Pointwise Binary Classification with Pairwise Confidence Comparisons: Appendix

A. Proof of Theorem 1

It is clear that each pair of examples $(\boldsymbol{x}, \boldsymbol{x}')$ is independently drawn from the following data distribution:

$$\widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') = p((\boldsymbol{x}, \boldsymbol{x}') \mid (y, y') \in \widetilde{\mathcal{Y}}) = \frac{p((\boldsymbol{x}, \boldsymbol{x}'), (y, y') \in \widetilde{\mathcal{Y}})}{p((y, y') \in \widetilde{\mathcal{Y}})},$$

where $p((y, y') \in \widetilde{\mathcal{Y}}) = \pi_+^2 + \pi_-^2 + \pi_+\pi_-$ and

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{x}', (y, y') \in \widetilde{\mathcal{Y}}) &= \sum\nolimits_{(y,y') \in \widetilde{\mathcal{Y}}} p(\boldsymbol{x}, \boldsymbol{x}' \mid (y, y')) \cdot p(y, y') \\ &= \pi_+^2 p_+(\boldsymbol{x}) p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}) p_-(\boldsymbol{x}') + \pi_+\pi_- p_+(\boldsymbol{x}) p_-(\boldsymbol{x}'). \end{aligned}$$

Finally, let $\widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') = p((\boldsymbol{x}, \boldsymbol{x}') \mid (y, y') \in \widetilde{\mathcal{Y}})$, the proof is completed. □

B. Proof of Theorem 2

In order to decompose the pairwise comparison data distribution into pointwise distribution, we marginalize $\widetilde{p}(\boldsymbol{x}, \boldsymbol{x}')$ with respect to $\boldsymbol{x}$ or $\boldsymbol{x}'$. Then we can obtain

$$\begin{aligned} \int \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') \mathrm{d}\boldsymbol{x}' =& \frac{1}{\widetilde{\pi}} \Big(\pi_+^2 p_+(\boldsymbol{x}) + \pi_-^2 p_-(\boldsymbol{x}) + \pi_+\pi_- p_+(\boldsymbol{x}) \Big) \\ =& \frac{\pi_+}{\pi_-^2 + \pi_+} p_+(\boldsymbol{x}) + \frac{\pi_-^2}{\pi_-^2 + \pi_+} p_-(\boldsymbol{x}) \\ =& \widetilde{p}_+(\boldsymbol{x}), \end{aligned}$$

and

$$\begin{aligned} \int \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') \mathrm{d}\boldsymbol{x} =& \frac{1}{\widetilde{\pi}} \big(\pi_+^2 p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}') + \pi_+\pi_- p_-(\boldsymbol{x}') \big) \\ =& \frac{\pi_+^2}{\pi_+^2 + \pi_-} p_+(\boldsymbol{x}') + \frac{\pi_-}{\pi_+^2 + \pi_-} p_-(\boldsymbol{x}') \\ =& \widetilde{p}_-(\boldsymbol{x}'), \end{aligned}$$

which concludes the proof of Theorem 2. □

C. Proof of Lemma 1

Based on Theorem 2, we can obtain the following linear equation:

$$\begin{bmatrix} \widetilde{p}_+(\boldsymbol{x}) \\ \widetilde{p}_-(\boldsymbol{x}) \end{bmatrix} = \frac{1}{\widetilde{\pi}} \begin{bmatrix} \pi_+ & \pi_-^2 \\ \pi_+^2 & \pi_- \end{bmatrix} \begin{bmatrix} p_+(\boldsymbol{x}) \\ p_-(\boldsymbol{x}) \end{bmatrix}.$$

By solving the above equation, we obtain

$$\begin{aligned} p_+(\boldsymbol{x}) &= \frac{1}{\pi_+ - \pi_-\pi_+^2} \big(\widetilde{\pi} \cdot \widetilde{p}_+(\boldsymbol{x}) - \pi_- \widetilde{\pi} \cdot \widetilde{p}_-(\boldsymbol{x}) \big) = \frac{1}{\pi_+} \big(\widetilde{p}_+(\boldsymbol{x}) - \pi_- \widetilde{p}_-(\boldsymbol{x}) \big), \\ p_-(\boldsymbol{x}) &= \frac{1}{\pi_- - \pi_+\pi_-^2} \big(\widetilde{\pi} \cdot \widetilde{p}_-(\boldsymbol{x}) - \pi_+ \widetilde{\pi} \cdot \widetilde{p}_+(\boldsymbol{x}) \big) = \frac{1}{\pi_-} \big(\widetilde{p}_-(\boldsymbol{x}) - \pi_+ \widetilde{p}_+(\boldsymbol{x}) \big), \end{aligned}$$

which concludes the proof of Lemma 1. □

D. Proof of Theorem 3

It is quite intuitive to derive

$$\begin{aligned}
R(f) &= \mathbb{E}_{p(\boldsymbol{x},y)}\big[\ell(f(\boldsymbol{x}),y)\big] \\
&= \pi_+\mathbb{E}_{p_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)\big] + \pi_-\mathbb{E}_{p_-(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),-1)\big] \\
&= \frac{\pi_+\widetilde{\pi}}{\pi_+ - \pi_-\pi_+^2}\mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)\big] - \frac{\pi_+\pi_-\widetilde{\pi}}{\pi_+ - \pi_-\pi_+^2}\mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}),+1)\big] \qquad \text{(Lemma 1)} \\
&\qquad + \frac{\pi_-\widetilde{\pi}}{\pi_- - \pi_+\pi_-^2}\mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}),-1)\big] - \frac{\pi_+\pi_-\widetilde{\pi}}{\pi_- - \pi_+\pi_-^2}\mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),-1)\big] \\
&= \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1) - \pi_+\ell(f(\boldsymbol{x}),-1)\big] + \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}),-1) - \pi_-\ell(f(\boldsymbol{x}),+1)\big] \\
&= R_{\mathrm{PC}}(f),
\end{aligned}$$

which concludes the proof of Theorem 3. □

E. Proof of Theorem 4

First of all, we introduce the following notations:

$$\begin{aligned}
R^+_{\mathrm{PC}}(f) &= \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1) - \pi_+\ell(f(\boldsymbol{x}),-1)\big], \\
\widehat{R}^+_{\mathrm{PC}}(f) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(f(\boldsymbol{x}_i),+1) - \pi_+\ell(f(\boldsymbol{x}_i),-1)\Big), \\
R^-_{\mathrm{PC}}(f) &= \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}'),-1) - \pi_-\ell(f(\boldsymbol{x}'),+1)\big], \\
\widehat{R}^-_{\mathrm{PC}}(f) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(f(\boldsymbol{x}'_i),-1) - \pi_-\ell(f(\boldsymbol{x}'_i),+1)\Big).
\end{aligned}$$

In this way, we could simply represent $R_{\mathrm{PC}}(f)$ and $\widehat{R}_{\mathrm{PC}}(f)$ as

$$R_{\mathrm{PC}}(f) = R^+_{\mathrm{PC}}(f) + R^-_{\mathrm{PC}}(f), \quad \widehat{R}_{\mathrm{PC}}(f) = \widehat{R}^+_{\mathrm{PC}}(f) + \widehat{R}^-_{\mathrm{PC}}(f).$$

Then we have the following lemma.

Lemma 2. *The following inequality holds:*

$$R(\widehat{f}_{\mathrm{PC}}) - R(f^\star) \le 2\sup_{f\in\mathcal{F}}\left|R^+_{\mathrm{PC}}(f) - \widehat{R}^+_{\mathrm{PC}}(f)\right| + 2\sup_{f\in\mathcal{F}}\left|R^-_{\mathrm{PC}}(f) - \widehat{R}^-_{\mathrm{PC}}(f)\right|. \tag{9}$$

Proof. We could intuitively express $R(\widehat{f}_{\mathrm{PC}}) - R(f^\star)$ as

$$\begin{aligned}
R(\widehat{f}_{\mathrm{PC}}) - R(f^\star) &= R(\widehat{f}_{\mathrm{PC}}) - \widehat{R}_{\mathrm{PC}}(\widehat{f}_{\mathrm{PC}}) + \widehat{R}_{\mathrm{PC}}(\widehat{f}_{\mathrm{PC}}) - \widehat{R}_{\mathrm{PC}}(f^\star) + \widehat{R}_{\mathrm{PC}}(f^\star) - R(f^\star) \\
&= R_{\mathrm{PC}}(\widehat{f}_{\mathrm{PC}}) - \widehat{R}_{\mathrm{PC}}(\widehat{f}_{\mathrm{PC}}) + \widehat{R}_{\mathrm{PC}}(\widehat{f}_{\mathrm{PC}}) - \widehat{R}_{\mathrm{PC}}(f^\star) + \widehat{R}_{\mathrm{PC}}(f^\star) - R_{\mathrm{PC}}(f^\star) \\
&\le \sup_{f\in\mathcal{F}}\left|R_{\mathrm{PC}}(f) - \widehat{R}_{\mathrm{PC}}(f)\right| + 0 + \sup_{f\in\mathcal{F}}\left|R_{\mathrm{PC}}(f) - \widehat{R}_{\mathrm{PC}}(f)\right| \\
&= 2\sup_{f\in\mathcal{F}}\left|R_{\mathrm{PC}}(f) - \widehat{R}_{\mathrm{PC}}(f)\right| \\
&\le 2\sup_{f\in\mathcal{F}}\left|R^+_{\mathrm{PC}}(f) - \widehat{R}^+_{\mathrm{PC}}(f)\right| + 2\sup_{f\in\mathcal{F}}\left|R^-_{\mathrm{PC}}(f) - \widehat{R}^-_{\mathrm{PC}}(f)\right|,
\end{aligned}$$

where the second equality holds due to Theorem 3. □

As suggested by Lemma 2, we need to further upper bound the right hand size of Eq. (9). Before doing that, we introduce the *uniform deviation bound*, which is useful to derive estimation error bounds. The proof can be found in some textbooks such as (Mohri et al., 2012) (Theorem 3.1).

Lemma 3. *Let Z be a random variable drawn from a probability distribution with density μ, $\mathcal{H} = \{h : \mathcal{Z} \mapsto [0, M]\}$ ($M > 0$) be a class of measurable functions, $\{z_i\}_{i=1}^n$ be i.i.d. examples drawn from the distribution with density μ. Then, for any $delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{h\in\mathcal{H}} \left| \mathbb{E}_{Z\sim\mu}\big[h(Z)\big] - \frac{1}{n}\sum_{i=1}^{n} h(z_i) \right| \leq 2\mathfrak{R}_n(\mathcal{H}) + M\sqrt{\frac{\log\frac{2}{\delta}}{2n}},$$

where $\mathfrak{R}_n(\mathcal{H})$ denotes the (expected) Rademacher complexity *(Bartlett & Mendelson, 2002) of $\mathcal{H}$ with sample size n over μ.*

Lemma 4. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_{b} such that $\|f\|_\infty \leq C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t=\pm1} \ell(C_{\mathrm{b}}, t)$. For any $\delta > 0$, with probability $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} \left| R^+_{\mathrm{PC}}(f) - \widehat{R}^+_{\mathrm{PC}}(f) \right| \leq (1+\pi_+)2\rho\widetilde{\mathfrak{R}}^+_n(\mathcal{F}) + (1+\pi_+)C_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2n}}.$$

Proof. By the definition of $R^+_{\mathrm{PC}}(f)$ and $\widehat{R}^+_{\mathrm{PC}}(f)$, we can obtain

$$\begin{aligned}\sup_{f\in\mathcal{F}} \left| R^+_{\mathrm{PC}}(f) - \widehat{R}^+_{\mathrm{PC}}(f) \right| \leq \sup_{f\in\mathcal{F}} &\left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)\big] - \frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}),+1) \right| \\ &+ \pi_+ \sup_{f\in\mathcal{F}} \left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),-1)\big] - \frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}),-1) \right|. \end{aligned} \tag{10}$$

By applying Lemma 3, we have for any $\delta > 0$, with probability $1 - \delta$,

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)\big] - \frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}),+1) \right| \leq 2\widetilde{\mathfrak{R}}^+_n(\ell\circ\mathcal{F}) + C_\ell\sqrt{\frac{\log\frac{2}{\delta}}{2n}}, \tag{11}$$

and for any for any $\delta > 0$, with probability $1 - \delta$,

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),-1)\big] - \frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}),-1) \right| \leq 2\widetilde{\mathfrak{R}}^+_n(\ell\circ\mathcal{F}) + C_\ell\sqrt{\frac{\log\frac{2}{\delta}}{2n}}, \tag{12}$$

where $\ell \circ \mathcal{F}$ means $\{\ell \circ f \mid f \in \mathcal{F}\}$. By Talagrand's lemma (Lemma 4.2 in (Mohri et al., 2012)),

$$\widetilde{\mathfrak{R}}^+_n(\ell\circ\mathcal{F}) \leq \rho\widetilde{\mathfrak{R}}^+_n(\mathcal{F}). \tag{13}$$

Finally, by combing Eqs. (10), (11), (12), and (13), we have for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}} \left| R^+_{\mathrm{PC}}(f) - \widehat{R}^+_{\mathrm{PC}}(f) \right| \leq (1+\pi_+)2\rho\widetilde{\mathfrak{R}}^+_n(\mathcal{F}) + (1+\pi_+)C_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2n}}, \tag{14}$$

which concludes the proof of Lemma 4. □

Lemma 5. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_{b} such that $\|f\|_\infty \leq C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t=\pm1} \ell(C_{\mathrm{b}}, t)$. For any $\delta > 0$, with probability $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} \left| R^-_{\mathrm{PC}}(f) - \widehat{R}^-_{\mathrm{PC}}(f) \right| \leq (1+\pi_-)2\rho\widetilde{\mathfrak{R}}^-_n(\mathcal{F}) + (1+\pi_-)C_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2n}}.$$

Proof. Lemma 5 can be proved similarly to Lemma 4. □

By combining Lemma 2, Lemma 4, and Lemma 5, Theorem 4 is proved. □

F. Proof of Theorem 5

Suppose there are n pairs of paired data points, which means there are in total $2n$ data points. For our Pcomp classification problem, we could simply regard $\boldsymbol{x}$ sampled from $\widetilde{p}_+(\boldsymbol{x})$ as (noisy) positive data and $\boldsymbol{x}'$ sampled from $\widetilde{p}_-(\boldsymbol{x}')$ as (noisy) negative data. Given n pairs of examples $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}_{i=1}^n$, for the n observed positive examples, there are actually $n \cdot p(y = +1|\widetilde{y} = +1)$ true positive examples; for the n observed negative examples, there are actually $n \cdot p(y = -1|\widetilde{y} = -1)$ true negative examples. From our defined data generation process in Theorem 1, it is intuitive to obtain

$$p(y = +1 \mid \widetilde{y} = +1) = \frac{\pi_+^2 + \pi_+\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-} = \frac{\pi_+}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-},$$
$$p(y = -1 \mid \widetilde{y} = -1) = \frac{\pi_-^2 + \pi_+\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-} = \frac{\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-}.$$

Since $\phi_+ = p(y = -1 \mid \widetilde{y} = +1) = 1 - p(y = +1 \mid \widetilde{y} = +1)$ and $\phi_- = p(y = +1 \mid \widetilde{y} = -1) = 1 - p(y = -1 \mid \widetilde{y} = -1)$, we can obtain

$$\phi_+ = p(y = -1 \mid \widetilde{y} = +1) = 1 - \frac{\pi_+}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-} = \frac{\pi_-^2}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-},$$
$$\phi_- = p(y = +1 \mid \widetilde{y} = -1) = 1 - \frac{\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-} = \frac{\pi_+^2}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-}.$$

In this way, we can further obtain the following noise transition ratios:

$$\rho_+ = p(\widetilde{y} = -1 \mid y = +1) = \frac{p(y = +1 \mid \widetilde{y} = -1)p(\widetilde{y} = -1)}{p(y = +1 \mid \widetilde{y} = -1)p(\widetilde{y} = -1) + p(y = +1 \mid \widetilde{y} = +1)p(\widetilde{y} = +1)} = \frac{\pi_+}{1 + \pi_+},$$
$$\rho_- = p(\widetilde{y} = +1 \mid y = -1) = \frac{p(y = -1 \mid \widetilde{y} = +1)p(\widetilde{y} = +1)}{p(y = -1 \mid \widetilde{y} = +1)p(\widetilde{y} = +1) + p(y = -1 \mid \widetilde{y} = -1)p(\widetilde{y} = -1)} = \frac{\pi_-}{1 + \pi_-},$$

where $p(\widetilde{y} = 1) = p(\widetilde{y} = -1) = \frac{1}{2}$, because we have the same number of observed positive examples and negative examples.

G. Proof of Theorem 7

First of all, we introduce the following notations:

$$R_{\mathrm{pPC}}^+(f) = \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), +1)\mathbb{I}[\boldsymbol{x} \in \widetilde{\mathrm{P}}]\big],$$
$$\widehat{R}_{\mathrm{pPC}}^+(f) = \frac{1}{n}\sum_{i=1}^n \big(\ell(f(\boldsymbol{x}_i), +1)\mathbb{I}[\boldsymbol{x}_i \in \widetilde{\mathrm{P}}]\big),$$
$$R_{\mathrm{pPC}}^-(f) = \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}'), -1)\mathbb{I}[\boldsymbol{x}' \in \widetilde{\mathrm{N}}]\big],$$
$$\widehat{R}_{\mathrm{pPC}}^-(f) = \frac{1}{n}\sum_{i=1}^n \big(\ell(f(\boldsymbol{x}'_i), -1)\mathbb{I}[\boldsymbol{x}'_i \in \widetilde{\mathrm{N}}]\big).$$

In this way, we could simply represent $R_{\mathrm{ppc}}(f)$ and $\widehat{R}_{\mathrm{pPC}}(f)$ as

$$R_{\mathrm{pPC}}(f) = \frac{1}{1-\rho_+}R_{\mathrm{pPC}}^+(f) + \frac{1}{1-\rho_-}R_{\mathrm{pPC}}^-(f), \quad \widehat{R}_{\mathrm{pPC}}(f) = \frac{1}{1-\rho_+}\widehat{R}_{\mathrm{pPC}}^+(f) + \frac{1}{1-\rho_-}\widehat{R}_{\mathrm{pPC}}^-(f).$$

Then we have the following lemma.

Lemma 6. *The following inequality holds:*

$$R(\widehat{f}_{\mathrm{pPC}}) - R(f^\star) \leq \frac{2}{1-\rho_+}\sup_{f\in\mathcal{F}}\left|R_{\mathrm{pPC}}^+(f) - \widehat{R}_{\mathrm{pPC}}^+(f)\right| + \frac{2}{1-\rho_-}\sup_{f\in\mathcal{F}}\left|R_{\mathrm{pPC}}^-(f) - \widehat{R}_{\mathrm{pPC}}^-(f)\right|. \tag{15}$$

Proof. We omit the proof of Lemma 6 since it is quite similar to that of Lemma 2. □

As suggested by Lemma 6, we need to further upper bound the right hand size of Eq. (15). According to Lemma 3, we have the following two lemmas.

Lemma 7. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_{b} such that $\|f\|_\infty \leq C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{z \leq C_{\mathrm{b}}, t=\pm 1} \ell(z,t)$. For any $\delta > 0$, with probability $1-\delta$,*

$$\sup_{f \in \mathcal{F}} \left| R^+_{\mathrm{pPC}}(f) - \widehat{R}^+_{\mathrm{pPC}}(f) \right| \leq 2\rho \widetilde{\mathfrak{R}}^+_n(\mathcal{F}) + C_\ell \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Lemma 8. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_{b} such that $\|f\|_\infty \leq C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{z \leq C_{\mathrm{b}}, t=\pm 1} \ell(z,t)$. For any $\delta > 0$, with probability $1-\delta$,*

$$\sup_{f \in \mathcal{F}} \left| R^-_{\mathrm{pPC}}(f) - \widehat{R}^-_{\mathrm{pPC}}(f) \right| \leq 2\rho \widetilde{\mathfrak{R}}^-_n(\mathcal{F}) + C_\ell \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

We omit the proofs of Lemma 7 and Lemma 8 since they are similar to that of Lemma 4.

By combing Lemma 6, Lemma 7, and Lemma 8, Theorem 7 is proved.

H. Supplementary Information of Experiments

We report the detailed information of the used datasets as follows.

MNIST[1] (LeCun et al., 1998). This is a grayscale image dataset composed of handwritten digits from 0 to 9 where the size of the each image is 28×28. It contains 60,000 training images and 10,000 test images. Because the original dataset has 10 classes, we regard the even digits as the positive class and the odd digits as the negative class. We generate 30,000 pointwise corrupted examples from MNIST for model training.

Fashion-MNIST[2] (Xiao et al., 2017). Similarly to MNIST, this is also a grayscale image dataset composed of fashion items ('T-shirt', 'trouser', 'pullover', 'dress', 'sandal', 'coat', 'shirt', 'sneaker', 'bag', and 'ankle boot'). It contains 60,000 training examples and 10,000 test examples. It is converted into a binary classification dataset as follows:

- The positive class is formed by 'T-shirt', 'pullover', 'coat', 'shirt', and 'bag'.
- The negative class is formed by 'trouser', 'dress', 'sandal', 'sneaker', and 'ankle boot'.

We generate 30,000 pointwise corrupted examples from Fashion-MNIST for model training.

Kuzushiji-MNIST[3] (Netzer et al., 2011). This is another grayscale image dataset that is similar to MNIST. It is a 10-class dataset of cursive Japanese ("Kuzushiji") characters. It consists of 60,000 training images and 10,000 test images. It is converted into a binary classification dataset as follows:

- The positive class is formed by 'o', 'su','na', 'ma', 're'.
- The negative class is formed by 'ki','tsu','ha', 'ya','wo'.

We generate 30,000 pointwise corrupted examples from Kuzushiji-MNIST for model training.

CIFAR-10[4] (Krizhevsky et al., 2009). This is also a color image dataset of 10 different objects ('airplane', 'bird', 'automobile', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', and 'truck'), where the size of each image is $32 \times 32 \times 3$. There are 5,000 training images and 1,000 test images per class. This dataset is converted into a binary classification dataset as follows:

[1] http://yann.lecun.com/exdb/mnist/
[2] https://github.com/zalandoresearch/fashion-mnist
[3] https://github.com/rois-codh/kmnist
[4] https://www.cs.toronto.edu/~kriz/cifar.html

- The positive class is formed by 'bird', 'deer', 'dog', 'frog', 'cat', and 'horse'.
- The negative class is formed by 'airplane', 'automobile', 'ship', and 'truck'.

We generate 30,000 pointwise corrupted examples from CIFAR-10 for model training.

USPS, Pendigits, Optdigits. These datasets are composed of handwritten digits from 0 to 9. Because each of the original datasets has 10 classes, we regard the even digits as the positive class and the odd digits as the negative class. We generate 4,000 pointwise corrupted examples from USPS (5,000 from Pendigits and 2,000 from Optdigits) for model training.

CNAE-9. This dataset contains 1,080 documents of free text business descriptions of Brazilian companies categorized into a subset of 9 categories cataloged in a table called National Classification of Economic Activities.

- The positive class is formed by '2', '4', '6' and '8'.
- The negative class is formed by '1', '3', '5', '7' and '9'.

We generate 400 pointwise corrupted examples from CNAE-9 for model training.

For the four datasets, USPS can be downloaded from the website of the late Sam Roweis[5], and the other three datasets can be downloaded from the UCI machine learning repository[6].

For MNIST, Kuzushiji-MNIST, and Fashion-MNIST, we set learning rate to $1e-3$ and weight decay to $1e-5$. For CIFAR-10, we set learning rate to $1e-3$ and weight decay to $1e-3$. For the four datasets including USPS, Pendigits, Optdigits, and CNAE-9, we search learning rate and weight decay from $\{1e-5, 1e-4, \ldots, 1e-1\}$ for all learning methods. For Pcomp-teacher, the regularization parameter is searched from $\{1e-3, 1e-2, \ldots, 1e+3\}$ and the exponential moving average decay is fixed at 0.97. Hyper-parameters for all learning methods are selected so as to maximize the accuracy of five-fold cross validation on the training set.

References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(11): 463–482, 2002.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[5] http://cs.nyu.edu/~roweis/data.html
[6] http://archive.ics.uci.edu/ml/