

Proof Details

Anonymous authors
Paper under double-blind review

1 Reviewer RaW6S

1.1 W3: "tight" bound

Thanks for your raising the question which allows us to further clarify our work. Our claim of being tighter is based on a comparison with SCR. Directly comparing GMD with SCR is challenging due to their distinct optimization objectives and forms. Therefore, we considered conducting an indirect comparison between the SCR and SD.

For indirect comparison, we illustrate that our GMD is equivalent to a lower bound (lowest value) of SCR term. As a result, using our GMD to upper bound the robust generation error will results in a more tighter upper bound than using SCR. In specific, the lower bound of SCR term is expressed as:

$$\begin{aligned}
 SCR &= \min_{\theta} \frac{t}{N} \sum_{i=1}^K \sum_{v \in \hat{N}_i} \|f_{\theta}(x_v^{adv}) - f_{\theta}(x_v) - \mathbb{E}[f_{\theta}(x^{adv}) - f_{\theta}(x) | x \in C_i]\|_2^2 \\
 &\geq \min_{\theta} \frac{t}{N} \left\| \sum_{i=1}^K \sum_{v \in \hat{N}_i} (f_{\theta}(x_v^{adv}) - f_{\theta}(x_v)) - N \sum_{i=1}^K \mathbb{E}[f_{\theta}(x^{adv}) - f_{\theta}(x) | x \in C_i] \right\|_2^2 \\
 &\geq \min_{\theta} \frac{t}{N} \left| \left\| \sum_{i=1}^K \sum_{v \in \hat{N}_i} (f_{\theta}(x_v^{adv}) - f_{\theta}(x_v)) \right\|_2^2 - N \sum_{i=1}^K \|\mathbb{E}[f_{\theta}(x^{adv}) - f_{\theta}(x) | x \in C_i]\|_2^2 \right| \\
 &= \min_{\theta} \frac{t}{N} \left| \|f(X_D^{adv}) - f(X_D)\|_2^2 - \|f(X^{adv}) - f(X)\|_2^2 \right|. \tag{1}
 \end{aligned}$$

When $f(X^{adv}) \rightarrow f(X)$ and $f(X_D^{adv}) \rightarrow f(X_D)$, the resulting term,

$$\min_{\theta} \frac{t}{N} \left| \|f(X_D^{adv}) - f(X_D)\|_2^2 - \|f(X^{adv}) - f(X)\|_2^2 \right|,$$

provides a more constrained optimization compared to SCR. This optimization encourages not only intra-class consistency, as is the case with SCR, but also improves inter-class separation, which is essential for better generalization and robustness against adversarial attacks.

For our GMD term, $\min_{\theta} [\|\nabla T_d\|_2 + \|\nabla T\|_2] = \min_{\theta} [\| [f_{\theta}(X_D^{adv}) - f_{\theta}(X_D)] [f_{\theta}(X_D^{adv}) + f_{\theta}(X_D)] \|_2 + \| [f_{\theta}(X^{adv}) - f_{\theta}(X)] [f_{\theta}(X^{adv}) + f_{\theta}(X)] \|_2]$, it also encourages both $f(X_D^{adv}) \rightarrow f(X_D)$ and $f(X^{adv}) \rightarrow f(X)$. Based on the above analysis, the optimization of GMD is a tighter and more constrained approach compared to SCR.

Moreover, as demonstrated in the experimental results, our proposed SD achieves better robust accuracy than SCR in most experiments, which validates the superiority resulting from the proposed bound given by SD. Meanwhile, as observed in Section 6.6 of the main paper for generalization analysis, the robust models trained by our SD present smaller robust accuracy gaps than other methods, which further verifies the achieved better robust generalization.

Finally, to clarify this point, we have added the above discussion in the appendix of the revised version.

2 Reviewer GQUJ

2.1 L_d and L_u

Thank you for pointing out the error on p.24, and the issues with L_d and L_u . We have carefully considered your suggestion and have decided not to use the constants L_d and L_u . The main reason is that these constants are difficult to express in a concrete analytical form, which might confuse the theoretical derivation. To avoid this issue, we revised the proof and introduced a new proof approach that simplifies and clarifies the derivation. Under the revised proof, we believe the above two weaknesses can be solved.

Specifically, we made improvements to the original proof on p.25 and derived the following new form through derivation:

$$\begin{aligned}
& \varepsilon_{GE} + \frac{t}{N} \|f_\theta(\mathbf{X}_d^{adv}) - f_\theta(\mathbf{X}_d)\|_2 + Kt \|\mathbb{E}[f_\theta(\mathbf{X}_u^{adv})] - \mathbb{E}[f_\theta(\mathbf{X}_u)]\|_2 + M \cdot \sqrt{\frac{2K \ln 2 + 2 \ln(\frac{1}{\sigma})}{N}}, \\
& = \varepsilon_{GE} + \frac{t}{N} \left[\|f_\theta(\mathbf{X}_d^{adv}) - f_\theta(\mathbf{X}_d)\|_2 + \|[f_\theta(\mathbf{X}_d^{adv})]^2 - [f_\theta(\mathbf{X}_d)]^2\|_2 - \|[f_\theta(\mathbf{X}_d^{adv})]^2 - [f_\theta(\mathbf{X}_d)]^2\|_2 \right] \\
& + Kt \left[\|\mathbb{E}[f_\theta(\mathbf{X}_u^{adv})] - \mathbb{E}[f_\theta(\mathbf{X}_u)]\|_2 + \|\mathbb{E}[[f_\theta(\mathbf{X}_u^{adv})]^2] - \mathbb{E}[[f_\theta(\mathbf{X}_u)]^2]\|_2 - \|\mathbb{E}[[f_\theta(\mathbf{X}_u^{adv})]^2] - \mathbb{E}[[f_\theta(\mathbf{X}_u)]^2]\|_2 \right] \\
& + M \cdot \sqrt{\frac{2K \ln 2 + 2 \ln(\frac{1}{\sigma})}{N}}.
\end{aligned}$$

In the further derivation, we obtained the following form:

$$\begin{aligned}
& \leq \varepsilon_{GE} + \frac{t}{N} \left\| [f_\theta(\mathbf{X}_D^{adv})]^2 - [f_\theta(\mathbf{X}_D)]^2 \right\|_2 + Kt \left\| \mathbb{E}[f_\theta(\mathbf{X}_u^{adv})]^2 - \mathbb{E}[f_\theta(\mathbf{X}_u)]^2 \right\|_2 \\
& + \frac{t}{N} \left[\|f_\theta(\mathbf{X}_D^{adv}) - f_\theta(\mathbf{X}_D)\|_2 (\|f_\theta(\mathbf{X}_D^{adv}) + f_\theta(\mathbf{X}_D)\|_2 + 1) \right] \\
& + Kt \left[\|\mathbb{E}[f_\theta(\mathbf{X}_u^{adv})] - \mathbb{E}[f_\theta(\mathbf{X}_u)]\|_2 (\|\mathbb{E}[f_\theta(\mathbf{X}_u^{adv})] + \mathbb{E}[f_\theta(\mathbf{X}_u)]\|_2 + 1) \right] + M \cdot \sqrt{\frac{2K \ln 2 + 2 \ln(\frac{1}{\sigma})}{N}}.
\end{aligned}$$

Finally, a further refinement of the upper bound can be achieved through the L -Lipschitz constant of $f_\theta(\cdot)$:

$$\begin{aligned}
& \leq \varepsilon_{GE} + \frac{t}{N} \|[f_\theta(\mathbf{X}_D^{adv})]^2 - [f_\theta(\mathbf{X}_D)]^2\|_2 + Kt \|\mathbb{E}[f_\theta(\mathbf{X}_u^{adv})]^2 - \mathbb{E}[f_\theta(\mathbf{X}_u)]^2\|_2 \\
& + 2t(N + K^2 + 1)HL\|\delta\|_2 + M \cdot \sqrt{\frac{2K \ln 2 + 2 \ln(\frac{1}{\sigma})}{N}},
\end{aligned}$$

where $H = \max_x \|f_\theta(x)\|_2$. This new derivation is clearer and avoids the use of constants that are difficult to express. We believe this improvement not only enhances the theoretical interpretability but also strengthens the rigor of the analysis.

2.2 $\text{tr}(\mathbf{G}) = N$

Regarding your question on why $\text{tr}(\mathbf{G}) = N$, we express the forms of \mathbf{D} and \mathbf{D}^* to better illustrate this relationship. Specifically, we present the following expressions to provide a clearer understanding of the structure and behavior of these matrices in the context of our proof.

The matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ can be written as:

$$\mathbf{D} = \mathbf{I} + \begin{bmatrix} \mathbf{Z}_1^\top \mathbf{Z}_1 & \mathbf{Z}_1^\top \mathbf{Z}_2 & \cdots & \mathbf{Z}_1^\top \mathbf{Z}_k \\ \mathbf{Z}_2^\top \mathbf{Z}_1 & \mathbf{Z}_2^\top \mathbf{Z}_2 & \cdots & \mathbf{Z}_2^\top \mathbf{Z}_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_k^\top \mathbf{Z}_1 & \mathbf{Z}_k^\top \mathbf{Z}_2 & \cdots & \mathbf{Z}_k^\top \mathbf{Z}_k \end{bmatrix}.$$

Furthermore, we define $\mathbf{D}^* \in \mathbb{R}^{N \times N}$ as:

$$\mathbf{D}^* = \mathbf{I} + \begin{bmatrix} \mathbf{Z}_1^\top \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2^\top \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_k^\top \mathbf{Z}_k \end{bmatrix}.$$

Thus, $\mathbf{G} \in \mathbb{R}^{N \times N}$ is given by:

$$\mathbf{G} = \begin{bmatrix} (\mathbf{I} + \mathbf{Z}_1^\top \mathbf{Z}_1)^{-1} (\mathbf{I} + \mathbf{Z}_1^\top \mathbf{Z}_1) & \cdots & (\mathbf{I} + \mathbf{Z}_1^\top \mathbf{Z}_1)^{-1} (\mathbf{I} + \mathbf{Z}_1^\top \mathbf{Z}_k) \\ \vdots & \ddots & \vdots \\ (\mathbf{I} + \mathbf{Z}_k^\top \mathbf{Z}_1)^{-1} (\mathbf{I} + \mathbf{Z}_k^\top \mathbf{Z}_1) & \cdots & (\mathbf{I} + \mathbf{Z}_k^\top \mathbf{Z}_k)^{-1} (\mathbf{I} + \mathbf{Z}_k^\top \mathbf{Z}_k) \end{bmatrix} = \begin{bmatrix} 1 & \emptyset & \cdots & \emptyset \\ \emptyset & 1 & \cdots & \emptyset \\ \vdots & \vdots & \ddots & \vdots \\ \emptyset & \emptyset & \cdots & 1 \end{bmatrix},$$

where \emptyset represents the irrelevant numbers for the trace calculation. Therefore, $\text{tr}(\mathbf{G}) = N$.