From the assumption, adding Gaussian perturbation on the weight space does not improve the test error. Moreover, from Theorem 3.2 in Chatterji et al. (2019), the following generalization bound holds under the perturbation:

$$E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_D(\mathbf{w} + \boldsymbol{\epsilon})] \leq E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_S(\mathbf{w} + \boldsymbol{\epsilon})] + \sqrt{\frac{1}{n-1}\left(\frac{1}{4}k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + \log\frac{n}{\delta} + C(n,\sigma,k)\right)}.$$

Therefore, the left hand side of the statement can be bounded as

$$\begin{aligned}L_D(\mathbf{w}) &\leq E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_S(\mathbf{w} + \boldsymbol{\epsilon})] + \sqrt{\frac{1}{n-1}\left(\frac{1}{4}k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + \log\frac{n}{\delta} + C\right)} \\ &\leq \left(1 - \frac{1}{\sqrt{n}}\right) \max_{\|T_{\mathbf{w}}^{-1}\epsilon\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \frac{1}{\sqrt{n}} + \sqrt{\frac{1}{n-1}\left(\frac{1}{4}k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + \log\frac{n}{\delta} + C\right)} \\ &\leq \max_{\|T_{\mathbf{w}}^{-1}\epsilon\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \sqrt{\frac{1}{n-1}\left(k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + 4\log\frac{n}{\delta} + 4C\right)}\end{aligned}$$

where the second inequality follows from Lemma 1 and $\|T_{\mathbf{w}}^{-1}\|_2 \leq \frac{1}{\eta}$.

□

B. Correlation Analysis

To capture the correlation between generalization measures, i.e., sharpness and adaptive sharpness, and actual generalization gap, we utilize Kendall rank correlation coefficient (Kendall, 1938). Formally, given the set of pairs of a measure and generalization gap observed $S = \{(m_1, g_1), \dots, (m_n, g_n)\}$, Kendall rank correlation coefficient τ is given by

$$\tau(S) = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(m_i - m_j)\text{sign}(g_i - g_j).$$

Since τ represents the difference between the proportion of concordant pairs, i.e., either both $m_i < m_j$ and $g_i < g_j$ or both $m_i > m_j$ and $g_i > g_j$ among the whole $\binom{n}{2}$ point pairs, and the proportion of discordant pairs, i.e., not concordant, the value of τ is in the range of $[-1, 1]$.

While the rank correlation coefficient aggregates the effects of all the hyper-parameters, granulated coefficient (Jiang et al., 2019) can consider the correlation with respect to the each hyper-parameter separately. If $\Theta = \prod_{i=1}^{N} \Theta_i$ is the Cartesian product of each hyper-parameter space Θ_i, granulated coefficient with respect to Θ_i is given by

$$\psi_i = \frac{1}{|\Theta_{-i}|} \sum_{\theta_1 \in \Theta_1} \cdots \sum_{\theta_{i-1} \in \Theta_{i-1}} \sum_{\theta_{i+1} \in \Theta_{i+1}} \cdots \sum_{\theta_N \in \Theta_N} \tau \left(\bigcup_{\theta_i \in \Theta_i} \{(m(\boldsymbol{\theta}), g(\boldsymbol{\theta})\} \right)$$

where $\Theta_{-i} = \Theta_1 \times \cdots \Theta_{i-1} \times \Theta_{i+1} \times \Theta_N$. Then the average $\Psi = \sum_{i=1}^{N} \psi_i / N$ of ψ_i indicates whether the correlation exists across all hyper-parameters.

We vary 4 hyper-parameters, mini-batch size, initial learning rate, weight decay coefficient and dropout rate, to produce different models. It is worth mentioning that changing one or two hyper-parameters for correlation analysis may cause spurious correlation (Jiang et al., 2019). For each hyper-parameter, we use 5 different values in Table 7 which implies that $5^4 = 625$ configurations in total.

Table 7. Hyper-parameter configurations.

mini-batch size	32, 64, 128, 256, 512
learning rate	0.0033, 0.01, 0.033, 0.1, 0.33
weight decay	5e−7, 5e−6, 5e−5, 5e−4, 5e−3
dropout rate	0, 0.125, 0.25, 0.375, 0.5

By using the above hyper-parameter configurations, we train WideResNet-28-2 model on CIFAR-10 dataset. We use SGD as an optimizer and set momentum to 0.9. We set the number of epochs to 200 and cosine learning rate decay (Loshchilov & Hutter, 2016) is adopted. Also, random resize, padding by four pixels, normalization and random horizontal flip are applied for data augmentation and label smoothing (Müller et al., 2019) is adopted with its factor of 0.1. Using model parameters with training accuracy higher than 99.0% among the generated models, we calculate sharpness and adaptive sharpness with respect to generalization gap.

To calculate adaptive sharpness, we fix normalization scheme to element-wise normalization. We calculate adaptive sharpness and sharpness with both $p = 2$ and $p = \infty$. We conduct a grid search over $\{5\text{e}{-6}, 1\text{e}{-5}, 5\text{e}{-5}, \dots, 5\text{e}{-1}, 1.0\}$ to obtain each ρ for sharpness and adaptive sharpness which maximizes correlation with generalization gap. As results of the grid search, we select $1\text{e}{-5}$ and $5\text{e}{-4}$ as ρs for sharpness of $p = 2$ and $p = \infty$, respectively, and select $5\text{e}{-1}$ and $5\text{e}{-3}$ as ρs for adaptive sharpness of $p = 2$ and $p = \infty$, respectively. To calculate maximizers of each loss function for calculation of sharpness and adaptive sharpness, we follow m-sharpness strategy suggested by Foret et al. (2021) and m is set to 8.