A Proof of Theorems

The proof of the key lemma (Lemma 5), which establishes a connection between the margin operator and the robust margin operator, is presented in the main content.

We still need to demonstrate that the properties in PAC-Bayes analysis hold for both the margin operator and the robust margin operator. The following proofs are adapted from the work of (Neyshabur et al., 2017b), with the steps being kept independent of the (robust) margin operator. We will begin by finishing the proofs of Lemma 6 and Lemma 7. Afterward, we will proceed to complete the proof of Theorem 1, which is our primary result.

A.1 Proof of Lemma 6

Proof of Lemma 6.1:

For any $i \in [k]$,

$$|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x})[i] - f_{\mathbf{w}}(\mathbf{x})[i]| \leq \|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})\|_2.$$

For any $i, j \in [k]$,

$$|M(f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}), i, j) - M(f_{\mathbf{w}}(\mathbf{x}), i, j)| \leq 2|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x})[i] - f_{\mathbf{w}}(\mathbf{x})[i]| \leq 2\|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})\|_2.$$

Therefore, it is left to bound $\|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})\|$. It is provided in (Neyshabur et al., 2017b), we provide the proof here for reference. Let $\Delta_i = \left|f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x})\right|_2$. We will prove using induction that for any $i \geq 0$:

$$\Delta_i \leq \left(1 + \frac{1}{d}\right)^i \left(\prod_{j=1}^{i} \|W_j\|_2\right) |\mathbf{x}|_2 \sum_{j=1}^{i} \frac{\|U_j\|_2}{\|W_j\|_2}.$$

The above inequality together with $\left(1 + \frac{1}{d}\right)^d \leq e$ proves the lemma statement. The induction base clearly holds since $\Delta_0 = |\mathbf{x} - \mathbf{x}|_2 = 0$. For any $i \geq 1$, we have the following:

$$\begin{aligned}
\Delta_{i+1} &= \left|(W_{i+1} + U_{i+1})\, \phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - W_{i+1}\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&= \left|(W_{i+1} + U_{i+1})\left(\phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right) + U_{i+1}\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&\leq (\|W_{i+1}\|_2 + \|U_{i+1}\|_2)\left|\phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 + \|U_{i+1}\|_2\left|\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&\leq (\|W_{i+1}\|_2 + \|U_{i+1}\|_2)\left|f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x})\right|_2 + \|U_{i+1}\|_2\left|f^i_{\mathbf{w}}(\mathbf{x})\right|_2 \\
&= \Delta_i\left(\|W_{i+1}\|_2 + \|U_{i+1}\|_2\right) + \|U_{i+1}\|_2\left|f^i_{\mathbf{w}}(\mathbf{x})\right|_2,
\end{aligned}$$

where the last inequality is by the Lipschitz property of the activation function and using $\phi(0) = 0$. The ℓ_2 norm of outputs of layer i is bounded by $|\mathbf{x}|_2 \Pi_{j=1}^{i} \|W_j\|_2$ and by the lemma assumption we have $\|U_{i+1}\|_2 \leq \frac{1}{d}\|W_{i+1}\|_2$. Therefore, using the induction step, we get the following bound:

$$\begin{aligned}
\Delta_{i+1} &\leq \Delta_i\left(1 + \frac{1}{d}\right)\|W_{i+1}\|_2 + \|U_{i+1}\|_2 |\mathbf{x}|_2 \prod_{j=1}^{i} \|W_j\|_2 \\
&\leq \left(1 + \frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\|W_j\|_2\right)|\mathbf{x}|_2 \sum_{j=1}^{i}\frac{\|U_j\|_2}{\|W_j\|_2} + \frac{\|U_{i+1}\|_2}{\|W_{i+1}\|_2}|\mathbf{x}|_2 \prod_{j=1}^{i+1}\|W_i\|_2 \\
&\leq \left(1 + \frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\|W_j\|_2\right)|\mathbf{x}|_2 \sum_{j=1}^{i+1}\frac{\|U_j\|_2}{\|W_j\|_2}.
\end{aligned}$$

Then we complete the proof of Lemma 6.1. By combining Lemma 6.1 and Lemma 5, we directly obtain Lemma 6.2.

□

A.2 Proof of Lemma 7

The proof of Lemma 7.1 and 7.2 is similar. We provide the proof of Lemma 7.2 below. The proof of Lemma 7.1 follows the proof of Lemma 7.2 by replacing the robust margin operator by the margin operator.

Let $\mathbf{w}' = \mathbf{w} + \mathbf{u}$. Let $\mathcal{S}_{\mathbf{w}}$ be the set of perturbations with the following property:

$$\mathcal{S}_{\mathbf{w}} \subseteq \left\{ \mathbf{w}' \;\middle|\; \max_{i,j\in[k],\mathbf{x}\in\mathcal{X}} |RM(f_{\mathbf{w}'}(\mathbf{x}), i, j) - RM(f_{\mathbf{w}}(\mathbf{x}), i, j)| < \frac{\gamma}{2} \right\}.$$

Let q be the probability density function over the parameters $\mathbf{w}'$. We construct a new distribution $\tilde{Q}$ over predictors $f_{\tilde{\mathbf{w}}}$ where $\tilde{\mathbf{w}}$ is restricted to $\mathcal{S}_{\mathbf{w}}$ with the probability density function:

$$\tilde{q}(\tilde{\mathbf{w}}) = \frac{1}{Z} \begin{cases} q(\tilde{\mathbf{w}}) & \tilde{\mathbf{w}} \in \mathcal{S}_{\mathbf{w}} \\ 0 & \text{otherwise.} \end{cases}$$

Here Z is a normalizing constant and by the lemma assumption $Z = \mathbb{P}\left[\mathbf{w}' \in \mathcal{S}_{\mathbf{w}}\right] \geq \frac{1}{2}$. By the definition of $\tilde{Q}$, we have:

$$\max_{i,j\in[k],\mathbf{x}\in\mathcal{X}} |RM(f_{\tilde{\mathbf{w}}}(\mathbf{x}), i, j) - RM(f_{\mathbf{w}}(\mathbf{x}), i, j)| < \frac{\gamma}{2}.$$

Since the above bound holds for any $\mathbf{x}$ in the domain $\mathcal{X}$, we can get the following a.s.:

$$R_0(f_{\mathbf{w}}) \leq R_{\frac{\gamma}{2}}(f_{\tilde{\mathbf{w}}})$$
$$\hat{R}_{\frac{\gamma}{2}}(f_{\tilde{\mathbf{w}}}) \leq \hat{R}_{\gamma}(f_{\mathbf{w}})$$

Now using the above inequalities together with the equation (5), with probability $1-\delta$ over the training set we have:

$$\begin{aligned} R_0(f_{\mathbf{w}}) &\leq \mathbb{E}_{\tilde{\mathbf{w}}}\left[R_{\frac{\gamma}{2}}(f_{\tilde{\mathbf{w}}})\right] \\ &\leq \mathbb{E}_{\tilde{\mathbf{w}}}\left[\hat{R}_{\frac{\gamma}{2}}(f_{\tilde{\mathbf{w}}})\right] + 2\sqrt{\frac{2(KL\left(\tilde{\mathbf{w}}\|P\right) + \ln\frac{2m}{\delta})}{m-1}} \\ &\leq \hat{R}_{\gamma}(f_{\mathbf{w}}) + 2\sqrt{\frac{2(KL\left(\tilde{\mathbf{w}}\|P\right) + \ln\frac{2m}{\delta})}{m-1}} \\ &\leq \hat{R}_{\gamma}(f_{\mathbf{w}}) + 4\sqrt{\frac{KL\left(\mathbf{w}'\|P\right) + \ln\frac{6m}{\delta}}{m-1}}, \end{aligned}$$

The last inequality follows from the following calculation.

Let $\mathcal{S}_{\mathbf{w}}^c$ denote the complement set of $\mathcal{S}_{\mathbf{w}}$ and $\tilde{q}^c$ denote the density function q restricted to $\mathcal{S}_{\mathbf{w}}^c$ and normalized. Then,

$$KL(q||p) = ZKL(\tilde{q}||p) + (1-Z)KL(\tilde{q}^c||p) - H(Z),$$

where $H(Z) = -Z\ln Z - (1-Z)\ln(1-Z) \leq 1$ is the binary entropy function. Since KL is always positive, we get,

$$KL(\tilde{q}||p) = \frac{1}{Z}\left[KL(q||p) + H(Z)) - (1-Z)KL(\tilde{q}^c||p)\right] \leq 2(KL(q||p) + 1).$$

A.3 Proof of Theorem 1

Given the local perturbation bound of the robust margin operator and Lemma 5, the proof of Theorem 1 follows the procedure of the proof of Theorem 2.

Let $\beta = \left(\prod_{i=1}^{d} \|W_i\|_2\right)^{1/d}$ and consider a network with the normalized weights $\widetilde{W}_i = \frac{\beta}{\|W_i\|_2} W_i$. Due to the homogeneity of the ReLU, we have that for feedforward networks with ReLU activations

$f_{\tilde{\mathbf{w}}} = f_{\mathbf{w}}$, and so the (empirical and expected) loss (including margin loss) is the same for $\mathbf{w}$ and $\widetilde{\mathbf{w}}$. We can also verify that $\left(\prod_{i=1}^{d} \|W_i\|_2\right) = \left(\prod_{i=1}^{d} \left\|\widetilde{W}_i\right\|_2\right)$ and $\frac{\|W_i\|_F}{\|W_i\|_2} = \frac{\left\|\tilde{W}_i\right\|_F}{\left\|\tilde{W}_i\right\|_2}$, and so the excess error in the Theorem statement is also invariant to this transformation. It is therefore sufficient to prove the Theorem only for the normalized weights $\tilde{\mathbf{w}}$, and hence we assume w.l.o.g. that the spectral norm is equal across layers, i.e. for any layer i, $\|W_i\|_2 = \beta$.

Choose the distribution of the prior P to be $\mathcal{N}(0, \sigma^2 I)$, and consider the random perturbation $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 I)$, with the same σ, which we will set later according to β. More precisely, since the prior cannot depend on the learned predictor $\mathbf{w}$ or its norm, we will set σ based on an approximation $\tilde{\beta}$. For each value of $\tilde{\beta}$ on a pre-determined grid, we will compute the PAC-Bayes bound, establishing the generalization guarantee for all $\mathbf{w}$ for which $|\beta - \tilde{\beta}| \leq \frac{1}{d}\beta$, and ensuring that each relevant value of β is covered by some $\tilde{\beta}$ on the grid. We will then take a union bound over all $\tilde{\beta}$ on the grid. For now, we will consider a fixed $\tilde{\beta}$ and the $\mathbf{w}$ for which $|\beta - \tilde{\beta}| \leq \frac{1}{d}\beta$, and hence $\frac{1}{e}\beta^{d-1} \leq \tilde{\beta}^{d-1} \leq e\beta^{d-1}$.

Since $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 I)$, we get the following bound for the spectral norm of U_i (Tropp, 2012):

$$\mathbb{P}_{U_i \sim N(0,\sigma^2 I)}\left[\|U_i\|_2 > t\right] \leq 2he^{-t^2/2h\sigma^2}.$$

Taking a union bond over the layers, we get that, with probability $\geq \frac{1}{2}$, the spectral norm of the perturbation U_i in each layer is bounded by $\sigma\sqrt{2h\ln(4dh)}$. Plugging this spectral norm bound into the Lipschitz of robust margin operator we have that with probability at least $\frac{1}{2}$,

$$\max_{i,j\in[k],\mathbf{x}\in\mathcal{X}} |RM(f_{\mathbf{w}'}(\mathbf{x}), i, j) - RM(f_{\mathbf{w}}(\mathbf{x}), i, j)| \tag{10}$$

$$\leq 2e(B+\epsilon)\beta^d \sum_i \frac{\|U_i\|_2}{\beta}$$

$$= e(B+\epsilon)\beta^{d-1} \sum_i \|U_i\|_2 \leq e^2 d(B+\epsilon)\tilde{\beta}^{d-1}\sigma\sqrt{2h\ln(4dh)} \leq \frac{\gamma}{2}, \tag{11}$$

where we choose $\sigma = \frac{\gamma}{42d(B+\epsilon)\tilde{\beta}^{d-1}\sqrt{h\ln(4hd)}}$ to get the last inequality, the first inequality is Lemma 6.2. The second inequality is the tail bound above. Hence, the perturbation $\mathbf{u}$ with the above value of σ satisfies the assumptions of the Lemma 4.

We now calculate the KL-term in Lemma 4 with the chosen distributions for P and $\mathbf{u}$, for the above value of σ.

$$\begin{aligned}
&KL(\mathbf{w}+\mathbf{u}||P) \\
\leq &\frac{|\mathbf{w}|^2}{2\sigma^2} = \frac{42^2 d^2 (B+\epsilon)^2 \tilde{\beta}^{2d-2} h\ln(4hd)}{2\gamma^2} \sum_{i=1}^{d} \|W_i\|_F^2 \\
\leq &\mathcal{O}\left((B+\epsilon)^2 d^2 h\ln(dh)\frac{\beta^{2d}}{\gamma^2}\sum_{i=1}^{d}\frac{\|W_i\|_F^2}{\beta^2}\right) \\
\leq &\mathcal{O}\left((B+\epsilon)^2 d^2 h\ln(dh)\frac{\Pi_{i=1}^{d}\|W_i\|_2^2}{\gamma^2}\sum_{i=1}^{d}\frac{\|W_i\|_F^2}{\|W_i\|_2^2}\right).
\end{aligned}$$

Hence, for any $\tilde{\beta}$, with probability $\geq 1-\delta$ and for all $\mathbf{w}$ such that, $|\beta - \tilde{\beta}| \leq \frac{1}{d}\beta$, we have:

$$R_0(f_{\mathbf{w}}) \leq \hat{R}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{(B+\epsilon)^2 d^2 h\ln(dh)\Pi_{i=1}^{d}\|W_i\|_2^2 \sum_{i=1}^{d}\frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \ln\frac{m}{\delta}}{\gamma^2 m}}\right). \tag{12}$$

For other ℓ_p attacks, the results are directly obtained by Lemma 4 of (Xiao et al., 2022a).

A.4 Proof of Theorem 8

It is based on a slight modification of the key lemma. if $g_{\mathbf{w}}(\mathbf{x})$ has a $(A_1|\mathbf{x}|, \cdots, A_d|\mathbf{x}|)$-local perturbation bound, *i.e.,*

$$|g_{\mathbf{w}}(\mathbf{x}) - g_{\mathbf{w}'}(\mathbf{x})| \leq \sum_{i=1}^{d} A_i|\mathbf{x}|\|W_i - W_i'\|,$$

the robustified function $\inf_{\mathbf{x}' \in C(\mathbf{x})} g_{\mathbf{w}}(\mathbf{x}')$ has a $(A_1 D, \cdots, A_d D)$-local perturbation bound.

Proof: Let

$$\mathbf{x}(\mathbf{w}) = \arg \inf_{\mathbf{x}' \in C(\mathbf{x})} g_{\mathbf{w}}(\mathbf{x}'),$$
$$\mathbf{x}(\mathbf{w}') = \arg \inf_{\mathbf{x}' \in C(\mathbf{x})} g_{\mathbf{w}'}(\mathbf{x}'),$$

Then,

$$|\inf_{\|\mathbf{x}-\mathbf{x}'\| \le \epsilon} g_{\mathbf{w}}(\mathbf{x}') - \inf_{\|\mathbf{x}-\mathbf{x}'\| \le \epsilon} g_{\mathbf{w}'}(\mathbf{x}')| \le$$
$$\max\{|g_{\mathbf{w}}(\mathbf{x}(\mathbf{w})) - g_{\mathbf{w}'}(\mathbf{x}(\mathbf{w}))|, |g_{\mathbf{w}}(\mathbf{x}(\mathbf{w}')) - g_{\mathbf{w}'}(\mathbf{x}(\mathbf{w}'))|\}.$$

It is because $g_{\mathbf{w}}(\mathbf{x}(\mathbf{w})) - g_{\mathbf{w}'}(\mathbf{x}(\mathbf{w}')) \le g_{\mathbf{w}}(\mathbf{x}(\mathbf{w}')) - g_{\mathbf{w}'}(\mathbf{x}(\mathbf{w}'))$ and $g_{\mathbf{w}'}(\mathbf{x}(\mathbf{w}')) - g_{\mathbf{w}}(\mathbf{x}(\mathbf{w})) \le g_{\mathbf{w}'}(\mathbf{x}(\mathbf{w})) - g_{\mathbf{w}}(\mathbf{x}(\mathbf{w}))$. Therefore,

$$|\inf_{\|\mathbf{x}-\mathbf{x}'\| \le \epsilon} g_{\mathbf{w}}(\mathbf{x}') - \inf_{\|\mathbf{x}-\mathbf{x}'\| \le \epsilon} g_{\mathbf{w}'}(\mathbf{x}')|$$
$$\le \sum_{i=1}^{d} A_i |\mathbf{x}(\mathbf{w})| \|W_i - W_i'\|$$
$$\le \sum_{i=1}^{d} A_i D \|W_i - W_i'\|.$$

Therefore, combining the local perturbation bound and Lemma 7.2, we complete the proof. □

A.5 Proof of Theorem 9

As shown in the proof of Lemma 6, it is left to bound $\|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})\|$. Let $\Delta_i = \left|f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x})\right|_2$. We will prove using induction that for any $i \ge 0$:

$$\Delta_i \le \left(1 + \frac{1}{d}\right)^i \left(\prod_{j=1}^{i} (\|W_j\|_2 + 1)\right) |\mathbf{x}|_2 \sum_{j=1}^{i} \frac{\|U_j\|_2}{(\|W_j\|_2 + 1)}.$$

The above inequality together with $\left(1 + \frac{1}{d}\right)^d \le e$ proves the lemma statement. The induction base clearly holds since $\Delta_0 = |\mathbf{x} - \mathbf{x}|_2 = 0$. For any $i \ge 1$, we have the following:

$$\begin{aligned}
\Delta_{i+1} &= \left|(W_{i+1} + U_{i+1})\, \phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - W_{i+1}\phi_i(f^i_{\mathbf{w}}(\mathbf{x})) + (f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&= \left|(W_{i+1} + U_{i+1})\left(\phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right) + U_{i+1}\phi_i(f^i_{\mathbf{w}}(\mathbf{x})) + (f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 \\
&\le (\|W_{i+1}\|_2 + \|U_{i+1}\|_2)\left|\phi_i(f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 + \|U_{i+1}\|_2 \left|\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))\right|_2 + \Delta_i \\
&\le (\|W_{i+1}\|_2 + \|U_{i+1}\|_2)\left|f^i_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f^i_{\mathbf{w}}(\mathbf{x})\right|_2 + \|U_{i+1}\|_2 \left|f^i_{\mathbf{w}}(\mathbf{x})\right|_2 + \Delta_i \\
&= \Delta_i \left(\|W_{i+1}\|_2 + \|U_{i+1}\|_2 + 1\right) + \|U_{i+1}\|_2 \left|f^i_{\mathbf{w}}(\mathbf{x})\right|_2,
\end{aligned}$$

where the last inequality is by the Lipschitz property of the activation function and using $\phi(0) = 0$. The ℓ_2 norm of outputs of layer i is bounded by $|\mathbf{x}|_2 \Pi_{j=1}^{i}(\|W_j\|_2 + 1)$ and by the lemma assumption we have $\|U_{i+1}\|_2 \le \frac{1}{d}\|W_{i+1}\|_2$. Therefore, using the induction step, we get the following bound:

$$\begin{aligned}
\Delta_{i+1} &\le \Delta_i \left(1 + \frac{1}{d}\right)(\|W_{i+1}\|_2 + 1) + \|U_{i+1}\|_2 |\mathbf{x}|_2 \prod_{j=1}^{i}(\|W_j\|_2 + 1) \\
&\le \left(1 + \frac{1}{d}\right)^{i+1} \left(\prod_{j=1}^{i+1}(\|W_j\|_2 + 1)\right) |\mathbf{x}|_2 \sum_{j=1}^{i} \frac{\|U_j\|_2}{(\|W_j\|_2 + 1)} + \frac{\|U_{i+1}\|_2}{(\|W_{i+1}\|_2 + 1)} |\mathbf{x}|_2 \prod_{j=1}^{i+1}(\|W_i\|_2 + 1) \\
&\le \left(1 + \frac{1}{d}\right)^{i+1} \left(\prod_{j=1}^{i+1}(\|W_j\|_2 + 1)\right) |\mathbf{x}|_2 \sum_{j=1}^{i+1} \frac{\|U_j\|_2}{(\|W_j\|_2 + 1)}.
\end{aligned}$$

Therefore, the margin operator of ResNet is locally $(A_1|\mathbf{x}|, \cdots, A_d|\mathbf{x}|)$-Lipschitz w.r.t. w, where

$$A_i = 2e \prod_{l=1}^{d} (\|W_l\|_2 + 1)/(\|W_i\|_2 + 1).$$

For any $\delta, \gamma > 0$, with probability $\geq 1 - \delta$ over a training set of size m, for any $\mathbf{w}$, we have:

$$\begin{aligned} &L_0(f_{\text{RN}}) - \hat{L}_\gamma(f_{\text{RN}}) \\ \leq &\mathcal{O}\left(\sqrt{\frac{B^2 d^2 h \ln(dh)\Phi(f_{\text{RN}}) + \ln\frac{dm}{\delta}}{\gamma^2 m}}\right); \end{aligned}$$

By a combination of Lemma 5 and Lemma 7, for any $\delta, \gamma > 0$, with probability $\geq 1 - \delta$ over a training set of size m, for any $\mathbf{w}$, we have:

$$\begin{aligned} &R_0(f_{\text{RN}}) - \hat{R}_\gamma(f_{\text{RN}}) \\ \leq &\mathcal{O}\left(\sqrt{\frac{(B+\epsilon)^2 d^2 h \ln(dh)\Phi(f_{\text{RN}}) + \ln\frac{dm}{\delta}}{\gamma^2 m}}\right), \end{aligned}$$

where $\Phi(f_{\text{RN}}) = \Pi_{i=1}^d (\|W_i\|_2 + 1)^2 \sum_{i=1}^d \frac{\|W_i\|_F^2}{(\|W_i\|_2+1)^2}$. □

B PAC-Bayesian Framework for Robust Generalization

PAC-Bayes analysis (McAllester, 1999) is a framework to provide generalization guarantees for randomized predictors drawn from a learned distribution Q (as opposed to a single predictor) that depends on the training data set. The expected generalization gap over the posterior distribution Q can be bounded in terms of the Kullback-Leibler divergence between the prior distribution P and the posterior distribution Q, $KL(P\|Q)$.

A direct corollary of Eq. (5) is that, the expected robust error of $f_{\mathbf{w}+\mathbf{u}}$ can be bounded as follows

$$\begin{aligned} &\mathbb{E}_{\mathbf{u}}[R_0^{adv}(f_{\mathbf{w}+\mathbf{u}})] \\ &\leq \mathbb{E}_{\mathbf{u}}[\hat{R}_0^{adv}(f_{\mathbf{w}+\mathbf{u}})] + 2\sqrt{\frac{2\left(KL\left(\mathbf{w}+\mathbf{u}\|P\right) + \ln\frac{2m}{\delta}\right)}{m-1}}. \end{aligned} \tag{13}$$

By a slight modification of Lemma 4, the following lemma given in the work of (Farnia et al., 2018) shows how to obtain an robust generalization bound.

Lemma 10 (Farnia et al. (2018))**.** *Let $f_{\mathbf{w}}(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^k$ be any predictor (not necessarily a neural network) with parameters $\mathbf{w}$, and P be any distribution on the parameters that is independent of the training data. Then, for any $\gamma, \delta > 0$, with probability $\geq 1-\delta$ over the training set of size m, for any $\mathbf{w}$, and any random perturbation $\mathbf{u}$ s.t. $\mathbb{P}_{\mathbf{u}}[\max_{\mathbf{x}\in\mathcal{X}} \left|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}+\delta^{adv}_{\mathbf{w}+\mathbf{u}}(\mathbf{x})) - f_{\mathbf{w}}(\mathbf{x}+\delta^{adv}_{\mathbf{w}}(\mathbf{x}))\right|_\infty < \frac{\gamma}{4}] \geq \frac{1}{2}$, we have:*

$$R_0^{adv}(f_{\mathbf{w}}) \leq \hat{R}_\gamma^{adv}(f_{\mathbf{w}}) + 4\sqrt{\frac{KL\left(\mathbf{w}+\mathbf{u}\|P\right) + \ln\frac{6m}{\delta}}{m-1}}.$$

Table 1: Comparison of the empirical results of the standard generalization bound and robust generalization in the experiment of training MNIST, CIFAR-10 and CIFAR-100 on VGG networks.

	MNIST	CIFAR-10	CIFAR-100
Standard Generalization Gap	1.13%	9.21%	23.61%
Bound in Theorem 2 (Neyshabur et al., 2017b)	1.33×10^4	1.34×10^9	3.41×10^{11}
Robust Generalization Gap	9.67%	51.41%	78.82%
Bound in Theorem 3 (Farnia et al., 2018)	NA	NA	NA
Bound in Theorem 1 (Ours)	3.23×10^4	5.97×10^{10}	1.66×10^{13}

C Empirical Study of the Generalization Bounds

The spectral complexity $\Phi(f_{\mathbf{w}})$ induced by adversarial training is significantly larger. We conducted experiments training MNIST, CIFAR-10, and CIFAR-100 datasets using VGG-19 networks, following

the training parameters described in (Neyshabur et al., 2017a).[4] The results are presented in Table 1. It is evident that adversarial training can induce a larger spectral complexity, resulting in a larger generalization bound.[5] We refer the readers to our previous work (Xiao et al., 2022a) for more experiments results about norm-based complexity of adversarially-trained models. These experiments align with the findings presented by (Bartlett et al., 2017), indicating: 1) spectral complexity scales with the difficulty of the learning task, and 2) the generalization bound is sensitive to this complexity.

[4]The settings of standard training follows the experiments in `https://github.com/bneyshabur/generalization-bounds`.

[5]The settings of adversarial training follows the experiments in `https://github.com/JiancongXiao/Adversarial-Rademacher-Complexity`.