

Appendix

A. Additional Discussion on Theoretical Analysis

On the interpretation of Theorem 1. In Theorem 1, the distribution $\mathcal{D}$ is arbitrary. For example, if the number of samples generated during training is finite and n , then the simplest way to instantiate Theorem 1 is to set $\mathcal{D}$ to represent the empirical measure $\frac{1}{n} \sum_{i=1}^m \delta_{(x_i, y_i)}$ for training data $((x_i, y_i))_{i=1}^m$ (where the Dirac measures $\delta_{(x_i, y_i)}$), which yields the following:

$$\begin{aligned} & \frac{1}{n^2} \sum_{i=1}^m \sum_{j=1}^m \mathbb{E}_{\substack{\tilde{x}, \tilde{x}', \tilde{x}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha, \alpha', \alpha'' \sim \mathcal{D}_{\alpha}}} [\ell_{\text{ctr}}(\mathbf{x}_i^+, \mathbf{x}_i^{++}, \mathbf{x}_j^-)] \\ &= \frac{1}{n^2} \sum_{i=1}^m \sum_{j \in S_{y_i}} \mathbb{E}_{\substack{\tilde{x}, \tilde{x}', \tilde{x}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha, \alpha', \alpha'' \sim \mathcal{D}_{\alpha}}} [\ell_{\text{cf}}(f(\mathbf{x}_i^+), y_i)] + \frac{1}{n^2} \sum_{i=1}^n [(n - |S_{y_i}|) \mathcal{E}_y], \end{aligned}$$

where $\mathbf{x}_i^+ = \mathbf{x}_i + \alpha \delta(\mathbf{x}_i, \tilde{x})$, $\mathbf{x}_i^{++} = \mathbf{x}_i + \alpha' \delta(\mathbf{x}_i, \tilde{x}')$, $\mathbf{x}_j^- = \tilde{x}_j + \alpha'' \delta(\tilde{x}_j, \tilde{x}'')$, $S_y = \{i \in [m] : y_i \neq y\}$, $f(\mathbf{x}_i^+) = \|(h(\mathbf{x}_i^+))^{-1} h(\mathbf{x}_i^+)^\top \tilde{w}\|$, and $[m] = \{1, \dots, m\}$. Here, we used the fact that $\bar{\rho}(y) = \frac{|S_y|}{n}$ where $|S_y|$ is the number of elements in the set S_y . In general, in Theorem 1, we can set the distribution $\mathcal{D}$ to take into account additional data augmentations (that generate infinite number of samples) and the different ways that we generate positive and negative pairs.

On the interpretation of Theorem 2 for deep neural networks. Consider the case of deep neural networks with ReLU in the form of $f(\mathbf{x}) = W^{(H)} \sigma^{(H-1)}(W^{(H-1)} \sigma^{(H-2)}(\dots \sigma^{(1)}(W^{(1)} \mathbf{x}) \dots))$, where $W^{(l)}$ is the weight matrix and $\sigma^{(l)}$ is the ReLU nonlinear function at the l -th layer. In this case, we have

$$\|\nabla f(\mathbf{x})\| = \|W^{(H)} \dot{\sigma}^{(H-1)} W^{(H-1)} \dot{\sigma}^{(H-2)} \dots \dot{\sigma}^{(1)} W^{(1)}\|,$$

where $\dot{\sigma}^{(l)} = \frac{\partial \sigma^{(l)}(q)}{\partial q} \big|_{q=W^{(l-1)} \sigma^{(l-2)}(\dots \sigma^{(1)}(W^{(1)} \mathbf{x}) \dots)}$ is a Jacobian matrix and hence $W^{(H)} \dot{\sigma}^{(H-1)} W^{(H-1)} \dot{\sigma}^{(H-2)} \dots \dot{\sigma}^{(1)} W^{(1)}$ is the sum of the product of path weights. Thus, regularizing $\|\nabla f(\mathbf{x})\|$ tends to promote generalization as it corresponds to the path weight norm used in generalization error bounds in previous work (Kawaguchi et al., 2017).

B. Proof

In this section, we present complete proofs for our theoretical results. We note that in the proofs and in theorems, the distribution $\mathcal{D}$ is arbitrary. As an simplest example of the practical setting, we can set $\mathcal{D}$ to represent the empirical measure $\frac{1}{n} \sum_{i=1}^m \delta_{(x_i, y_i)}$ for training data $((x_i, y_i))_{i=1}^m$ (where the Dirac measures $\delta_{(x_i, y_i)}$), which yields the following:

$$\mathbb{E}_{\substack{\mathbf{x}, \tilde{x} \sim \mathcal{D}_{\mathbf{x}}, \\ \tilde{x}, \tilde{x}', \tilde{x}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha, \alpha', \alpha'' \sim \mathcal{D}_{\alpha}}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] = \frac{1}{n^2} \sum_{i=1}^m \sum_{j=1}^m \mathbb{E}_{\substack{\tilde{x}, \tilde{x}', \tilde{x}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha, \alpha', \alpha'' \sim \mathcal{D}_{\alpha}}} [\ell_{\text{ctr}}(\mathbf{x}_i^+, \mathbf{x}_i^{++}, \mathbf{x}_j^-)], \quad (11)$$

where $\mathbf{x}_i^+ = \mathbf{x}_i + \alpha \delta(\mathbf{x}_i, \tilde{x})$, $\mathbf{x}_i^{++} = \mathbf{x}_i + \alpha' \delta(\mathbf{x}_i, \tilde{x}')$, and $\mathbf{x}_j^- = \tilde{x}_j + \alpha'' \delta(\tilde{x}_j, \tilde{x}'')$. In equation (11), we can more easily see that for each single point x_i , we have the m negative examples as:

$$\sum_{j=1}^m \mathbb{E}_{\substack{\tilde{x}, \tilde{x}', \tilde{x}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha, \alpha', \alpha'' \sim \mathcal{D}_{\alpha}}} [\ell_{\text{ctr}}(\mathbf{x}_i^+, \mathbf{x}_i^{++}, \mathbf{x}_j^-)].$$

Thus, for each single point x_i , all points generated based on all other points $\tilde{x}_j$ for $j = 1, \dots, m$ are treated as negatives, whereas the positives are the ones generated based on the particular point x_i . The ratio of negatives increases as the number of original data points increases and our proofs apply for any number of original data points.

B.1. Proof of Theorem 1

We begin by introducing additional notation to be used in our proof. For two vectors $\mathbf{q}$ and $\mathbf{q}'$, define

$$\overline{\text{cov}}[\mathbf{q}, \mathbf{q}'] = \sum_k \text{cov}(\mathbf{q}_k, \mathbf{q}'_k)$$

Let $\rho_y = \mathbb{E}_{\bar{y}|y}[1_{[\bar{y}=y]}] = \sum_{\bar{y} \in \{0,1\}} p_{\bar{y}}(\bar{y} | y) 1_{[\bar{y}=y]} = \Pr(\bar{y} = y | y)$. For the completeness, we first recall the following well known fact:

Lemma 1. For any $y \in \{0, 1\}$ and $q \in \mathbb{R}$,

$$\ell(q, y) = -\log \left(\frac{\exp(yq)}{1 + \exp(q)} \right)$$

Proof. By simple arithmetic manipulations,

$$\begin{aligned} \ell(q, y) &= -y \log \left(\frac{1}{1 + \exp(-q)} \right) - (1 - y) \log \left(1 - \frac{1}{1 + \exp(-q)} \right) \\ &= -y \log \left(\frac{1}{1 + \exp(-q)} \right) - (1 - y) \log \left(\frac{\exp(-q)}{1 + \exp(-q)} \right) \\ &= -y \log \left(\frac{\exp(q)}{1 + \exp(q)} \right) - (1 - y) \log \left(\frac{1}{1 + \exp(q)} \right) \\ &= \begin{cases} -\log \left(\frac{\exp(q)}{1 + \exp(q)} \right) & \text{if } y = 1 \\ -\log \left(\frac{1}{1 + \exp(q)} \right) & \text{if } y = 0 \end{cases} \\ &= -\log \left(\frac{\exp(yq)}{1 + \exp(q)} \right). \end{aligned}$$

□

Before starting the main parts of the proof, we also prepare the following simple facts:

Lemma 2. For any $(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)$, we have

$$\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-) = \ell(\text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^{++})] - \text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^-)], 1)$$

Proof. By simple arithmetic manipulations,

$$\begin{aligned} \ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-) &= -\log \frac{\exp(\text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^{++})])}{\exp(\text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^{++})]) + \exp(\text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^-)])} \\ &= -\log \frac{1}{1 + \exp(\text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^-)] - \text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^{++})])} \\ &= -\log \frac{\exp(\text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^{++})] - \text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^-)])}{1 + \exp(\text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^{++})] - \text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^-)])} \end{aligned}$$

Using Lemma 1 with $q = \text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^{++})] - \text{sim}[h(\mathbf{x}^+), h(\mathbf{x}^-)]$, this yields the desired statement. □

Lemma 3. For any $y \in \{0, 1\}$ and $q \in \mathbb{R}$,

$$\ell(-q, 1) = \ell(q, 0).$$

Proof. Using Lemma 1,

$$\ell(-q, 1) = -\log \left(\frac{\exp(-q)}{1 + \exp(-q)} \right) = -\log \left(\frac{1}{1 + \exp(q)} \right) = \ell(q, 0).$$

□

With these facts, we are now ready to start our proof. We first prove the relationship between the contrastive loss and classification loss under an ideal situation:

☐

Using the above the relationship under the ideal situation, we now proves the relationship under the practical situation:

Lemma 5. Assume that $\mathbf{x}^+ = \mathbf{x} + \alpha\delta(\mathbf{x}, \tilde{\mathbf{x}})$, $\mathbf{x}^{++} = \mathbf{x} + \alpha'\delta(\mathbf{x}, \tilde{\mathbf{x}}')$, $\mathbf{x}^- = \bar{\mathbf{x}} + \alpha''\delta(\bar{\mathbf{x}}, \tilde{\mathbf{x}}'')$, and $\text{sim}[z, z'] = \frac{z^\top z'}{\zeta(z)\zeta(z')}$ where $\zeta : z \mapsto \zeta(z) \in \mathbb{R}$. Then for any $(\alpha, \tilde{\mathbf{x}}, \delta, \zeta, y)$, we have that

$$\begin{aligned} & \mathbb{E}_{\bar{y}|y} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y, \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}}} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= (1 - \rho_y) \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y, \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y} \neq y}} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} \left[\ell \left(\frac{h(\mathbf{x}^+)^\top \tilde{w}}{\zeta(h(\mathbf{x}^+))}, y \right) \right] + \rho_y E \end{aligned}$$

where

$$\begin{aligned} E &= \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim \mathcal{D}_y} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} \left[\log \left(1 + \exp \left[-\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} - \frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right) \right] \right) \right] \\ &\geq \log \left(1 + \exp \left[-\overline{\text{cov}}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))}, \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \right] \right) \end{aligned}$$

Proof. Using Lemma 4,

$$\begin{aligned} & \mathbb{E}_{\bar{y}|y} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y, \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}}} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= \sum_{\bar{y} \in \{0,1\}} p_{\bar{y}}(\bar{y} | y) \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y, \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}}} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= \Pr(\bar{y} = 0 | y) \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}=0}, \\ \tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] + \Pr(\bar{y} = 1 | y) \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}=1}, \\ \tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= \Pr(\bar{y} \neq y | y) \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y} \neq y}, \\ \tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] + \Pr(\bar{y} = y | y) \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}=y}, \\ \tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= (1 - \rho_y) \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y} \neq y}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] + \rho_y \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}=y}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= (1 - \rho_y) \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y} \neq y}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} \left[\ell \left(\frac{h(\mathbf{x}^+)^\top \tilde{w}}{\zeta(h(\mathbf{x}^+))}, y \right) \right] + \rho_y \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}=y}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)], \end{aligned}$$

which obtain the desired statement for the first term. We now focus on the second term. Using Lemmas 1 and 2, with

$$q = \frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} - \frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right),$$

$$\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-) = \ell(q, 1) = -\log \left(\frac{\exp(q)}{1 + \exp(q)} \right) = -\log \left(\frac{1}{1 + \exp(-q)} \right) = \log(1 + \exp(-q)).$$

Therefore,

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y} \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\bar{y}=y}} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim \mathcal{D}_y} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} \left[\log \left(1 + \exp \left[-\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} - \frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right) \right] \right) \right] = E, \end{aligned}$$

which proves the desired statement with E . We now focus on the lower bound on E . By using the convexity of $q \mapsto \log(1 + \exp(-q))$ and Jensen's inequality,

$$E \geq \log \left(1 + \exp \left[\mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim \mathcal{D}_y} \mathbb{E}_{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{\mathbf{x}}}, \alpha' \sim \mathcal{D}_\alpha} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \left(\frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} - \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right) \right] \right] \right)$$

$$\begin{aligned}
 &= \log \left(1 + \exp \left[\mathbb{E} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right] - \mathbb{E} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \right] \right) \\
 &= \log \left(1 + \exp \left[\mathbb{E} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \right] \mathbb{E} \left[\frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right] - \mathbb{E} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \right] \right)
 \end{aligned}$$

Here, we have

$$\begin{aligned}
 &\mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \\
 &= \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \sum_k \left(\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))} \right)_k \left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right)_k \\
 &= \sum_k \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \left(\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))} \right)_k \left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right)_k \\
 &= \sum_k \mathbb{E} \left[\left(\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))} \right)_k \right] \mathbb{E} \left[\left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right)_k \right] + \sum_k \text{cov} \left(\left(\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))} \right)_k, \left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right)_k \right) \\
 &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \right] \mathbb{E}_{\substack{\tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] + \overline{\text{cov}} \left[\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))}, \frac{h(\mathbf{x})}{\zeta(h(\mathbf{x}))} \right]
 \end{aligned}$$

Since $\mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \left[\left(\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right) \right] = \mathbb{E}_{\substack{\tilde{\mathbf{x}} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha'' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right]$,

$$\begin{aligned}
 &\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \right] \mathbb{E}_{\substack{\tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha'' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right] - \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \\
 &= \mathbb{E} \left[\frac{h(\mathbf{x}^+)^\top}{\zeta(h(\mathbf{x}^+))} \right] \left(\mathbb{E}_{\substack{\tilde{\mathbf{x}} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha'' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^-)}{\zeta(h(\mathbf{x}^-))} \right] - \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}}' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha' \sim \mathcal{D}_\alpha}} \left[\frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \right) - \overline{\text{cov}} \left[\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))}, \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \\
 &= -\overline{\text{cov}} \left[\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))}, \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right]
 \end{aligned}$$

Substituting this to the above inequality on E ,

$$E \geq \log \left(1 + \exp \left[-\overline{\text{cov}} \left[\frac{h(\mathbf{x}^+)}{\zeta(h(\mathbf{x}^+))}, \frac{h(\mathbf{x}^{++})}{\zeta(h(\mathbf{x}^{++}))} \right] \right] \right),$$

which proves the desired statement for the lower bound on E . $\square$

With these lemmas, we are now ready to prove Theorem 1:

Proof of Theorem 1. From Lemma 5, we have that

$$\begin{aligned}
 &\mathbb{E}_{\tilde{y}|y} \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\tilde{y}}}} \mathbb{E}_{\substack{\tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\
 &= (1 - \rho_y) \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\tilde{y}}}} \mathbb{E}_{\substack{\tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} \left[\ell_{\text{cf}} \left(\frac{h(\mathbf{x}^+)^\top \tilde{w}}{\zeta(h(\mathbf{x}^+))}, y \right) \right] + \rho_y E
 \end{aligned}$$

By taking expectation over y in both sides,

$$\mathbb{E}_{y, \tilde{y}} \mathbb{E}_{\substack{\mathbf{x} \sim \mathcal{D}_y, \\ \tilde{\mathbf{x}} \sim \mathcal{D}_{\tilde{y}}}} \mathbb{E}_{\substack{\tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)]$$

$$= \mathbb{E}_y \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y, \substack{\tilde{\mathbf{x}} \sim \mathcal{D}_{\tilde{y} \neq y} \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} \left[(1 - \rho_y) \ell_{\text{cf}} \left(\frac{h(\mathbf{x}^+)^\top \tilde{w}}{\zeta(h(\mathbf{x}^+))}, y \right) \right] + \mathbb{E}_y [\rho_y E]$$

Since $\mathbb{E}_y \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y} [\varphi(x)] = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\varphi(x)] = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_x} [\varphi(x)]$ given a function φ of x , we have

$$\begin{aligned} & \mathbb{E}_{\substack{\mathbf{x}, \tilde{\mathbf{x}} \sim \mathcal{D}_x, \\ \tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} [\ell_{\text{ctr}}(\mathbf{x}^+, \mathbf{x}^{++}, \mathbf{x}^-)] \\ &= \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \mathbb{E}_{\substack{\tilde{\mathbf{x}} \sim \mathcal{D}_{\tilde{y}}, \\ \tilde{\mathbf{x}}', \tilde{\mathbf{x}}'' \sim \mathcal{D}_{\tilde{x}}, \\ \alpha', \alpha'' \sim \mathcal{D}_\alpha}} \left[\bar{\rho}(y) \ell_{\text{cf}} \left(\frac{h(\mathbf{x}^+)^\top \tilde{w}}{\zeta(h(\mathbf{x}^+))}, y \right) \right] + \mathbb{E}_y [(1 - \bar{\rho}(y)) E] \end{aligned}$$

Taking expectations over $\tilde{\mathbf{x}} \sim \mathcal{D}_{\tilde{x}}$ and $\alpha \sim \mathcal{D}_\alpha$ in both sides yields the desired statement. $\square$

B.2. Proof of Theorem 2

We begin by introducing additional notation. Define $\ell_{f,y}(q) = \ell(f(q), y)$ and $\ell_y(q) = \ell(q, y)$. Note that $\ell(f(q), y) = \ell_{f,y}(q) = (\ell_y \circ f)(q)$. The following shows that the contrastive pre-training is related to minimizing the standard classification loss $\ell(f(\mathbf{x}), y)$ while regularizing the change of the loss values in the direction of $\delta(\mathbf{x}, \tilde{\mathbf{x}})$:

Lemma 6. Assume that $\ell_{f,y}$ is twice differentiable. Then there exists a function φ such that $\lim_{q \rightarrow 0} \varphi(q) = 0$ and

$$\ell(f(\mathbf{x}^+), y) = \ell(f(\mathbf{x}), y) + \alpha \nabla \ell_{f,y}(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \delta(\mathbf{x}, \tilde{\mathbf{x}})^\top \nabla^2 \ell_{f,y}(\mathbf{x}) \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \alpha^2 \varphi(\alpha).$$

Proof. Let $\mathbf{x}$ be an arbitrary point in the domain of f . Let $\varphi_0(\alpha) = \ell(f(\mathbf{x}^+), y) = \ell_{f,y}(\mathbf{x} + \alpha \delta(\mathbf{x}, \tilde{\mathbf{x}}))$. Then, using the definition of the twice-differentiability of function φ_0 , there exists a function φ such that

$$\ell(f(\mathbf{x}^+), y) = \varphi_0(\alpha) = \varphi_0(0) + \varphi'_0(0)\alpha + \frac{1}{2} \varphi''_0(0)\alpha^2 + \alpha^2 \varphi(\alpha), \quad (12)$$

where $\lim_{\alpha \rightarrow 0} \varphi(\alpha) = 0$. By chain rule,

$$\begin{aligned} \varphi'_0(\alpha) &= \frac{\partial \ell(f(\mathbf{x}^+), y)}{\partial \alpha} = \frac{\partial \ell(f(\mathbf{x}^+), y)}{\partial \mathbf{x}^+} \frac{\partial \mathbf{x}^+}{\partial \alpha} = \frac{\partial \ell(f(\mathbf{x}^+), y)}{\partial \mathbf{x}^+} \delta(\mathbf{x}, \tilde{\mathbf{x}}) = \nabla \ell_{f,y}(\mathbf{x}^+)^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) \\ \varphi''_0(\alpha) &= \delta(\mathbf{x}, \tilde{\mathbf{x}})^\top \left[\frac{\partial}{\partial \alpha} \left(\frac{\partial \ell(f(\mathbf{x}^+), y)}{\partial \mathbf{x}^+} \right)^\top \right] = \delta(\mathbf{x}, \tilde{\mathbf{x}})^\top \left[\frac{\partial}{\partial \mathbf{x}^+} \left(\frac{\partial \ell(f(\mathbf{x}^+), y)}{\partial \mathbf{x}^+} \right)^\top \right] \frac{\partial \mathbf{x}^+}{\partial \alpha} \\ &= \delta(\mathbf{x}, \tilde{\mathbf{x}})^\top \nabla^2 \ell_{f,y}(\mathbf{x}^+) \delta(\mathbf{x}, \tilde{\mathbf{x}}) \end{aligned}$$

Therefore,

$$\begin{aligned} \varphi'_0(0) &= \nabla \ell_{f,y}(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) \\ \varphi''_0(0) &= \delta(\mathbf{x}, \tilde{\mathbf{x}})^\top \nabla^2 \ell_{f,y}(\mathbf{x}) \delta(\mathbf{x}, \tilde{\mathbf{x}}). \end{aligned}$$

By substituting this to the above equation based on the definition of twice differentiability,

$$\ell(f(\mathbf{x}^+), y) = \varphi_0(\alpha) = \ell(f(\mathbf{x}), y) + \alpha \nabla \ell_{f,y}(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \delta(\mathbf{x}, \tilde{\mathbf{x}})^\top \nabla^2 \ell_{f,y}(\mathbf{x}) \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \alpha^2 \varphi(\alpha).$$

$\square$

Whereas the above lemma is at the level of loss, we now analyze the phenomena at the level of model:

Lemma 7. Let $\mathbf{x}$ be a fixed point in the domain of f . Given the fixed $\mathbf{x}$, let $w \in \mathcal{W}$ be a point such that $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ exist. Assume that $f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \mathbf{x}$ and $\nabla^2 f(\mathbf{x}) = 0$. Then we have

$$\begin{aligned} & \ell(f(\mathbf{x}^+), y) \\ &= \ell(f(\mathbf{x}), y) + \alpha (\psi(f(\mathbf{x})) - y) \nabla f(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x})) |\nabla f(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}})|^2 + \alpha^2 \varphi(\alpha), \end{aligned}$$

where $\psi'(\cdot) = \psi(\cdot)(1 - \psi(\cdot)) > 0$.

Proof. Under these conditions,

$$\nabla \ell_{f,y}(\mathbf{x}) = \nabla(\ell_y \circ f)(\mathbf{x}) = \ell'_y(f(\mathbf{x}))\nabla f(\mathbf{x})$$

$$\nabla^2 \ell_{f,y}(\mathbf{x}) = \ell''_y(f(\mathbf{x}))\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top + \ell'_y(f(\mathbf{x}))\nabla^2 f(\mathbf{x}) = \ell''_y(f(\mathbf{x}))\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top$$

Substituting these into Lemma 6 yields

$$\begin{aligned} \ell(f(\mathbf{x}^+), y) &= \ell(f(\mathbf{x}), y) + \alpha \ell'_y(f(\mathbf{x}))\nabla f(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \ell''_y(f(\mathbf{x}))\delta(\mathbf{x}, \tilde{\mathbf{x}})^\top [\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top] \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \alpha^2 \varphi(\alpha) \\ &= \ell(f(\mathbf{x}), y) + \alpha \ell'_y(f(\mathbf{x}))\nabla f(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \ell''_y(f(\mathbf{x}))[\nabla f(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}})]^2 + \alpha^2 \varphi(\alpha) \end{aligned}$$

Using Lemma 1, we can rewrite this loss as follows:

$$\ell(f(\mathbf{x}), y) = -\log \frac{\exp(yf(\mathbf{x}))}{1 + \exp(f(\mathbf{x}))} = \log[1 + \exp(f(\mathbf{x}))] - yf(\mathbf{x}) = \psi_0(f(\mathbf{x})) - yf(\mathbf{x})$$

where $\psi_0(q) = \log[1 + \exp(q)]$. Thus,

$$\ell'_y(f(\mathbf{x})) = \psi'_0(f(\mathbf{x})) - y = \psi(f(\mathbf{x})) - y$$

$$\ell''_y(f(\mathbf{x})) = \psi''_0(f(\mathbf{x})) = \psi'(f(\mathbf{x}))$$

Substituting these into the above equation, we have

$$\begin{aligned} \ell(f(\mathbf{x}^+), y) &= \ell(f(\mathbf{x}), y) + \alpha(\psi(f(\mathbf{x})) - y)\nabla f(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x}))[\nabla f(\mathbf{x})^\top \delta(\mathbf{x}, \tilde{\mathbf{x}})]^2 + \alpha^2 \varphi(\alpha) \end{aligned}$$

□

The following lemma shows that Mixup version is related to minimize the standard classification loss plus the regularization term on $\|\nabla f(\mathbf{x})\|$.

Lemma 8. *Let $\delta(\mathbf{x}, \tilde{\mathbf{x}}) = \tilde{\mathbf{x}} - \mathbf{x}$. Let $\mathbf{x}$ be a fixed point in the domain of f . Given the fixed $\mathbf{x}$, let $w \in \mathcal{W}$ be a point such that $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ exist. Assume that $f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \mathbf{x}$ and $\nabla^2 f(\mathbf{x}) = 0$. Assume that $\mathbb{E}_{\tilde{\mathbf{x}}}[\tilde{\mathbf{x}}] = 0$. Then, if $yf(\mathbf{x}) + (y - 1)f(\mathbf{x}) \geq 0$,*

$$\begin{aligned} \mathbb{E}_{\tilde{\mathbf{x}}} \ell(f(\mathbf{x}^+), y) &= \ell(f(\mathbf{x}), y) + c_1(\mathbf{x})\|\nabla f(\mathbf{x})\|_2 + c_2(\mathbf{x})\|\nabla f(\mathbf{x})\|_2^2 + c_3(\mathbf{x})\|\nabla f(\mathbf{x})\|_{\mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\tilde{\mathbf{x}}}}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top]}^2 + O(\alpha^3), \end{aligned}$$

where

$$c_1(\mathbf{x}) = \alpha |\cos(\nabla f(\mathbf{x}), \mathbf{x})| |y - \psi(f(\mathbf{x}))| \|\mathbf{x}\|_2 \geq 0$$

$$c_2(\mathbf{x}) = \frac{\alpha^2 |\cos(\nabla f(\mathbf{x}), \mathbf{x})|^2 \|\mathbf{x}\|_2^2}{2} |\psi'(f(\mathbf{x}))| \geq 0$$

$$c_3(\mathbf{x}) = \frac{\alpha^2}{2} |\psi'(f(\mathbf{x}))| > 0.$$

Proof. Using Lemma 7 with $\delta(\mathbf{x}, \tilde{\mathbf{x}}) = \tilde{\mathbf{x}} - \mathbf{x}$,

$$\begin{aligned} \ell(f(\mathbf{x}^+), y) &= \ell(f(\mathbf{x}), y) + \alpha(\psi(f(\mathbf{x})) - y)\nabla f(\mathbf{x})^\top (\tilde{\mathbf{x}} - \mathbf{x}) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x}))|\nabla f(\mathbf{x})^\top (\tilde{\mathbf{x}} - \mathbf{x})|^2 + \alpha^2 \varphi(\alpha) \\ &= \ell(f(\mathbf{x}), y) - \alpha(\psi(f(\mathbf{x})) - y)\nabla f(\mathbf{x})^\top (\mathbf{x} - \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x}))|\nabla f(\mathbf{x})^\top (\mathbf{x} - \tilde{\mathbf{x}})|^2 + \alpha^2 \varphi(\alpha) \\ &= \ell(f(\mathbf{x}), y) - \alpha(\psi(f(\mathbf{x})) - y)(f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x}))|f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}|^2 + \alpha^2 \varphi(\alpha) \end{aligned}$$

$$= \ell(f(\mathbf{x}), y) + \alpha(y - \psi(f(\mathbf{x}))(f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x})) |f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}|^2 + \alpha^2 \varphi(\alpha)$$

Therefore, using $\mathbb{E}_{\tilde{\mathbf{x}}} \tilde{\mathbf{x}} = 0$,

$$\begin{aligned} & \mathbb{E}_{\tilde{\mathbf{x}}} \ell(f(\mathbf{x}^+), y) \\ &= \ell(f(\mathbf{x}), y) + \alpha[y - \psi(f(\mathbf{x}))]f(\mathbf{x}) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x})) \mathbb{E}_{\tilde{\mathbf{x}}} |f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}|^2 + \mathbb{E}_{\tilde{\mathbf{x}}} \alpha^2 \varphi(\alpha) \end{aligned}$$

Since $|f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}|^2 = f(\mathbf{x})^2 - 2f(\mathbf{x})\nabla f(\mathbf{x})^\top \tilde{\mathbf{x}} + (\nabla f(\mathbf{x})^\top \tilde{\mathbf{x}})^2$,

$$\begin{aligned} \mathbb{E}_{\tilde{\mathbf{x}}} |f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}|^2 &= f(\mathbf{x})^2 + \mathbb{E}_{\tilde{\mathbf{x}}} (\nabla f(\mathbf{x})^\top \tilde{\mathbf{x}})^2 \\ &= f(\mathbf{x})^2 + \nabla f(\mathbf{x})^\top \mathbb{E}_{\tilde{\mathbf{x}}} [\tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top] \nabla f(\mathbf{x}). \end{aligned}$$

Thus,

$$\begin{aligned} & \mathbb{E}_{\tilde{\mathbf{x}}} \ell(f(\mathbf{x}^+), y) \\ &= \ell(f(\mathbf{x}), y) + \alpha[y - \psi(f(\mathbf{x}))]f(\mathbf{x}) + \frac{\alpha^2}{2} |\psi'(f(\mathbf{x}))| [f(\mathbf{x})^2 + \nabla f(\mathbf{x})^\top \mathbb{E}_{\tilde{\mathbf{x}}} [\tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top] \nabla f(\mathbf{x})] + \mathbb{E}_{\tilde{\mathbf{x}}} \alpha^2 \varphi(\alpha) \end{aligned}$$

The assumption that $yf(\mathbf{x}) + (y - 1)f(\mathbf{x}) \geq 0$ implies that $f(\mathbf{x}) \geq 0$ if $y = 1$ and $f(\mathbf{x}) \leq 0$ if $y = 0$. Thus, if $y = 1$,

$$[y - \psi(f(\mathbf{x}))]f(\mathbf{x}) = [1 - \psi(f(\mathbf{x}))]f(\mathbf{x}) \geq 0,$$

since $f(\mathbf{x}) \geq 0$ and $(1 - \psi(f(\mathbf{x}))) \geq 0$ due to $\psi(f(\mathbf{x})) \in (0, 1)$. If $y = 0$,

$$[y - \psi(f(\mathbf{x}))]f(\mathbf{x}) = -\psi(f(\mathbf{x}))f(\mathbf{x}) \geq 0,$$

since $f(\mathbf{x}) \leq 0$ and $-\psi(f(\mathbf{x})) < 0$. Therefore, in both cases,

$$[y - \psi(f(\mathbf{x}))]f(\mathbf{x}) \geq 0,$$

which implies that,

$$\begin{aligned} y - \psi(f(\mathbf{x})) &= [y - \psi(f(\mathbf{x}))]f(\mathbf{x}) \\ &= |y - \psi(f(\mathbf{x}))| |\nabla f(\mathbf{x})^\top \mathbf{x}| \\ &= |y - \psi(f(\mathbf{x}))| \|\nabla f(\mathbf{x})\| \|\mathbf{x}\| \cos(\nabla f(\mathbf{x}), \mathbf{x}) \end{aligned}$$

Therefore, substituting this and using $f(\mathbf{x}) = \|\nabla f(\mathbf{x})\| \|\mathbf{x}\| \cos(\nabla f(\mathbf{x}), \mathbf{x})$

$$\begin{aligned} & \mathbb{E}_{\tilde{\mathbf{x}}} \ell(f(\mathbf{x}^+), y) \\ &= \ell(f(\mathbf{x}), y) + c_1(\mathbf{x}) \|\nabla f(\mathbf{x})\|_2 + c_2(\mathbf{x}) \|\nabla f(\mathbf{x})\|_2^2 + c_3(\mathbf{x}) \nabla f(\mathbf{x})^\top \mathbb{E}_{\tilde{\mathbf{x}}} [\tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top] \nabla f(\mathbf{x}) + \mathbb{E}_{\tilde{\mathbf{x}}} [\alpha^2 \varphi(\alpha)]. \end{aligned}$$

□

In the case of Gaussian-noise, we have $\delta(\mathbf{x}, \tilde{\mathbf{x}}) = \tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)$:

Lemma 9. *Let $\delta(\mathbf{x}, \tilde{\mathbf{x}}) = \tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)$. Let $\mathbf{x}$ be a fixed point in the domain of f . Given the fixed $\mathbf{x}$, let $w \in \mathcal{W}$ be a point such that $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ exist. Assume that $f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \mathbf{x}$ and $\nabla^2 f(\mathbf{x}) = 0$. Then*

$$\mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)} \ell(f(\mathbf{x}^+), y) = \ell(f(\mathbf{x}), y) + \sigma^2 c_3(\mathbf{x}) \|\nabla f(\mathbf{x})\|_2^2 + \alpha^2 \varphi(\alpha)$$

where

$$c_3(\mathbf{x}) = \frac{\alpha^2}{2} |\psi'(f(\mathbf{x}))| > 0.$$

Proof. With $\delta(\mathbf{x}, \tilde{\mathbf{x}}) = \tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)$, Lemma 7 yields

$$\begin{aligned} & \ell(f(\mathbf{x}^+), y) \\ &= \ell(f(\mathbf{x}), y) + \alpha(\psi(f(\mathbf{x})) - y) \nabla f(\mathbf{x})^\top \tilde{\mathbf{x}} + \frac{\alpha^2}{2} \psi'(f(\mathbf{x})) |\nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}|^2 + \alpha^2 \varphi(\alpha), \end{aligned}$$

Thus,

$$\begin{aligned} & \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)} \ell(f(\mathbf{x}^+), y) \\ &= \ell(f(\mathbf{x}), y) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x})) \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)} |\nabla f(\mathbf{x})^\top \tilde{\mathbf{x}}|^2 + \alpha^2 \varphi(\alpha) \\ &= \ell(f(\mathbf{x}), y) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x})) \nabla f(\mathbf{x})^\top \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)} [\tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top] \nabla f(\mathbf{x}) + \alpha^2 \varphi(\alpha) \\ &= \ell(f(\mathbf{x}), y) + \frac{\alpha^2}{2} \psi'(f(\mathbf{x})) \|\nabla f(\mathbf{x})\|_{\mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)} [\tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top]}^2 + \alpha^2 \varphi(\alpha) \end{aligned}$$

By noticing that $\|w\|_{\mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2 I)} [\tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top]}^2 = \sigma^2 w^\top I w = \sigma^2 \|w\|_2^2$, this implies the desired statement. $\square$

Combining Lemmas 8–9 yield the statement of Theorem 2.

B.3. Proof of Theorem 3

Proof. Applying the standard result (Bartlett & Mendelson, 2002) yields that with probability at least $1 - \delta$,

$$\mathbb{E}_{(\mathbf{x}, y)} [1_{[(2y-1) \neq \text{sign}(f(\mathbf{x}))]}] - \frac{1}{n} \sum_{i=1}^n \phi((2y_i - 1)f(\mathbf{x}_i)) \leq 4L_\phi \mathcal{R}_n(\mathcal{F}_b^{(\text{mix})}) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

The rest of the proof bounds the Rademacher complexity $\mathcal{R}_n(\mathcal{F}_b^{(\text{mix})})$.

$$\begin{aligned} \hat{\mathcal{R}}_n(\mathcal{F}_b^{(\text{mix})}) &= \mathbb{E}_\xi \sup_{f \in \mathcal{F}_b} \frac{1}{n} \sum_{i=1}^n \xi_i f(\mathbf{x}_i) \\ &= \mathbb{E}_\xi \sup_{w: \|w\|_{\mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\mathbf{x}}} [\tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top]}^2 \leq b} \frac{1}{n} \sum_{i=1}^n \xi_i w^\top \mathbf{x}_i \\ &= \mathbb{E}_\xi \sup_{w: w^\top \Sigma_X w \leq b} \frac{1}{n} \sum_{i=1}^n \xi_i (\Sigma_X^{1/2} w)^\top \Sigma_X^{\dagger/2} \mathbf{x}_i \\ &\leq \frac{1}{n} \mathbb{E}_\xi \sup_{w: w^\top \Sigma_X w \leq b} \|\Sigma_X^{1/2} w\|_2 \left\| \sum_{i=1}^n \xi_i \Sigma_X^{\dagger/2} \mathbf{x}_i \right\|_2 \\ &\leq \frac{\sqrt{b}}{n} \mathbb{E}_\xi \sqrt{\sum_{i=1}^n \sum_{j=1}^n \xi_i \xi_j (\Sigma_X^{\dagger/2} \mathbf{x}_i)^\top (\Sigma_X^{\dagger/2} \mathbf{x}_j)} \\ &\leq \frac{\sqrt{b}}{n} \sqrt{\mathbb{E}_\xi \sum_{i=1}^n \sum_{j=1}^n \xi_i \xi_j (\Sigma_X^{\dagger/2} \mathbf{x}_i)^\top (\Sigma_X^{\dagger/2} \mathbf{x}_j)} \\ &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n (\Sigma_X^{\dagger/2} \mathbf{x}_i)^\top (\Sigma_X^{\dagger/2} \mathbf{x}_i)} \\ &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \mathbf{x}_i^\top \Sigma_X^\dagger \mathbf{x}_i} \end{aligned}$$

Therefore,

$$\begin{aligned}
 \mathcal{R}_n(\mathcal{F}_b^{(\text{mix})}) &= \mathbb{E}_S \hat{\mathcal{R}}_n(\mathcal{F}_b^{(\text{mix})}) = \mathbb{E}_S \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \mathbf{x}_i^\top \Sigma_X^\dagger \mathbf{x}_i} \\
 &\leq \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \mathbb{E}_{\mathbf{x}_i} \mathbf{x}_i^\top \Sigma_X^\dagger \mathbf{x}_i} \\
 &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \mathbb{E}_{\mathbf{x}_i} \sum_{k,l} (\Sigma_X^\dagger)_{kl} (\mathbf{x}_i)_k (\mathbf{x}_i)_l} \\
 &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \sum_{k,l} (\Sigma_X^\dagger)_{kl} \mathbb{E}_{\mathbf{x}_i} (\mathbf{x}_i)_k (\mathbf{x}_i)_l} \\
 &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \sum_{k,l} (\Sigma_X^\dagger)_{kl} (\Sigma_X)_{kl}} \\
 &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \text{tr}(\Sigma_X^\top \Sigma_X^\dagger)} \\
 &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \text{tr}(\Sigma_X \Sigma_X^\dagger)} \\
 &= \frac{\sqrt{b}}{n} \sqrt{\sum_{i=1}^n \text{rank}(\Sigma_X)} \\
 &\leq \frac{\sqrt{b} \sqrt{\text{rank}(\Sigma_X)}}{\sqrt{n}}
 \end{aligned}$$

□

C. Best Hyperparameter Values for Various Experiments

In general, we found that our method works well for a large range of α values ($\alpha \in [0.6, 0.9]$) and ρ values ($\rho \in [0.1, 0.5]$). In Table 5, 6 and 7, we present the best hyperparameter values for the experiments in Section 5.

Method	Fashion-MNIST	CIFAR10
Gaussian-noise	Gaussian-mean=0.1, $\tau=1.0$	Gaussian-mean=0.05, $\tau=1.0$
DACL	$\alpha=0.9$, $\tau=1.0$	$\alpha=0.9$, $\tau=1.0$
DACL+	$\alpha=0.6$, $\tau=1$, $\rho=0.1$	$\alpha=0.7$, $\tau=1.0$, $\rho=0.5$

Table 5. Best hyperparamter values for experiments on Tabular data (Table 1)

Method	CIFAR10	CIFAR100
Gaussian-noise	Gaussian-mean=0.05, $\tau=0.1$	Gaussian-mean=0.05, $\tau=0.1$
DACL	$\alpha=0.9, \tau=1.0$	$\alpha=0.9, \tau=1.0$
DACL+	$\alpha=0.9, \rho=0.1, \tau=1.0$	$\alpha=0.9, \rho=0.5, \tau=1.0$
SimCLR	$\tau=0.5$	$\tau=0.5$
SimCLR+DACL	$\alpha=0.7, \tau=1.0$	$\alpha=0.7, \tau=1.0$

Table 6. Best hyperparameter values for experiment of CIFAR10/100 dataset (Table 2)

Method	ImageNet
Gaussian-noise	Gaussian-mean=0.1, $\tau=1.0$
DACL	$\alpha=0.9, \tau=1.0$
SimCLR	$\tau=0.1$
SimCLR+DACL	$\alpha=0.9, \tau=0.1$

Table 7. Best hyperparameter values for experiments on ImageNet data (Table 3)