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:

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

where $\boldsymbol{x}_i^+=\boldsymbol{x}_i+\alpha\delta(\boldsymbol{x}_i,\tilde{\boldsymbol{x}})$, $\boldsymbol{x}_i^{++}=\boldsymbol{x}_i+\alpha'\delta(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}')$, $\boldsymbol{x}_j^-=\bar{\boldsymbol{x}}_j+\alpha''\delta(\bar{\boldsymbol{x}}_j,\tilde{\boldsymbol{x}}'')$, $S_y=\{i\in[m]:y_i\neq y\}$, $f(\boldsymbol{x}_i^+)=\|(h(\boldsymbol{x}_i^+)\|^{-1}h(\boldsymbol{x}_i^+)^\top\tilde{w}$, 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 $f(\boldsymbol{x})=W^{(H)}\sigma^{(H-1)}(W^{(H-1)}\sigma^{(H-2)}(\cdots\sigma^{(1)}(W^{(1)}\boldsymbol{x})\cdots))$, 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

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

where $\dot{\sigma}^{(l)}=\frac{\partial\sigma^{(l)}(q)}{\partial q}|_{q=W^{(l-1)}\sigma^{(l-2)}(\cdots\sigma^{(1)}(W^{(1)}x)\cdots)}$ is a Jacobian matrix and hence $W^{(H)}\dot{\sigma}^{(H-1)}W^{(H-1)}\dot{\sigma}^{(H-2)}\cdots\dot{\sigma}^{(1)}W^{(1)}$ is the sum of the product of path weights. Thus, regularizing$\|\nabla f(\boldsymbol{x})\|$ 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:

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

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

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

Thus, for each single point x_i, all points generated based on all other points $\bar{x}_j$ 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 $\boldsymbol{q}$ and $\boldsymbol{q}'$, define

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

Let $\rho_y = \mathbb{E}_{\bar{y}|y}[1_{[\bar{y}=y]}] = \sum_{\bar{y}\in\{0,1\}} p_{\bar{y}}(\bar{y} \mid y)1_{[\bar{y}=y]} = \Pr(\bar{y} = y \mid y)$. 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 $(\boldsymbol{x}^+, \boldsymbol{x}^{++}, \boldsymbol{x}^-)$, we have*

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

Proof. By simple arithmetic manipulations,

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

Using Lemma 1 with $q = \text{sim}[h(\boldsymbol{x}^+), h(\boldsymbol{x}^{++})] - \text{sim}[h(\boldsymbol{x}^+), h(\boldsymbol{x}^-)]$, 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:

Lemma 4. *Assume that $\boldsymbol{x}^+ = \boldsymbol{x} + \alpha\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, $\boldsymbol{x}^{++} = \boldsymbol{x} + \alpha'\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}}')$, $\boldsymbol{x}^- = \bar{\boldsymbol{x}} + \alpha''\delta(\bar{\boldsymbol{x}}, \tilde{\boldsymbol{x}}'')$, and $\mathrm{sim}[z, z'] = \frac{z^\top z'}{\zeta(z)\zeta(z')}$ where $\zeta : z \mapsto \zeta(z) \in \mathbb{R}$. Then for any $(\alpha, \tilde{\boldsymbol{x}}, \delta, \zeta)$ and $(y, \bar{y})$ such that $y \neq \bar{y}$, we have that*

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

Proof. Using Lemma 2 and the assumption on sim,

$$\begin{aligned}
\ell_{\text{ctr}}(\boldsymbol{x}^+, \boldsymbol{x}^{++}, \boldsymbol{x}^-) &= \ell(\mathrm{sim}[h(\boldsymbol{x}^+), h(\boldsymbol{x}^{++})] - \mathrm{sim}[h(\boldsymbol{x}^+), h(\boldsymbol{x}^-)], 1) \\
&= \ell\left(\frac{h(\boldsymbol{x}^+)^\top h(\boldsymbol{x}^{++})}{\zeta(h(\boldsymbol{x}^+))\zeta(h(\boldsymbol{x}^{++}))} - \frac{h(\boldsymbol{x}^+)^\top h(\boldsymbol{x}^-)}{\zeta(h(\boldsymbol{x}^+))\zeta(h(\boldsymbol{x}^-))}, 1\right) \\
&= \ell\left(\frac{h(\boldsymbol{x}^+)^\top}{\zeta(h(\boldsymbol{x}^+))}\left(\frac{h(\boldsymbol{x}^{++})}{\zeta(h(\boldsymbol{x}^{++}))} - \frac{h(\boldsymbol{x}^-)}{\zeta(h(\boldsymbol{x}^-))}\right), 1\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_y}\mathbb{E}_{\bar{\boldsymbol{x}}\sim\mathcal{D}_{\bar{y}\neq y}}\mathbb{E}_{\substack{\boldsymbol{x}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}[\ell_{\text{ctr}}(\boldsymbol{x}^+, \boldsymbol{x}^{++}, \boldsymbol{x}^-)] \\
&= \mathbb{E}_{\substack{\boldsymbol{x}\sim\mathcal{D}_y,\\ \bar{\boldsymbol{x}}\sim\mathcal{D}_{\bar{y}\neq y}}}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}'',\\ \alpha',\alpha''}}\left[\ell\left(\frac{h(\boldsymbol{x}+\alpha\delta(\boldsymbol{x},\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}+\alpha\delta(\boldsymbol{x},\tilde{\boldsymbol{x}})))}\left(\frac{h(\boldsymbol{x}+\alpha'\delta(\boldsymbol{x},\tilde{\boldsymbol{x}}'))}{\zeta(h(\boldsymbol{x}+\alpha'\delta(\boldsymbol{x},\tilde{\boldsymbol{x}}')))} - \frac{h(\bar{\boldsymbol{x}}+\alpha''\delta(\bar{\boldsymbol{x}},\tilde{\boldsymbol{x}}''))}{\zeta(h(\bar{\boldsymbol{x}}+\alpha''\delta(\bar{\boldsymbol{x}},\tilde{\boldsymbol{x}}'')))}\right), 1\right)\right] \\
&= \begin{cases}
\mathbb{E}_{\substack{\boldsymbol{x}^1\sim\mathcal{D}_1,\\ \boldsymbol{x}^0\sim\mathcal{D}_0}}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}''\\ \alpha',\alpha''}}\left[\ell\left(\frac{h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}})))}\left(\frac{h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}'))}{\zeta(h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}')))} - \frac{h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}''))}{\zeta(h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}'')))}\right), 1\right)\right] & \text{if } y = 1 \\
\mathbb{E}_{\substack{\boldsymbol{x}^0\sim\mathcal{D}_0,\\ \boldsymbol{x}^1\sim\mathcal{D}_1}}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}'',\\ \alpha',\alpha''}}\left[\ell\left(\frac{h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}})))}\left(\frac{h(\boldsymbol{x}^0+\alpha'\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}'))}{\zeta(h(\boldsymbol{x}^0+\alpha'\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}')))} - \frac{h(\boldsymbol{x}^1+\alpha''\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}''))}{\zeta(h(\boldsymbol{x}^1+\alpha''\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}'')))}\right), 1\right)\right] & \text{if } y = 0
\end{cases} \\
&= \begin{cases}
\mathbb{E}_{\substack{\boldsymbol{x}^1\sim\mathcal{D}_1,\\ \boldsymbol{x}^0\sim\mathcal{D}_0}}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}'',\\ \alpha',\alpha''}}\left[\ell\left(\frac{h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}})))}\left(\frac{h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}'))}{\zeta(h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}')))} - \frac{h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}''))}{\zeta(h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}'')))}\right), 1\right)\right] & \text{if } y = 1 \\
\mathbb{E}_{\substack{\boldsymbol{x}^0\sim\mathcal{D}_0,\\ \boldsymbol{x}^1\sim\mathcal{D}_1}}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}'',\\ \alpha',\alpha''}}\left[\ell\left(\frac{h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}})))}\left(\frac{h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}''))}{\zeta(h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}'')))} - \frac{h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}'))}{\zeta(h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}')))}\right), 1\right)\right] & \text{if } y = 0
\end{cases} \\
&= \begin{cases}
\mathbb{E}_{\boldsymbol{x}^1\sim\mathcal{D}_1}\mathbb{E}_{\boldsymbol{x}^0\sim\mathcal{D}_0}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}\left[\ell\left(\frac{h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}})))}\widetilde{W}(\boldsymbol{x}^1,\boldsymbol{x}^0), 1\right)\right] & \text{if } y = 1 \\
\mathbb{E}_{\boldsymbol{x}^0\sim\mathcal{D}_0}\mathbb{E}_{\boldsymbol{x}^1\sim\mathcal{D}_1}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}\left[\ell\left(-\frac{h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}})))}\widetilde{W}(\boldsymbol{x}^1,\boldsymbol{x}^0), 1\right)\right] & \text{if } y = 0
\end{cases}
\end{aligned}$$

where

$$\widetilde{W}(\boldsymbol{x}^1, \boldsymbol{x}^0) = \frac{h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}'))}{\zeta(h(\boldsymbol{x}^1+\alpha'\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}')))} - \frac{h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}''))}{\zeta(h(\boldsymbol{x}^0+\alpha''\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}'')))}.$$

Using Lemma 3,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_y}\mathbb{E}_{\bar{\boldsymbol{x}}\sim\mathcal{D}_{\bar{y}\neq y}}\mathbb{E}_{\substack{\boldsymbol{x}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}[\ell_{\text{ctr}}(\boldsymbol{x}^+, \boldsymbol{x}^{++}, \boldsymbol{x}^-)] \\
&= \begin{cases}
\mathbb{E}_{\boldsymbol{x}^1\sim\mathcal{D}_1}\mathbb{E}_{\boldsymbol{x}^0\sim\mathcal{D}_0}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}\left[\ell\left(\frac{h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}})))}\widetilde{W}(\boldsymbol{x}^1,\boldsymbol{x}^0), 1\right)\right] & \text{if } y = 1 \\
\mathbb{E}_{\boldsymbol{x}^0\sim\mathcal{D}_0}\mathbb{E}_{\boldsymbol{x}^1\sim\mathcal{D}_1}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}\left[\ell\left(\frac{h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}})))}\widetilde{W}(\boldsymbol{x}^1,\boldsymbol{x}^0), 0\right)\right] & \text{if } y = 0
\end{cases} \\
&= \begin{cases}
\mathbb{E}_{\boldsymbol{x}^1\sim\mathcal{D}_1}\mathbb{E}_{\boldsymbol{x}^0\sim\mathcal{D}_0}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}\left[\ell\left(\frac{h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^1+\alpha\delta(\boldsymbol{x}^1,\tilde{\boldsymbol{x}})))}\widetilde{W}(\boldsymbol{x}^1,\boldsymbol{x}^0), y\right)\right] & \text{if } y = 1 \\
\mathbb{E}_{\boldsymbol{x}^0\sim\mathcal{D}_0}\mathbb{E}_{\boldsymbol{x}^1\sim\mathcal{D}_1}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}\left[\ell\left(\frac{h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}^0+\alpha\delta(\boldsymbol{x}^0,\tilde{\boldsymbol{x}})))}\widetilde{W}(\boldsymbol{x}^1,\boldsymbol{x}^0), y\right)\right] & \text{if } y = 0
\end{cases} \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_y}\mathbb{E}_{\bar{\boldsymbol{x}}\sim\mathcal{D}_{\bar{y}\neq y}}\mathbb{E}_{\substack{\boldsymbol{x}',\tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha',\alpha''\sim\mathcal{D}_\alpha}}\left[\ell\left(\frac{h(\boldsymbol{x}+\alpha\delta(\boldsymbol{x},\tilde{\boldsymbol{x}}))^\top}{\zeta(h(\boldsymbol{x}+\alpha\delta(\boldsymbol{x},\tilde{\boldsymbol{x}})))}\tilde{w}, y\right)\right]
\end{aligned}$$

$\square$

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

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

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

where

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

Proof. Using Lemma 4,

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

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

Therefore,

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

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,

$$
E \geq \log\left(1+\exp\left[\mathbb{E}_{\boldsymbol{x},\bar{\boldsymbol{x}}}\mathbb{E}_{\substack{\tilde{\boldsymbol{x}}',\tilde{\boldsymbol{x}}'',\\ \alpha',\alpha''}}\left[\frac{h(\boldsymbol{x}^+)^\top}{\zeta(h(\boldsymbol{x}^+))}\left(\frac{h(\boldsymbol{x}^-)}{\zeta(h(\boldsymbol{x}^-))} - \frac{h(\boldsymbol{x}^{++})}{\zeta(h(\boldsymbol{x}^{++}))}\right)\right]\right]\right)
$$

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

Here, we have

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

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

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_y}\left[\frac{h(\boldsymbol{x}^+)^\top}{\zeta(h(\boldsymbol{x}^+))}\right]\mathbb{E}_{\substack{\bar{\boldsymbol{x}}\sim\mathcal{D}_y,\\ \tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha''\sim\mathcal{D}_\alpha}}\left[\frac{h(\boldsymbol{x}^-)}{\zeta(h(\boldsymbol{x}^-))}\right]-\mathbb{E}_{\substack{\boldsymbol{x}\sim\mathcal{D}_y,\\ \tilde{\boldsymbol{x}}'\sim\mathcal{D}_{\tilde{x}},\\ \alpha'\sim\mathcal{D}_\alpha}}\left[\frac{h(\boldsymbol{x}^+)^\top}{\zeta(h(\boldsymbol{x}^+))}\frac{h(\boldsymbol{x}^{++})}{\zeta(h(\boldsymbol{x}^{++}))}\right]$$
$$= \mathbb{E}\left[\frac{h(\boldsymbol{x}^+)^\top}{\zeta(h(\boldsymbol{x}^+))}\right]\left(\mathbb{E}_{\substack{\bar{\boldsymbol{x}}\sim\mathcal{D}_y,\\ \tilde{\boldsymbol{x}}''\sim\mathcal{D}_{\tilde{x}},\\ \alpha''\sim\mathcal{D}_\alpha}}\left[\frac{h(\boldsymbol{x}^-)}{\zeta(h(\boldsymbol{x}^-))}\right]-\mathbb{E}_{\substack{\boldsymbol{x}\sim\mathcal{D}_y,\\ \tilde{\boldsymbol{x}}'\sim\mathcal{D}_{\tilde{x}},\\ \alpha'\sim\mathcal{D}_\alpha}}\left[\frac{h(\boldsymbol{x}^{++})}{\zeta(h(\boldsymbol{x}^{++}))}\right]\right)-\overline{\operatorname{cov}}\left[\frac{h(\boldsymbol{x}^+)}{\zeta(h(\boldsymbol{x}^+))},\frac{h(\boldsymbol{x}^{++})}{\zeta(h(\boldsymbol{x}^{++}))}\right]$$
$$= -\overline{\operatorname{cov}}\left[\frac{h(\boldsymbol{x}^+)}{\zeta(h(\boldsymbol{x}^+))},\frac{h(\boldsymbol{x}^{++})}{\zeta(h(\boldsymbol{x}^{++}))}\right]$$

Substituting this to the above inequality on E,

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

which proves the desired statement for the lower bound on E. □

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

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

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

By taking expectation over y in both sides,

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

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

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

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

Taking expectations over $\tilde{\boldsymbol{x}} \sim \mathcal{D}_{\tilde{x}}$ 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 $\ell_{f,y}(q) = \ell\left(f(q), y\right)$ and $\ell_y(q) = \ell(q, y)$. Note that $\ell\left(f(q), y\right) = \ell_{f,y}(q) = (\ell_y \circ f)(q)$. The following shows that the contrastive pre-training is related to minimizing the standard classification loss $\ell(f(\boldsymbol{x}), y)$ while regularizing the change of the loss values in the direction of $\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})$:

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

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

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

$$\ell\left(f(\boldsymbol{x}^+), y\right) = \varphi_0(\alpha) = \varphi_0(0) + \varphi_0'(0)\alpha + \frac{1}{2}\varphi_0''(0)\alpha^2 + \alpha^2\varphi(\alpha), \tag{12}$$

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

$$\varphi_0'(\alpha) = \frac{\partial\ell\left(f(\boldsymbol{x}^+), y\right)}{\partial\alpha} = \frac{\partial\ell\left(f(\boldsymbol{x}^+), y\right)}{\partial\boldsymbol{x}^+}\frac{\partial\boldsymbol{x}^+}{\partial\alpha} = \frac{\partial\ell\left(f(\boldsymbol{x}^+), y\right)}{\partial\boldsymbol{x}^+}\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \nabla\ell_{f,y}(\boldsymbol{x}^+)^\top\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

$$\begin{aligned}
\varphi_0''(\alpha) = \delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})^\top\left[\frac{\partial}{\partial\alpha}\left(\frac{\partial\ell\left(f(\boldsymbol{x}^+), y\right)}{\partial\boldsymbol{x}^+}\right)^\top\right] &= \delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})^\top\left[\frac{\partial}{\partial\boldsymbol{x}^+}\left(\frac{\partial\ell\left(f(\boldsymbol{x}^+), y\right)}{\partial\boldsymbol{x}^+}\right)^\top\right]\frac{\partial\boldsymbol{x}^+}{\partial\alpha} \\
&= \delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})^\top\nabla^2\ell_{f,y}(\boldsymbol{x}^+)\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})
\end{aligned}$$

Therefore,

$$\begin{aligned}
\varphi_0'(0) &= \nabla\ell_{f,y}(\boldsymbol{x})^\top\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \\
\varphi_0''(0) &= \delta(\boldsymbol{x}, \tilde{\boldsymbol{x}})^\top\nabla^2\ell_{f,y}(\boldsymbol{x})\delta(\boldsymbol{x}, \tilde{\boldsymbol{x}}).
\end{aligned}$$

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

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

$\square$

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

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

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

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

Proof. Under these conditions,

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

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

Substituting these into Lemma 6 yields

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

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

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

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

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

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

Substituting these into the above equation, we have

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

□

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

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

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

where

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

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

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

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

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

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

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

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

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

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

Thus,

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

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

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

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

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

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

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

which implies that,

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

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

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

□

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

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

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

where

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

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

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

Thus,

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

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

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 $1 - \delta$,

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

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

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

Therefore,

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

□

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 rho 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	Gausssian-mean=0.1, τ=1.0	Gausssian-mean=0.05, τ=1.0
DACL	α=0.9, τ=1.0	α=0.9, τ=1.0
DACL+	α=0.6, τ=1, ρ=0.1	α=0.7, τ=1.0, ρ=0.5

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

Method	CIFAR10	CIFAR100
Gaussian-noise	Gaussian-mean=0.05, τ=0.1	Gaussian-mean=0.05, τ=0.1
DACL	α=0.9,τ=1.0	α=0.9, τ=1.0
DACL+	α=0.9, ρ=0.1, τ=1.0	α=0.9, ρ=0.5, τ=1.0
SimCLR	τ=0.5	τ=0.5
SimCLR+DACL	α=0.7, τ=1.0	α=0.7, τ=1.0

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

Method	ImageNet
Gaussian-noise	Gaussian-mean=0.1, τ=1.0
DACL	α=0.9, τ=1.0
SimCLR	τ=0.1
SimCLR+DACL	α=0.9, τ=0.1

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