Supplementary Material for Mitigating Overfitting in Supervised Classification from Two Unlabeled Datasets: A Consistent Risk Correction Approach

A Proofs

In this appendix, we prove all theorems.

A.1 Proof of Lemma 2

Let

$$p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}}) = p_{\mathrm{tr}}(x_1)\cdots p_{\mathrm{tr}}(x_n), \quad p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}}) = p'_{\mathrm{tr}}(x'_1)\cdots p'_{\mathrm{tr}}(x'_{n'})$$

be the probability density functions of $\mathcal{X}_{\mathrm{tr}}$ and $\mathcal{X}'_{\mathrm{tr}}$ (due to the i.i.d. sample assumption). Then, the measure of $\mathfrak{D}^-(g)$ is defined by

$$\Pr(\mathfrak{D}^-(g)) = \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}},$$

where Pr denotes the probability, $\mathrm{d}\mathcal{X}_{\mathrm{tr}} = \mathrm{d}x_1\cdots\mathrm{d}x_n$ and $\mathrm{d}\mathcal{X}'_{\mathrm{tr}} = \mathrm{d}x'_1\cdots\mathrm{d}x'_{n'}$. Since $\widehat{R}_{\mathrm{uu}}(g)$ is unbiased and $\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g) = 0$ on $\mathfrak{D}^+(g)$, the bias of $\widehat{R}_{\mathrm{cc}}(g)$ can be formulated as:

$$\begin{aligned}
\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)] - R(g) &= \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)] \\
&= \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^+(g)} \left(\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)\right) p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}} \\
&\quad + \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} \left(\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)\right) p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}} \\
&= \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} \left(\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)\right) p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}}
\end{aligned}$$

Thus we have $\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)] - R(g) > 0$ if and only if $\int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}} > 0$ due to the fact that $\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g) > 0$ on $\mathfrak{D}^-(g)$. That is, the bias of $\widehat{R}_{\mathrm{cc}}(g)$ is positive if and only if the measure of $\mathfrak{D}^-(g)$ is non-zero.

Next we study the probability measure of $\mathfrak{D}^-(g)$ by *the method of bounded differences*. Since $R^+_{\mathrm{p}}(g) \geq \alpha_g/\pi_{\mathrm{p}}$ and $R^-_{\mathrm{n}}(g) \geq \beta_g/\pi_{\mathrm{n}}$, then

$$\mathbb{E}[A - C] = \pi_{\mathrm{p}}R^+_{\mathrm{p}}(g) \geq \alpha_g, \quad \mathbb{E}[D - B] = \pi_{\mathrm{n}}R^-_{\mathrm{n}}(g) \geq \beta_g.$$

We have assumed that $0 \leq \ell(z) \leq C_\ell$, and thus the change of $a\widehat{R}^+_{\mathrm{u}}(g)$ and $b\widehat{R}^-_{\mathrm{u}}(g)$ will be no more than aC_ℓ/n and bC_ℓ/n if some $x_i \in \mathcal{X}_{\mathrm{tr}}$ is replaced, or the change of $c\widehat{R}^+_{\mathrm{u}'}(g)$ and $d\widehat{R}^-_{\mathrm{u}'}(g)$ will be no more than cC_ℓ/n' and dC_ℓ/n' if some $x'_j \in \mathcal{X}'_{\mathrm{tr}}$ is replaced. Subsequently, *McDiarmid's inequality* (McDiarmid, 1989) implies

$$\begin{aligned}
\Pr\{\pi_{\mathrm{p}}R^+_{\mathrm{p}}(g) - (A - C) \geq \alpha_g\} &\leq \exp\left(-\frac{2\alpha_g^2}{n(aC_\ell/n)^2 + n'(cC_\ell/n')^2}\right) \\
&= \exp\left(-\frac{2\alpha_g^2/C_\ell^2}{a^2/n + c^2/n'}\right),
\end{aligned}$$

and

$$\Pr\{\pi_{\mathrm{n}}R_{\mathrm{n}}^{-}(g)-(D-B)\geq\beta_g\}\leq\exp\left(-\frac{2\beta_g^2}{n'(dC_\ell/n')^2+n(bC_\ell/n)^2}\right)$$
$$=\exp\left(-\frac{2\beta_g^2/C_\ell^2}{b^2/n'+d^2/n}\right).$$

Then the probability measure of $\mathfrak{D}^{-}(g)$ can be bounded by

$$\begin{aligned}\Pr(\mathfrak{D}^{-}(g))&\leq\Pr\{A-C\leq 0\}+\Pr\{D-B<0\}\\&\leq\Pr\{A-C\leq\pi_{\mathrm{p}}R_{\mathrm{p}}^{+}(g)-\alpha_g\}+\Pr\{D-B\leq\pi_{\mathrm{n}}R_{\mathrm{n}}^{-}(g)-\beta_g\}\\&=\Pr\{\pi_{\mathrm{p}}R_{\mathrm{p}}^{+}(g)-(A-C)\geq\alpha_g\}+\Pr\{\pi_{\mathrm{n}}R_{\mathrm{n}}^{-}(g)-(D-B)\geq\beta_g\}\\&\leq\exp\left(-\frac{2\alpha_g^2/C_\ell^2}{a^2/n+c^2/n'}\right)+\exp\left(-\frac{2\beta_g^2/C_\ell^2}{b^2/n'+d^2/n}\right),\end{aligned}$$

we complete the proof. □

A.2 Proof of Theorem 3

Based on Lemma 2, we can show the exponential decay of the bias and also the consistency of the proposed non-negative risk estimator $\widehat{R}_{\mathrm{cc}}(g)$. It has been proved in Lemma 2 that

$$\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]-R(g)=\int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}\left(\widehat{R}_{\mathrm{cc}}(g)-\widehat{R}_{\mathrm{uu}}(g)\right)p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}}.$$

Therefore the exponential decay of the bias can be obtained via

$$\begin{aligned}\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]-R(g)&\leq\sup_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}\left(\widehat{R}_{\mathrm{cc}}(g)-\widehat{R}_{\mathrm{uu}}(g)\right)\cdot\int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}}\\&=\sup_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}(f_1(A-C)+f_2(D-B)-(A-C)-(D-B))\cdot\Pr(\mathfrak{D}^{-}(g))\\&\leq\sup_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}(|f_1(A-C)|+|f_2(D-B)|+|A-C|+|D-B|)\cdot\Pr(\mathfrak{D}^{-}(g))\\&\leq\sup_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}(L_f|A-C|+L_f|D-B|+|A-C|+|D-B|)\cdot\Pr(\mathfrak{D}^{-}(g))\\&=\sup_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}((L_f+1)|A-C|+(L_f+1)|D-B|)\cdot\Pr(\mathfrak{D}^{-}(g))\\&\leq\sup_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}((L_f+1)(a+c)C_\ell+(L_f+1)(d+b)C_\ell)\cdot\Pr(\mathfrak{D}^{-}(g))\\&=(L_f+1)(a+b+c+d)C_\ell\Delta_g,\end{aligned}$$

where we employed the Lipschitz condition, i.e., $|f_1(x)-f_1(y)|\leq L_f|x-y|$ (also holds for f_2), and the assumption $f(0)=0$ in Definition 1. Then the deviation bound (9) is due to

$$\begin{aligned}|\widehat{R}_{\mathrm{cc}}(g)-R(g)|&\leq|\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|+|\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]-R(g)|\\&\leq|\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|+(L_f+1)(a+b+c+d)C_\ell\Delta_g.\end{aligned}$$

Denote by A', B', C' and D' that differs from A, B ,C and D on a single example. Then

$$\begin{aligned}&|f_1(A-C)+f_2(D-B)-f_1(A'-C)-f_2(D-B')|\\&\leq|f_1(A-C)-f_1(A'-C)|+|f_2(D-B)-f_2(D-B')|\\&\leq L_f|A-C-A'+C|+L_f|D-B-D+B'|\\&=L_f|A-A'|+L_f|B'-B|\\&\leq(a+b)L_fC_\ell/n.\end{aligned}\tag{12}$$

Similarily, we can obtain

$$|f_1(A-C)+f_2(D-B)-f_1(A-C')-f_2(D'-B)|\leq(c+d)L_fC_\ell/n'.\tag{13}$$

Therefore the change of $\widehat{R}_{\mathrm{cc}}(g)$ will be no more than $(a+b)L_f C_\ell/n$ if some $x_i \in \mathcal{X}_{\mathrm{tr}}$ is replaced, or it will be no more than $(c+d)L_f C_\ell/n'$ if some $x'_j \in \mathcal{X}'_{\mathrm{tr}}$ is replaced, and McDiarmid's inequality gives us

$$\Pr\{|\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]| \geq \epsilon\} \leq 2\exp\left(-\frac{2\epsilon^2}{n((a+b)L_f C_\ell/n)^2 + n'((c+d)L_f C_\ell/n')^2}\right).$$

Setting the above right-hand side to be equal to δ and solving for ϵ yields immediately the following bound. For any $\delta > 0$, the following inequality holds with probability at least $1-\delta$,

$$\begin{aligned}|\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]| &\leq \sqrt{\frac{\ln(2/\delta)C_\ell^2 L_f^2}{2}\left(\frac{(a+b)^2}{n} + \frac{(c+d)^2}{n'}\right)} \\ &\leq C_\delta\left(\frac{(a+b)}{\sqrt{n}} + \frac{(c+d)}{\sqrt{n'}}\right) \\ &= C_\delta \cdot \chi_{n,n'},\end{aligned}$$

where $C_\delta = C_\ell L_f\sqrt{\ln(2/\delta)/2}$ and $\chi_{n,n'} = (a+b)/\sqrt{n} + (c+d)/\sqrt{n'}$. Thus we obtain

$$|\widehat{R}_{\mathrm{cc}}(g) - R(g)| \leq C_\delta \cdot \chi_{n,n'} + (L_f+1)(a+b+c+d)C_\ell\Delta_g.$$

On the other hand, the deviation bound (10) is due to

$$|\widehat{R}_{\mathrm{cc}}(g) - R(g)| \leq |\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)| + |\widehat{R}_{\mathrm{uu}}(g) - R(g)|,$$

where $|\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)| > 0$ with probability at most Δ_g, and $|\widehat{R}_{\mathrm{uu}}(g) - R(g)|$ shares the same concentration inequality with $|\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|$. □

A.3 Proof of Theorem 4

First, we introduce the definitions of Rademacher complexity.

Definition 6 (Rademacher complexity)**.** *Let $\mathcal{G} = \{g : \mathcal{Z} \to \mathbb{R}\}$ be a class of measurable functions, $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a fixed sample of size n i.i.d. drawn from a probability distribution p, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ be Rademacher variables, i.e., independent uniform random variables taking values in $\{-1, +1\}$. For any integer $n \geq 1$, the Rademacher complexity of $\mathcal{G}$ (Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014a) is defined as*

$$\mathfrak{R}_{n,p}(\mathcal{G}) = \mathbb{E}_{\mathcal{X}}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum\nolimits_{x_i\in\mathcal{X}}\varepsilon_i g(x_i)\right].$$

An alternative definition of the Rademacher complexity (Koltchinskii, 2001; Bartlett and Mendelson, 2002) will be used in the proof is:

$$\mathfrak{R}'_{n,p}(\mathcal{G}) = \mathbb{E}_{\mathcal{X}}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum\nolimits_{x_i\in\mathcal{X}}\varepsilon_i g(x_i)\right|\right].$$

Then, we list all the lemmas that will be used to derive the estimation error bound in Theorem 4.

Lemma 7. *For arbitrary $\mathcal{G}$, $\mathfrak{R}'_{n,p}(\mathcal{G}) \geq \mathfrak{R}_{n,p}(\mathcal{G})$; if $\mathcal{G}$ is closed under negation, $\mathfrak{R}'_{n,p}(\mathcal{G}) = \mathfrak{R}_{n,p}(\mathcal{G})$.*

Lemma 8 (Theorem 4.12 in Ledoux and Talagrand (1991))**.** *If $\psi : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous function with a Lipschitz constant L_ψ and satisfies $\psi(0) = 0$, we have*

$$\mathfrak{R}'_{n,p}(\psi \circ \mathcal{G}) \leq 2L_\psi\mathfrak{R}'_{n,p}(\mathcal{G}),$$

where $\psi \circ \mathcal{G} = \{\psi \circ g | g \in \mathcal{G}\}$ and $\circ$ is a composition operator.

Lemma 9. *Under the assumptions of Theorem 4, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\begin{aligned}\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{cc}}(g) - R(g)| \leq\ & 4(a+b)L_f L_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G}) + 4(c+d)L_f L_\ell\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}) \\ & + (L_f+1)(a+b+c+d)C_\ell\Delta + C'_\delta \cdot \chi_{n,n'}. \qquad (14)\end{aligned}$$

Proof. Firstly, we deal with the bias of $\widehat{R}_{\mathrm{cc}}(g)$. Noticing that the assumptions $\inf_{g\in\mathcal{G}} R_{\mathrm{p}}^{+}(g) \geq \alpha/\pi_{\mathrm{p}} > 0$ and $\inf_{g\in\mathcal{G}} R_{\mathrm{n}}^{-}(g) \geq \beta/\pi_{\mathrm{n}} > 0$ imply $\Delta = \sup_{g\in\mathcal{G}} \Delta_g$. By (8) we have:

$$\begin{aligned}\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - R(g)| &\leq \sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]| + \sup_{g\in\mathcal{G}} |\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)] - R(g)| \\ &\leq \sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]| + (L_f+1)(a+b+c+d)C_\ell\Delta. \end{aligned} \tag{15}$$

Secondly, we consider the double-sided uniform deviation $\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|$. Denote by $\mathcal{X}_{\mathrm{s}} = \{(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}})\}$, and $\mathcal{X}'_{\mathrm{s}}$ that differs from $\mathcal{X}_{\mathrm{s}}$ on a single example. Then we have

$$\begin{aligned}&\big|\sup_{g\in\mathcal{G}} \big|\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{s}}) - \mathbb{E}_{\mathcal{X}_{\mathrm{s}}}[\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{s}})]\big| - \sup_{g\in\mathcal{G}} \big|\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}'_{\mathrm{s}}) - \mathbb{E}_{\mathcal{X}'_{\mathrm{s}}}[\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}'_{\mathrm{s}})]\big|\big| \\ &\quad\leq \sup_{g\in\mathcal{G}} \big||\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{s}}) - \mathbb{E}_{\mathcal{X}_{\mathrm{s}}}[\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{s}})]| - |\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}'_{\mathrm{s}}) - \mathbb{E}_{\mathcal{X}'_{\mathrm{s}}}[\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}'_{\mathrm{s}})]|\big| \\ &\quad\leq \sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{s}}) - \widehat{R}_{\mathrm{cc}}(g;\mathcal{X}'_{\mathrm{s}})|,\end{aligned}$$

where we applied the *triangle inequality*. According to (12) and (13), we see that the change of $\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|$ will be no more than $(a+b)L_fC_\ell/n$ if some $x_i \in \mathcal{X}_{\mathrm{tr}}$ is replaced, or it will be no more than $(c+d)L_fC_\ell/n'$ if some $x'_j \in \mathcal{X}'_{\mathrm{tr}}$ is replaced. Similar to the proof technique of Theorem 3, by applying McDiarmid's inequality to the uniform deviation we have with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]| - \mathbb{E}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|] \\ &\quad\leq \sqrt{\frac{\ln(1/\delta)C_\ell^2L_f^2}{2}\left(\frac{(a+b)^2}{n} + \frac{(c+d)^2}{n'}\right)} \\ &\quad= C'_\delta \cdot \chi_{n,n'},\end{aligned} \tag{16}$$

where $C'_\delta = C_\ell L_f\sqrt{\ln(1/\delta)/2}$. Thirdly, we make *symmetrization* (Vapnik, 1998). Suppose that $(\mathcal{X}_{\mathrm{tr}}^{gh}, \mathcal{X}_{\mathrm{tr}}^{'gh})$ is a *ghost sample*, then

$$\begin{aligned}&\mathbb{E}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|] \\ &\quad= \mathbb{E}_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}}) - \mathbb{E}_{(\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})}\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})|] \\ &\quad\leq \mathbb{E}_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})}[\sup_{g\in\mathcal{G}} \mathbb{E}_{(\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})}|\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}}) - \widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})|] \\ &\quad\leq \mathbb{E}_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}}),(\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}}) - \widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})|],\end{aligned}$$

where we applied *Jensen's inequality* twice since the absolute value and the supremum are convex. By decomposing the difference $|\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}}) - \widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})|$, we can know that

$$\begin{aligned}&|\widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}}) - \widehat{R}_{\mathrm{cc}}(g;\mathcal{X}_{\mathrm{tr}}^{gh},\mathcal{X}_{\mathrm{tr}}^{'gh})| \\ &\quad= \Big|f_1\left(a\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}) - c\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}'_{\mathrm{tr}})\right) - f_1\left(a\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) - c\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right) \\ &\quad+ f_2\left(-b\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}) + d\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}'_{\mathrm{tr}})\right) - f_2\left(-b\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) + d\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right)\Big| \\ &\quad\leq \Big|f_1\left(a\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}) - c\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}'_{\mathrm{tr}})\right) - f_1\left(a\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) - c\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right)\Big| \\ &\quad+ \Big|f_2\left(-b\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}) + d\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}'_{\mathrm{tr}})\right) - f_2\left(-b\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) + d\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right)\Big| \\ &\quad\leq \Big|L_f\left(a\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}) - c\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}'_{\mathrm{tr}}) - a\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) + c\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right)\Big| \\ &\quad+ \Big|L_f\left(-b\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}) + d\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}'_{\mathrm{tr}}) + b\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) - d\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right)\Big| \\ &\quad\leq \Big|aL_f\left(\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}) - \widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh})\right)\Big| + \Big|cL_f\left(\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}'_{\mathrm{tr}}) - \widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right)\Big| \\ &\quad+ \Big|bL_f\left(\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}) - \widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{gh})\right)\Big| + \Big|dL_f\left(\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}'_{\mathrm{tr}}) - \widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\right)\Big|,\end{aligned}$$

where we employed the Lipschitz condition. This decomposition results in

$$\begin{aligned}\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|] \leq\ & aL_f\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh})\Big)\Big|\Big]\\ &+cL_f\mathbb{E}_{\mathcal{X}'_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{'gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}'_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\Big)\Big|\Big]\\ &+bL_f\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{gh})\Big)\Big|\Big]\\ &+dL_f\mathbb{E}_{\mathcal{X}'_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{'gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}'_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{'gh})\Big)\Big|\Big].\end{aligned}$$

Fourthly, we relax those expectations to Rademacher complexities. The original ℓ may miss the origin, i.e., $\ell(0,y)\neq 0$, with which we need to cope. Let

$$\bar{\ell}(t,y)=\ell(t,y)-\ell(0,y)$$

be a *shifted loss* so that $\bar{\ell}(0,y)=0$. Hence,

$$\begin{aligned}\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) &= (1/n)\textstyle\sum_{x_i\in\mathcal{X}_{\mathrm{tr}}}\ell(g(x_i),+1)-(1/n)\sum_{x_i^{gh}\in\mathcal{X}_{\mathrm{tr}}^{gh}}\ell(g(x_i^{gh}),+1)\\ &= (1/n)\textstyle\sum_{i=1}^{n}(\ell(g(x_i),+1)-\ell(g(x_i^{gh}),+1))\\ &= (1/n)\textstyle\sum_{i=1}^{n}(\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1)).\end{aligned}$$

This is already a standard form where we can attach Rademacher variables to every $\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1)$, so we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}[\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh})|]\\ &=\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}(\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1))\Big|\Big]\\ &=\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1))\Big|\Big]\\ &\leq\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i),+1)\Big|\Big]\\ &+\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i^{gh}),+1)\Big|\Big]\\ &=2\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i),+1)\Big|\Big]\\ &=2\mathfrak{R}'_{n,p_{\mathrm{tr}}}(\bar{\ell}(\cdot,+1)\circ\mathcal{G})\end{aligned}$$

The other three expectations can be handled analogously. As a result,

$$\begin{aligned}\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|] \leq\ & 2aL_f\mathfrak{R}'_{n,p_{\mathrm{tr}}}(\bar{\ell}(\cdot,+1)\circ\mathcal{G})+2cL_f\mathfrak{R}'_{n',p'_{\mathrm{tr}}}(\bar{\ell}(\cdot,+1)\circ\mathcal{G})\\ &+2bL_f\mathfrak{R}'_{n,p_{\mathrm{tr}}}(\bar{\ell}(\cdot,-1)\circ\mathcal{G})+2dL_f\mathfrak{R}'_{n',p'_{\mathrm{tr}}}(\bar{\ell}(\cdot,-1)\circ\mathcal{G}).\end{aligned}$$

Finally, we transform the Rademacher complexities of composite function classes to the original function class. It is obvious that $\bar{\ell}$ shares the same Lipschitz constant L_ℓ with ℓ, and consequently

$$\begin{aligned}\mathfrak{R}'_{n,p_{\mathrm{tr}}}(\bar{\ell}(\cdot,+1)\circ\mathcal{G}) &\leq 2L_\ell\mathfrak{R}'_{n,p_{\mathrm{tr}}}(\mathcal{G})=2L_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})\\ \mathfrak{R}'_{n',p'_{\mathrm{tr}}}(\bar{\ell}(\cdot,+1)\circ\mathcal{G}) &\leq 2L_\ell\mathfrak{R}'_{n',p'_{\mathrm{tr}}}(\mathcal{G})=2L_\ell\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G})\\ \mathfrak{R}'_{n,p_{\mathrm{tr}}}(\bar{\ell}(\cdot,-1)\circ\mathcal{G}) &\leq 2L_\ell\mathfrak{R}'_{n,p_{\mathrm{tr}}}(\mathcal{G})=2L_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})\\ \mathfrak{R}'_{n',p'_{\mathrm{tr}}}(\bar{\ell}(\cdot,-1)\circ\mathcal{G}) &\leq 2L_\ell\mathfrak{R}'_{n',p'_{\mathrm{tr}}}(\mathcal{G})=2L_\ell\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G})\end{aligned}$$

where we used the assumption that $\mathcal{G}$ is closed under negation, Lemma 7 and Lemma 8. So we have

$$\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|]\leq 4(a+b)L_fL_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})+4(c+d)L_fL_\ell\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}). \tag{17}$$

Combining (15), (16) and (17) finishes the proof of the uniform deviation bound (14). □

We are now ready to prove our estimation error bound based on the uniform deviation bound in Lemma 9.

$$\begin{aligned}R(\widehat{g}_{\mathrm{cc}})-R(g^*)&=\left(\widehat{R}_{\mathrm{cc}}(\widehat{g}_{\mathrm{cc}})-\widehat{R}_{\mathrm{cc}}(g^*)\right)+\left(R(\widehat{g}_{\mathrm{cc}})-\widehat{R}_{\mathrm{cc}}(\widehat{g}_{\mathrm{cc}})\right)+\left(\widehat{R}_{\mathrm{cc}}(g^*)-R(g^*)\right)\\&\leq 0+2\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{cc}}(g)-R(g)|\\&\leq 8(a+b)L_fL_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})+8(c+d)L_fL_\ell\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G})\\&\quad+2(L_f+1)(a+b+c+d)C_\ell\Delta+2C'_\delta\cdot\chi_{n,n'},\end{aligned}$$

where $\widehat{R}_{\mathrm{cc}}(\widehat{g}_{\mathrm{cc}})\leq\widehat{R}_{\mathrm{cc}}(g^*)$ by the definition of g^* and $\widehat{g}_{\mathrm{cc}}$. □

A.4 Proof of Corollary 5

We further get bounds on the Rademacher complexity of deep neural networks by the following Theorem.

Theorem 10 (Theorem 1 in Golowich et al. (2017)). *Assume the Frobenius norm of the weight matrices W_j are at most $M_F(j)$, and the activation function σ satisfying the assumption that it is 1-Lipschitz, positive-homogeneous which is applied element-wise (such as the ReLU). Let x is upper bounded by C_x. Then,*

$$\mathfrak{R}_{n,p}(\mathcal{G})\leq\frac{1}{n}\prod_{j=1}^{m}M_F(j)\cdot(\sqrt{2m\log 2}+1)\sqrt{\sum_{i=1}^{n}\|\mathbf{x}_i\|^2}\leq\frac{C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m}M_F(j)}{\sqrt{n}}. \tag{18}$$

Based on Theorem 10, we proved

$$\begin{aligned}R(\widehat{g}_{\mathrm{cc}})-R(g^*)&\leq 8(a+b)L_fL_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})+8(c+d)L_fL_\ell\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G})\\&\quad+2(L_f+1)(a+b+c+d)C_\ell\Delta+2C'_\delta\cdot\chi_{n,n'}\\&\leq 8(a+b)L_fL_\ell\frac{C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m}M_F(j)}{\sqrt{n}}\\&\quad+8(c+d)L_fL_\ell\frac{C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m}M_F(j)}{\sqrt{n'}}\\&\quad+2(L_f+1)(a+b+c+d)C_\ell\Delta+2C'_\delta\cdot\chi_{n,n'}\\&=\left(8L_fL_\ell C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m}M_F(j)+2C'_\delta\right)\cdot\chi_{n,n'}\\&\quad+2(L_f+1)(a+b+c+d)C_\ell\Delta.\end{aligned}$$

B Supplementary information on Figure 1

In Sec. 3.1, we illustrated the overfitting issue of state-of-the-art unbiased UU method using different datasets, different models, different optimizers and different loss functions. The details of these demonstration results are presented here.

In the upper row, the dataset used was MNIST and we artifically corrupt it into a binary classification dataset: even digits form the P class and odd digits form the N class. The models used were a linear-in-input model (Linear) $g(x)=\boldsymbol{\omega}^Tx+b$ where $\boldsymbol{\omega}\in\mathbb{R}^{784}$ and $b\in\mathbb{R}$, and a 5-layer *multi-layer perceptron* (MLP): d-300-300-300-300-1. And the optimizer was SGD with momentum (momentum=0.9) with logistic loss $\ell_{\log}(z)=\ln(1+\exp(-z))$ or sigmoid loss $\ell_{\mathrm{sig}}(z)=1/(1+\exp(z))$. For linear model experiments, the batch size was fixed to be 1000 and the initial learning rate was $5e-2$. For MLP model experiments, the batch size was fixed to be 3000 and the initial learning rate was $1e-3$.

In the bottom row, the dataset used was CIFAR-10 and we artifically corrupt it into a binary classification dataset: the P class is composed of 'bird', 'deer', 'dog', 'frog', 'ship' and 'truck', and the N class is composed of 'airplane', 'automobile', 'cat' and 'horse'. The models used were *all convolutional net* (AllConvNet) (Springenberg et al., 2015) as follows:

0th (input) layer:	(32*32*3)-
1st to 3rd layers:	[C(3*3, 96)]*2-C(3*3, 96, 2)-
4th to 6th layers:	[C(3*3, 192)]*2-C(3*3, 192, 2)-
7th to 9th layers:	C(3*3, 192)-C(1*1, 192)-C(1*1, 10)-
10th to 12th layers:	1000-1000-1

where C(3*3, 96) means 96 channels of 3*3 convolutions followed by ReLU, [·]*2 means 2 such layers, C(3*3, 96, 2) means a similar layer but with stride 2, etc; and a 32-layer *residual networks* (ResNet32) (He et al., 2016) as follows:

0th (input) layer:	(32*32*3)-
1st to 11th layers:	C(3*3, 16)-[C(3*3, 16), C(3*3, 16)]*5-
12th to 21st layers:	[C(3*3, 32), C(3*3, 32)]*5-
22nd to 31st layers:	[C(3*3, 64), C(3*3, 64)]*5-
32nd layer:	Global Average Pooling-1

where [·, ·] means a building block (He et al., 2016). Batch normalization (Ioffe and Szegedy, 2015) was applied before hidden layers. An ℓ_2-regularization was added, where the regularization parameter was fixed to 5e-3. The models were trained by Adam (Kingma and Ba, 2015) with the default momentum parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and the loss function was $\ell_{\mathrm{sig}}(z)$ or $\ell_{\log}(z)$. For AllConvNet experiments, the batch size and the learning rate were fixed to be 500 and $1e-5$ respectively. For ResNet32 experiments, the batch size and the learning rate were fixed to be 3000 and $3e-5$ respectively.

The two training distributions were created following (3) with class priors $\theta = 0.6$ and $\theta' = 0.4$. Subsequently, the two sets of U training data were sampled from those distributions with sample sizes $n = 30000$ and $n' = 30000$.

The results demonstrate the concurrence of empirical training risk going negative (blue dashed line) and the test accuracy overfitting (green dashed line) regardless of datasets, models, optimizers and loss functions.

C Supplementary information on the experiments

MNIST (LeCun et al., 1998) This is a grayscale image dataset of handwritten digits from 0 to 9 where the size of the images is 28*28. It contains 60,000 training images and 10,000 test images. See http://yann.lecun.com/exdb/mnist/ for details. Since it has 10 classes originally, we used the even digits as the P class and the odd digits as the N class, respectively.

The simple model used for training MNIST was a linear-in-input model $g(x) = \boldsymbol{\omega}^T x + b$ where $\boldsymbol{\omega} \in \mathbb{R}^{784}$ and $b \in \mathbb{R}$ with ℓ_2-regularization (the regularization parameter was fixed to be $1e-4$). The batch size and learning rate were set to be 3000 and $1e-3$ respectively. The deep model uased was a 5-layer FC with ReLU as the activation function: d-300-300-300-300-1 with ℓ_2-regularization (the regularization parameter was fixed to be $5e-3$). The batch size and learning rate were set to be 3000 and $5e-5$ respectively. For both models, batch normalization (Ioffe and Szegedy, 2015) with the default *momentum* $= 0.99$ and $\epsilon = 1e-3$ was applied before hidden layers, and the model was trained by Adam with the default momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For all the experiments, the generalized leaky ReLU hyperparameter λ was selected from -0.01 to 1.

Fashion-MNIST (Xiao et al., 2017) This is also a grayscale fashion image dataset similarly to MNIST, but here each data is associated with a label from 10 fashion item classes. See https://github.com/zalandoresearch/fashion-mnist for details. It was converted into a binary classification dataset as follows:

- the P class is formed by 'T-shirt', 'Trouser', 'Shirt', and 'Sneaker';
- the N class is formed by 'Pullover', 'Dress', 'Coat', 'Sandal', 'Bag', and 'Ankle boot'.

The models and optimizers were the same as MNIST, where the learning rate for the simple and deep models were set to be $5e-3$ and $3e-5$ and the other hyperparameters remain the same.

Kuzushiji-MNIST (Clanuwat et al., 2018) This is another variant of MNIST dataset consisting of 60,000 training images and 10,000 test images of cursive Japanese (Kuzushiji) characters. See https://github.com/rois-codh/kmnist for details. For Kuzushi-MNIST dataset,

- 'ki', 're', 'wo' made up the P class;
- 'o', 'su', 'tsu', 'na', 'ha', 'ma', 'ya' made up the N class.

The models and optimizers were the same as MNIST, where the learning rate for the deep models was set to be $3e-5$ and the other hyperparameters remain the same.

CIFAR-10 (Krizhevsky, 2009) This dataset consists of 60,000 32∗32 color images in 10 classes, and there are 5,000 training images and 1,000 test images per class. See https://www.cs.toronto.edu/~kriz/cifar.html for details. For CIFAR-10 dataset,

- the P class is composed of 'bird', 'deer', 'dog', 'frog', 'ship' and 'truck';
- the N class is composed of 'airplane', 'automobile', 'cat' and 'horse'.

The simple model used for training CIFAR-10 was also a linear-in-input model $g(x) = \boldsymbol{\omega}^T x + b$ where $\boldsymbol{\omega} \in \mathbb{R}^{3072}$ and $b \in \mathbb{R}$ with ℓ_2-regularization (the regularization parameter was fixed to be $5e-3$). The batch size and learning rate were set to be 3000 and $5e-3$ respectively. The deep model was again ResNet-32 (He et al., 2016) that can be find in Appendix B. The batch size and learning rate were set to be 3000 and $5e-3$ respectively. For both models, batch normalization (Ioffe and Szegedy, 2015) with the default $momentum = 0.99$ and $\epsilon = 1e-3$ was applied before hidden layers, and the model was trained by Adam with the default momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

D Supplementary experiments on general-purpose regularization

Regularization is the most common technique that lower the complexity of a neural network model during training, and thus prevent the overfitting (Goodfellow et al., 2016). In this section, we demonstrate that the general-purpose regularization methods fail to mitigate the overfitting in UU classification scenario.

We tested the unbiased UU method using two most popular regularization techniques, i.e., dropout and weight decay. The dataset and model used were again MNIST (the class priors θ and θ' were set to be 0.6 and 0.4) and the 5-layer MLP d-300-300-300-300-1 with ℓ_2-regularization, where dropout layers were added between the existing layers. The optimizer was SGD with momentum ($momentum = 0.9$) and logistic loss. Batch size was fixed to be 3000 and the initial learning rate was $1e-3$. For dropout experiments, we fixed the weight decay parameter to be $1e-4$ and change the dropout parameter from 0 to 0.8. For weight decay experiments, we fix the dropout parameter to be 0.2 and change the weight decay parameter from 0.0005 to 5.

Empirical results in Figure 3 show that the unbiased UU method with slightly strong regularizations outperforms the one with weaker regularizations, but still suffers from overfitting. It is because adding strong regularization may prohibit the high representation power of deep models, which in turn may cause underfitting.

Discussion Instead of the general-purpose regularization, our proposed method explicitly utilizes the additional knowledge that the empirical risk goes negative. By that, we can more "effectively" constrain the model without too much sacrificing the representation power of deep models. Note that our correction can also be regarded as a regularization in its general sense for fighting against overfitting, but differently from weight decay or dropout, it is exclusively designed for UU classification and hence it is no surprising that our regularization fits UU classification better than other regularizations as discussed in Sec. 3 theoretically and demonstrated in Sec. 4 empirically.

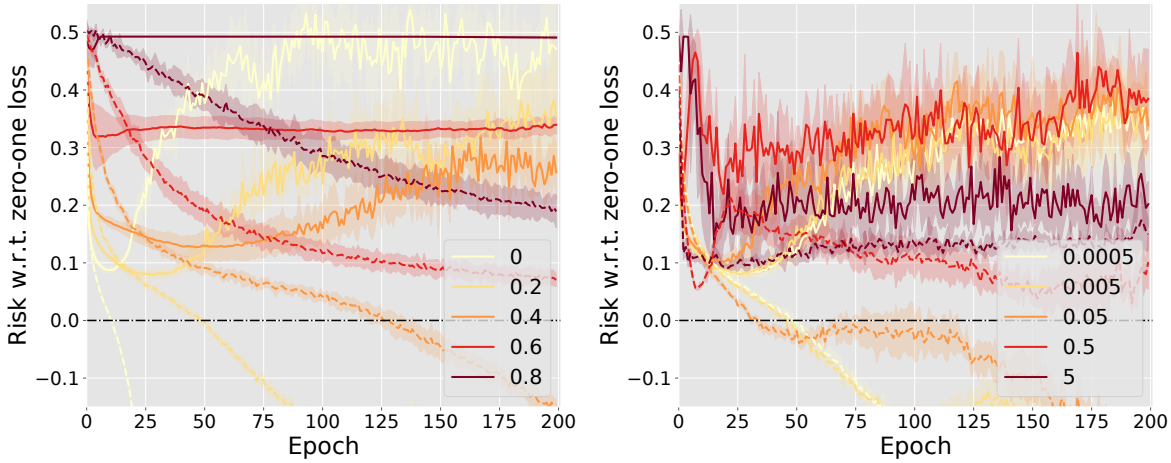

Figure 3: Supplementary experimental results on general regularization. Left: dropout. Right: weight decay. Solid curves are $\widehat{R}_{\mathrm{uu}}(g)$ on test data and dashed curves are $\widehat{R}_{\mathrm{uu}}(g)$ on training data.