Accurate and Robust Positive-Unlabeled Learning against Adversarial Perturbations

Ryo Shibazaki[1*], Kazuhiko Kawamoto[2] and Hiroshi Kera[2]

[1*]Graduate School of Science and Engineering, Chiba University, 1-33 Yayoichō, Inage-ku, Chiba-shi, 263-8522, Chiba, Japan.

[2]Graduate School of Informatics, Chiba University, 1-33 Yayoichō, Inage-ku, Chiba-shi, 263-8522, Chiba, Japan.

*Corresponding author(s). E-mail(s): ryo.shibazaki0517@chiba-u.jp;
Contributing authors: kawa@faculty.chiba-u.jp; kera@chiba-u.jp;

Abstract

Labeling costs are high in domains such as medical image analysis, where Positive and Unlabeled (PU) learning, which trains using only positive and unlabeled data, is effective. Medical images often contain small perturbations due to sensor noise and variations in acquisition conditions, which can cause a classifier to misclassify images that should be positive as negative. Therefore, in settings where even minor misclassifications may lead to critical misdiagnoses, high robustness is required. In this study, we focus on adversarial perturbations, which are known as worst-case noise among such perturbations, and aim to improve robustness within the PU learning framework. However, directly applying standard adversarial training methods to PU learning often severely degrades standard accuracy, making the trade-off between robustness and standard accuracy more pronounced. To address this issue, we propose PU-TRADES, a new learning method that extends the TRADES framework and integrates it with PU learning. Our method introduces label-independent adversarial perturbations and optimizes the balance between robustness and standard accuracy by combining a PU loss with a Kullback–Leibler loss. Furthermore, we theoretically derive an upper bound on the estimation error for the proposed loss and clarify conditions under which PU learning can outperform supervised learning when the number of unlabeled samples is sufficiently large. Finally, experiments on multiple benchmark datasets and a medical imaging dataset demonstrate that the proposed method provides an effective framework for robust learning in PU settings.

Keywords: positive-unlabeled learning, adversarial robustness, risk estimation, empirical risk minimization

1 Introduction

In recent years, machine learning has achieved remarkable success across a wide range of tasks, driven by advances in large-scale data and high-capacity models. However, in real-world applications, it is often difficult to obtain high-quality ...; moreover, in settings where the training set contains only positive and unlabeled data, one must learn from incomplete information in which the unlabeled set is a mixture of true positives and negatives. Therefore, unlike fully supervised learning, PU learning requires careful risk estimation and additional techniques to stabilize training.

Furthermore, in practical applications, robustness is an essential requirement in addition to label scarcity. In particular, medical images and sensor data are affected by variations in acquisition conditions, noise, and device ... We focus on adversarial perturbations[**?**] and aim to improve robustness within the PU learning framework.

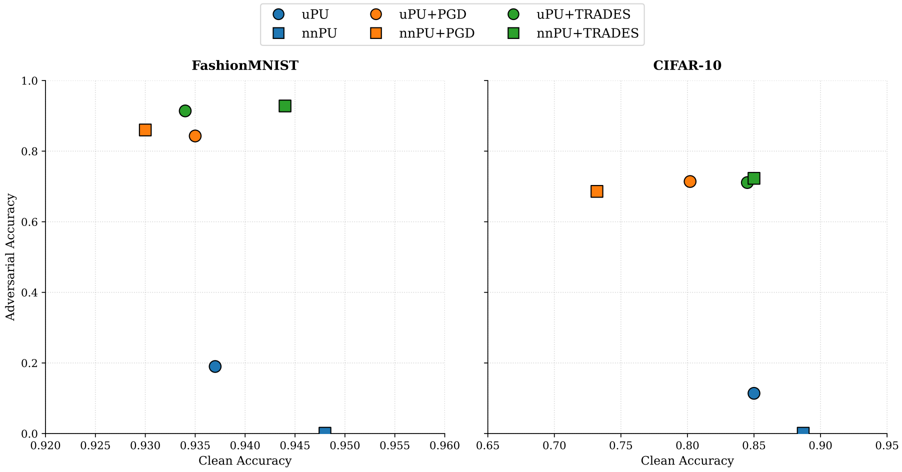
However, in preliminary experiments, directly applying adversarial training[**?**]—a standard approach to enhancing resilience against adversarial perturbations—to PU learning can substantially degrade classification performance on clean data. This is because, although unlabeled data contain a mixture of true positives and negatives, they are treated uniformly as negatives in the loss. As a result, the objectives of adversarial perturbation optimization and classification become misaligned, which destabilizes the updates associated with the unlabeled term. Figure 1 illustrates the relationship between clean accuracy and adversarial accuracy when PGD-based adversarial training is naively applied to PU learning. In particular, on CIFAR-10, adversarial accuracy improves while clean accuracy drops significantly. These observations indicate that, under the PU setting, introducing adversarial training may impair clean accuracy, which remains a key challenge.

In this study, we extend the TRADES framework[**?**], a representative adversarial training method, and propose a new learning method, PU-TRADES, by integrating it with PU learning. Our approach handles adversarial perturbations in a label-independent manner and combines a PU loss with a Kullback–Leibler loss, aiming to improve robustness while suppressing degradation in clean accuracy.

To validate the effectiveness of the proposed method, we conducted experiments on multiple benchmark datasets and medical imaging data. We evaluated both clean accuracy and ... and compared them with baseline methods. The results show that the proposed method substantially improves robustness to adversarial perturbations while maintaining accuracy on clean data.

In addition, we performed a theoretical analysis to better understand robustness in PU learning. Specifically, under binary classification with linear classifiers and adversarial perturbations, we derived upper bounds on the estimation error of the risks minimized by supervised learning and PU learning, thereby clarifying conditions under which PU learning can be advantageous over supervised learning. These conditions are consistent with practical scenarios, and they suggest that simply increasing the amount of unlabeled data can potentially achieve higher robustness than supervised learning.

Our contributions are summarized as follows.



	FashionMNIST			CIFAR-10		
Method	Clean	Adv	ΔClean	Clean	Adv	ΔClean
uPU	0.937	0.190	–	0.850	0.114	–
nnPU	**0.948**	0.001	–	**0.887**	0.001	–
uPU+PGD	0.935	0.843	-0.002	0.802	0.714	-0.048
nnPU+PGD	0.930	0.860	-0.018	0.732	0.686	**-0.155**

Fig. 1 Performance changes when PGD-based adversarial training is naively applied to PU learning (FashionMNIST / CIFAR-10). Left: a scatter plot showing the relationship between clean accuracy and adversarial accuracy for each method. Right: a numerical summary and the change in clean accuracy, ΔClean, before and after applying PGD (difference from the corresponding PU method without PGD). While PGD substantially improves adversarial accuracy, it degrades clean accuracy; this degradation is particularly pronounced for nnPU on CIFAR-10, where clean accuracy drops from 0.887 to 0.732 (ΔClean$= -0.155$).

- We propose a new learning framework (PU-TRADES) that integrates TRADES-style regularization into PU learning, improving robustness to adversarial perturbations while maintaining classification accuracy on clean samples.
- Assuming linear classifiers under adversarial perturbations, we derive upper bounds on the estimation error of the risks minimized by supervised learning and PU learning, and theoretically identify conditions under which PU learning becomes more favorable than supervised learning.
- Through experiments on benchmark datasets and medical imaging data, we empirically demonstrate that the proposed method acquires robustness to adversarial samples while preserving clean accuracy.

2 Related Work

In this chapter, we review prior studies related to this work, focusing on (i) Positive-Unlabeled (PU) learning and (ii) adversarial training. We then summarize existing research that combines PU learning with adversarial robustness.

2.1 Positive-Unlabeled (PU) Learning

PU learning is a classification framework in which only positive-labeled and unlabeled data are available. A representative line of work is risk-estimation-based PU learning, which constructs an (unbiased) estimator of the supervised classification risk using the class prior and the mixture structure of the unlabeled set. In particular, uPU (unbiased PU learning) estimates the true risk without bias, while nnPU (non-negative PU learning) introduces a non-negativity constraint to prevent the empirical risk from becoming negative, thereby mitigating overfitting. Many extensions have also been proposed, including methods that exploit high-confidence samples from the unlabeled set and approaches that incorporate various correction mechanisms.

2.2 Adversarial Training

Adversarial examples are inputs that are intentionally perturbed to cause a model to misclassify, and they have attracted extensive attention as a major threat to machine learning systems. A standard defense is adversarial training, which improves robustness by training the model on adversarially perturbed samples. Representative methods such as PGD-based adversarial training can be formulated as a min–max optimization problem consisting of an outer minimization (parameter optimization) and an inner maximization (perturbation generation). In addition, TRADES [**?**] is a prominent method that introduces a regularization term based on the Kullback–Leibler divergence between the model outputs on clean and perturbed inputs, aiming to balance clean accuracy and adversarial robustness.

3 Preliminaries

In this chapter, we introduce PU learning, adversarial examples and representative attacks, and adversarial training. Hereafter, we refer to learning a binary classifier from fully labeled positive and negative data as Positive–Negative (PN) learning.

3.1 Positive-Unlabeled (PU) Learning

We denote the input space by $\mathcal{X} \subseteq \mathbb{R}^d$ and the label space by $\mathcal{Y} = \{-1, +1\}$. Let $p(\boldsymbol{x}, y)$ be the joint distribution over $(\mathcal{X}, \mathcal{Y})$. Let the total number of samples be $n \in \mathbb{N}$, and let n_{P} and n_{N} denote the numbers of positive (P) and negative (N) samples, respectively. Each set is represented as follows:

$$\begin{aligned} \mathscr{X}_{\mathrm{P}} &= \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{P}}(\boldsymbol{x}), \\ \mathscr{X}_{\mathrm{N}} &= \{\boldsymbol{x}_i^{\mathrm{N}}\}_{i=1}^{n_{\mathrm{N}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{N}}(\boldsymbol{x}). \end{aligned} \tag{1}$$

Here, $p_{\mathrm{P}}(\boldsymbol{x})$ and $p_{\mathrm{N}}(\boldsymbol{x})$ denote the class-conditional densities for the positive and negative classes, respectively. The full dataset $\mathscr{X} = \mathscr{X}_{\mathrm{P}} \cup \mathscr{X}_{\mathrm{N}}$ is written as

$$\begin{aligned} \mathscr{X} &= \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}), \\ p(\boldsymbol{x}) &= \pi_{\mathrm{P}}\, p_{\mathrm{P}}(\boldsymbol{x}) + \pi_{\mathrm{N}}\, p_{\mathrm{N}}(\boldsymbol{x}), \end{aligned} \tag{2}$$

where $\pi_{\mathrm{P}} = p(y = +1)$ and $\pi_{\mathrm{N}} = ...p(y = -1)$ are the class priors satisfying $\pi_{\mathrm{P}} + \pi_{\mathrm{N}} = 1$.

In PU learning, the training set consists of positive (P) samples and unlabeled (U) samples. Since the marginal distribution of unlabeled data is $p_{\mathrm{U}}(\boldsymbol{x}) = \pi_{\mathrm{P}} p_{\mathrm{P}}(\boldsymbol{x}) + \pi_{\mathrm{N}} p_{\mathrm{N}}(\boldsymbol{x})$, letting n_{U} be the number of unlabeled samples, the unlabeled set is given by

$$\mathscr{X}_{\mathrm{U}} = \{\boldsymbol{x}_i^{\mathrm{U}}\}_{i=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{U}}(\boldsymbol{x}) = p(\boldsymbol{x}). \tag{3}$$

That is, the unlabeled samples are drawn i.i.d. from the marginal distribution of inputs, which is a mixture of positive and negative class-conditional distributions.

Unbiased PU Learning (uPU). uPU assumes that the positive class prior π_{P} is known and estimates the negative risk indirectly from the unlabeled data. Specifically, it minimizes the following empirical risk:

$$\widehat{R}_{\mathrm{uPU}}(g) = \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{i=1}^{n_{\mathrm{P}}} \tilde{\ell}\big(g\big(\boldsymbol{x}_i^{\mathrm{P}}\big), +1\big) \; + \; \frac{1}{n_{\mathrm{U}}} \sum_{i=1}^{n_{\mathrm{U}}} \ell\big(g\big(\boldsymbol{x}_i^{\mathrm{U}}\big), -1\big). \tag{4}$$

Here, the composite loss $\tilde{\ell}(g(\boldsymbol{x}), y)$ is defined by $\tilde{\ell}(g(\boldsymbol{x}), y) = \ell(g(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), -y)$.

Non-Negative PU Learning (nnPU). nnPU was introduced to address the overfitting issue in uPU, where the empirical risk can take negative values[**?**]. Specifically, when the estimated term involving the negative risk in PU learning becomes negative, nnPU clips its negative contribution to zero, yielding a non-negative risk estimator that keeps the overall estimate bounded below by 0. This modification prevents the empirical risk from diverging to negative values during empirical minimization and enables stable training. The empirical risk of nnPU is defined as follows:

$$\begin{aligned} \widehat{R}_{\mathrm{nnPU}}(g) = {} & \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{i=1}^{n_{\mathrm{P}}} \ell\left(g\left(\boldsymbol{x}_i^{\mathrm{P}}\right), +1\right) \\ & + \max\left\{0, -\frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{i=1}^{n_{\mathrm{P}}} \ell\left(g\left(\boldsymbol{x}_i^{\mathrm{P}}\right), -1\right) + \frac{1}{n_{\mathrm{U}}} \sum_{i=1}^{n_{\mathrm{U}}} \ell\left(g\left(\boldsymbol{x}_i^{\mathrm{U}}\right), -1\right)\right\}. \end{aligned} \tag{5}$$

3.2 Adversarial Examples

Adversarial examples are inputs constructed by adding small, carefully designed perturbations to an image that a classifier originally classified correctly, thereby intentionally causing misclassification...w2014ExplainingAH. For example, Fig. 2 adds a tiny perturbation to a cat image and generates an adversarial example...

3.3 Adversarial attack: FGSM and PGD

We refer to methods for generating adversarial examples as *adversarial attacks*, and many variants have been studied. In this work, we focus on representative first-order attacks: the Fast Gradient Sign Method (FGSM) [**?**] and Projected Gradient Descent

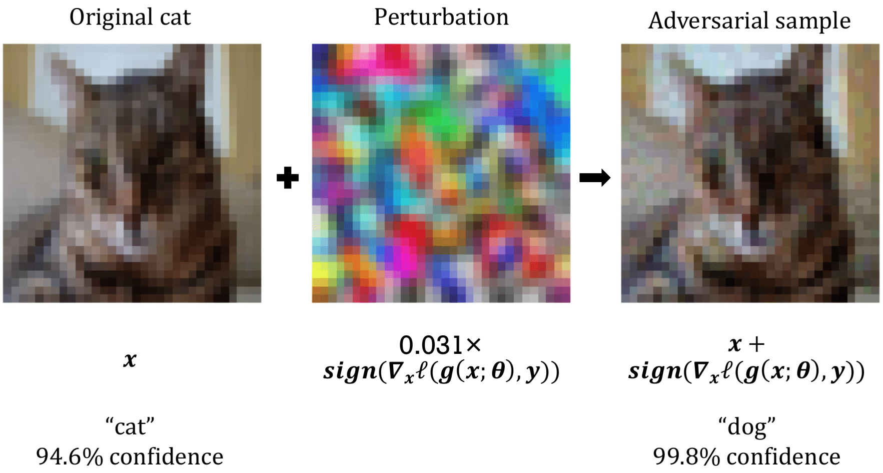

Original cat
Perturbation
Adversarial sample
$\boldsymbol{x}$
0.031×
$sign(\boldsymbol{\nabla}_x \ell(\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{y}))$
$\boldsymbol{x}+$
$sign(\boldsymbol{\nabla}_x \ell(\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{y}))$
“cat”
94.6% confidence
“dog”
99.8% confidence

Fig. 2 An example of generating adversarial examples. Starting from the clean image on the left, we add a small perturbation using PGD to obtain an adversarial input... The classifier correctly predicts the clean image as a cat (confidence 94.6%), while it misclassifies the adversarial example as a dog (confidence 99.8%). This illustrates that predictions can change drastically even when the input appears almost identical to humans[**?**].

(PGD) [**?**]. Here, sign : $\mathbb{R} \to [-1, 1]$ is applied element-wise to the argument vector.

$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \cdot \mathrm{sign}\left(\nabla_{\boldsymbol{x}} \ell(g(\boldsymbol{x}; \boldsymbol{\theta}), y)\right) \tag{6}$$

This produces an input that increases the loss when fed into the model. A stronger iterative variant of this method is PGD [**?**], which updates the input in the direction that increases the loss with step size α, similarly to FGSM, and then applies the projection $\Pi_{\dots}$. In particular,

$$\mathcal{B}_\infty(\boldsymbol{x}, \epsilon) := \{\boldsymbol{z} \in \mathbb{R}^d \mid \|\boldsymbol{z} - \boldsymbol{x}\|_\infty \le \epsilon\}$$

Then, $\Pi_{\mathcal{B}_\infty(\boldsymbol{x},\epsilon)}$ denotes the projection onto the ℓ_∞-ball, which guarantees $\|\boldsymbol{x}' - \boldsymbol{x}\|_\infty \le \epsilon$. Under this setting, the PGD update is given by:

$$\boldsymbol{x}' \leftarrow \Pi_{\mathcal{B}_\infty(\boldsymbol{x},\epsilon)} \left[\boldsymbol{x} + \alpha \ \mathrm{sign}\left(\nabla_{\boldsymbol{x}} \ell(g(\boldsymbol{x}; \boldsymbol{\theta}), y)\right)\right]. \tag{7}$$

By repeating Eq. (7) multiple times, we can generate samples that more strongly increase the loss within the ϵ-ball.

3.4 Adversarial Training

Adversarial training improves model robustness by training on adversarial examples[**?**]. It is typically formulated as the following min–max optimization problem: one first generates adversarial examples for each input, and then learns parameters that minimize the average loss over these adversarial inputs.

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\boldsymbol{x}_i' - \boldsymbol{x}_i\|_\infty \le \epsilon} \ell(g(\boldsymbol{x}_i'; \boldsymbol{\theta}), y_i) \,. \tag{8}$$

In addition, as a method to further enhance robustness against adversarial examples, TRADES (TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization), proposed by Zhang *et al.*[**?**], is a prominent approach. TRADES explicitly models the trade-off between accuracy on clean samples and robustness to adversarial samples, and aims to minimize the following loss function:

$$\begin{aligned} \mathcal{L}_{\mathrm{TR}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \Big[& \ell\big(g(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i\big) \\ & + \beta \cdot \max_{\|\boldsymbol{x}_i' - \boldsymbol{x}_i\|_\infty \le \epsilon} \ell_{\mathrm{KL}}\big(g(\boldsymbol{x}_i; \boldsymbol{\theta}), g(\boldsymbol{x}_i'; \boldsymbol{\theta})\big) \Big]. \end{aligned} \tag{9}$$

Here, $\ell_{\mathrm{KL}}(\cdot, \cdot)$ denotes the Kullback–Leibler (KL) divergence between predictive distributions, i.e.,

$$\ell_{\mathrm{KL}}\big(g(\boldsymbol{x}_i; \boldsymbol{\theta}),\, g(\boldsymbol{x}_i'; \boldsymbol{\theta})\big) := \mathrm{KL}\big(p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}_i) \;\big\|\; p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}_i')\big) \,.$$

The first term of Eq. (9), $\ell\big(g(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i\big)$, is the standard classification loss on the clean input $\boldsymbol{x}_i$.

On the other hand, the second term $\ell_{\mathrm{KL}}\big(g(\boldsymbol{x}_i; \boldsymbol{\theta}), g(\boldsymbol{x}_i'; \boldsymbol{\theta})\big)$ constrains the model so that the output distributions for $\boldsymbol{x}_i$ and its perturbed version $\boldsymbol{x}_i'$ are close, and this term plays a key role in improving robustness. Thus, TRADES is designed to enhance robustness while maintaining classification accuracy. In this study, we apply this framework to PU learning to achieve both high performance on clean samples and robustness to adversarial examples.

4 Accurate and Robust PU Learning

In this chapter, we first clarify the issues that arise when uPU learning is naively combined with PGD-based adversarial training. We then propose a new learning method, PU+TRADES, which adapts the TRADES framework to PU learning.

4.1 uPU+PGD

In uPU learning, the empirical risk $\widehat{R}_{\mathrm{uPU}}(g)$ is estimated by minimizing

$$\widehat{R}_{\mathrm{uPU}}(g) = \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}} \sum_{i=1}^{n_{\mathrm{P}}} \tilde{\ell}\left(g\big(\boldsymbol{x}_i^{\mathrm{P}}\big), +1\right) + \frac{1}{n_{\mathrm{U}}} \sum_{i=1}^{n_{\mathrm{U}}} \ell\left(g\big(\boldsymbol{x}_i^{\mathrm{U}}\big), -1\right). \tag{10}$$

Here, the loss is taken differently for P and U samples, which can be summarized as

$$\mathscr{L}(\boldsymbol{x}) := \begin{cases} \tilde{\ell}(g(\boldsymbol{x}), +1), & \boldsymbol{x} \in \mathscr{X}_{\mathrm{P}}, \\ \ell(g(\boldsymbol{x}), -1), & \boldsymbol{x} \in \mathscr{X}_{\mathrm{U}}. \end{cases} \tag{11}$$

Using this loss $\mathscr{L}$, we generate an adversarial example for each sample via PGD. A single PGD update step is given by

$$\boldsymbol{x}' \leftarrow \mathrm{Clip}_{(\boldsymbol{x}-\epsilon,\ \boldsymbol{x}+\epsilon)} \Big[\boldsymbol{x}' + \alpha\ \mathrm{sign}\big(\nabla_{\boldsymbol{x}'} \mathscr{L}(\boldsymbol{x}')\big)\Big]. \tag{12}$$

Issues with uPU+PGD

In uPU, the loss for unlabeled data is computed as if the label were always $y = -1$. However, in reality, the unlabeled set contains a mixture of positives and negatives. This property is incompatible with PGD-based adversarial training.

- **Negative U samples.** Since the loss $\ell(g(\boldsymbol{x}^{\mathrm{U}}), -1)$ is consistent with the true label, it pushes the input in a direction that increases the loss for the negative class. Consequently, PGD generates appropriate adversarial perturbations, contributing to improved robustness.
- **Positive U samples.** If perturbations are generated using $\ell(g(\boldsymbol{x}^{\mathrm{U}}), -1)$ even though the sample is truly positive, PGD updates the input so as to maximize the *negative-class* loss. As a result, the input may be pushed not toward the decision boundary, but rather toward a region where it is classified as positive with higher confidence. Therefore, PGD fails to produce perturbations in the "most misclassifiable direction," and the training can break down.

Hence, to generate adversarial perturbations appropriately in PU learning, it is essential to use a *label-independent* perturbation generation mechanism. This motivates PU+TRADES, introduced in the next section.

4.2 PU+TRADES

In this work, we propose **uPU+TRADES** and **nnPU+TRADES**, which integrate TRADES into uPU and nnPU, respectively. By introducing the TRADES framework into PU learning, we endow the model with robustness.

The objective function of uPU+TRADES is given by

$$\min_{g} \left[\widehat{R}_{\mathrm{uPU}}(g) + \beta \cdot \frac{1}{n} \sum_{i=1}^{n} \max_{\|\boldsymbol{x}' - \boldsymbol{x}_i\|_\infty \leq \epsilon} \ell_{\mathrm{KL}}\left(g\left(\boldsymbol{x}_i\right)\ \|\ g\left(\boldsymbol{x}'\right)\right) \right], \tag{13}$$

and the objective function of nnPU+TRADES is given by

$$\min_{g}\left[\widehat{R}_{\mathrm{nnPU}}(g)+\beta\cdot\frac{1}{n}\sum_{i=1}^{n}\max_{\|\boldsymbol{x}'-\boldsymbol{x}_i\|_\infty\leq\epsilon}\ell_{\mathrm{KL}}\left(g\left(\boldsymbol{x}_i\right)\ \|\ g\left(\boldsymbol{x}'\right)\right)\right]. \tag{14}$$

In binary classification, the network outputs a one-dimensional logit $g(\boldsymbol{x};\boldsymbol{\theta})$. We convert it into a Bernoulli probability vector and compute the KL divergence:

$$p(\boldsymbol{x})=\big[\sigma(g(\boldsymbol{x};\boldsymbol{\theta})),\,1-\sigma(g(\boldsymbol{x};\boldsymbol{\theta}))\big], \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The KL loss is then defined as

$$\begin{aligned}\ell_{\mathrm{KL}}\left(g\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right),g\left(\boldsymbol{x}_i';\boldsymbol{\theta}\right)\right)&=\mathrm{KL}\big(p(\boldsymbol{x}_i)\,\|\,p(\boldsymbol{x}_i')\big)\\&=\sum_{c\in\{0,1\}}p_c(\boldsymbol{x}_i)\,\log\frac{p_c(\boldsymbol{x}_i)}{p_c(\boldsymbol{x}_i')}.\end{aligned} \tag{16}$$

This term encourages the model outputs to remain stable under small perturbations of $\boldsymbol{x}_i$, thereby providing robustness against adversarial perturbations.

5 Appendix

5.1 Theoretical Analysis

In this chapter, we consider a binary classification problem where input data may be subject to adversarial perturbations, and derive upper bounds on the gap between the true adversarial risk and its empirical estimator for both supervised learning and PU learning. By comparing these bounds, we theoretically clarify conditions under which PU learning can be advantageous over supervised learning in the finite-sample regime.

Below we define the problem setting used throughout this chapter.

Problem Setting 5.1. (Adversarial Binary Classification Setting)

Let the input space be $\mathcal{X}\subseteq\mathbb{R}^d$, and the label space be $\mathcal{Y}=\{-1,+1\}$. Assume the input $\boldsymbol{x}\in\mathcal{X}$ is bounded, i.e., there exists a constant $C_x>0$ such that

$$\|\boldsymbol{x}\|_\infty\leq C_x$$

holds.

Assume $(\boldsymbol{x},y)$ is generated from a joint distribution $p(\boldsymbol{x},y)$, and define the class-conditional distributions as

$$p_{\mathrm{P}}(\boldsymbol{x})=p(\boldsymbol{x}\mid y=+1),\quad p_{\mathrm{N}}(\boldsymbol{x})=p(\boldsymbol{x}\mid y=-1)$$

as above. Let the class prior probabilities be $\pi_{\mathrm{P}}=p(y=+1)$ $\pi_{\mathrm{N}}=p(y=-1)$ and let with $\pi_{\mathrm{P}}+\pi_{\mathrm{N}}=1$.

In supervised learning, we use labeled data from the positive (P) and negative (N) classes. The corresponding datasets are

$$\begin{aligned}\mathscr{X}_{\mathrm{P}} &= \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{P}}(\boldsymbol{x}),\\ \mathscr{X}_{\mathrm{N}} &= \{\boldsymbol{x}_i^{\mathrm{N}}\}_{i=1}^{n_{\mathrm{N}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{N}}(\boldsymbol{x})\end{aligned} \tag{17}$$

given by

In PU learning, we use positive data and unlabeled (U) data. The marginal distribution of unlabeled data is

$$p_{\mathrm{U}}(\boldsymbol{x}) = \pi_{\mathrm{P}} p_{\mathrm{P}}(\boldsymbol{x}) + \pi_{\mathrm{N}} p_{\mathrm{N}}(\boldsymbol{x})$$

given by and the unlabeled dataset is

$$\mathscr{X}_{\mathrm{U}} = \{\boldsymbol{x}_i^{\mathrm{U}}\}_{i=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{U}}(\boldsymbol{x}) \tag{18}$$

.

Let the classifier be $g : \mathbb{R}^d \to \mathbb{R}$, and consider a linear classifier parameterized by a weight vector $\boldsymbol{w} \in \mathbb{R}^d$:

$$g(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$$

For $p \geq 1$ and $W > 0$, define the hypothesis class as

$$\mathscr{G} = \{g(\boldsymbol{x}) : \|\boldsymbol{w}\|_p \leq W\}$$

as above.

Let the loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ take as arguments the classifier output and the true label.

Moreover, as the adversarial regularization term in TRADES we use

$$\ell_{\mathrm{KL}}\big(g(\boldsymbol{x}), g(\boldsymbol{x}')\big)$$

where ℓ_{KL} denotes the Kullback–Leibler divergence between probability distributions induced by the classifier outputs.

Throughout this chapter, we assume that ℓ and ℓ_{KL} satisfy the regularity conditions needed for the analysis (boundedness and Lipschitz continuity); see the assumptions in Theorem5.2 for details.

5.2 Preliminaries (Notation and Assumptions)

To derive the estimation-error upper bounds in this chapter, we use the Rademacher complexity to bound the expected uniform deviation, and we use McDiarmid's inequality to obtain high-probability guarantees that hold with probability at least $1 - \delta$. Below we summarize the definitions and properties used in this chapter.

Rademacher Complexity

In our discussion, we need a measure of the complexity of a function class $\mathscr{G}$. We adopt the Rademacher complexity and recall its definition. A random variable σ satisfying $\Pr(\sigma = +1) = \Pr(\sigma = -1) = 1/2$ is called a Rademacher variable. Given a set of n-dimensional vectors $S \subseteq \mathbb{R}^n$, the Rademacher complexity of S is defined as

$$\mathfrak{R}(S) = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{(s_1,\ldots,s_n)\in S} \frac{1}{n}\sum_{i=1}^{n} \sigma_i s_i\right]$$

as above. It can be interpreted as the expected correlation between random noise and the best-matching element of S.

Next, let $S_n = \{\boldsymbol{x}_i\}_{i=1}^n$ be a dataset of size n. For a function class $\mathscr{G}$, define

$$\mathscr{G} \circ S_n = \{(g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_n)) \mid g \in \mathscr{G}\}$$

Then, the empirical Rademacher complexity of $\mathscr{G}$ on S_n is

$$\mathfrak{R}_{S_n}(\mathscr{G}) = \mathfrak{R}(\mathscr{G} \circ S_n) = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i)\right]$$

given by Furthermore, when $S_n = \{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \nu(\boldsymbol{x})$, we define the Rademacher complexity of $\mathscr{G}$ as follows (where $\nu(\boldsymbol{x})$ denotes the distribution of $\boldsymbol{x}$).

Definition 5.1 (Rademacher Complexity)**.** Let n be the sample size and let $S_n = \{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \nu(\boldsymbol{x})$. Then the Rademacher complexity of $\mathscr{G}$ is

$$\mathfrak{R}_{n,\nu}(\mathscr{G}) = \mathbb{E}_{S_n\sim\nu^n}\left[\mathfrak{R}_{S_n}(\mathscr{G})\right] = \mathbb{E}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i)\right] \tag{19}$$

is defined as above.

Standard results used in this chapter (Talagrand's contraction lemma, vector contraction, Rademacher complexity bounds for linear function classes, additional bounds for adversarial inputs, McDiarmid's inequality, etc.) are summarized in AppendixA.1. *(Proofs of theorems and lemmas are provided in Appendix A.)*

5.3 Upper Bound on the Estimation Error of Supervised TRADES

In this section, we derive an upper bound on the estimation error of supervised TRADES. The proof proceeds by (i) deriving an upper bound on the uniform deviation (lemma), and (ii) applying the standard ERM argument to obtain the estimation-error bound (theorem), following the standard flow.

Supervised TRADES risk (population risk and empirical risk)

We define the population risk of supervised TRADES as

$$\begin{aligned} R_{\text{PN-TR}}(g) := \pi_{\text{P}}\mathbb{E}_{\text{P}}&\left[\ell\big(g(\boldsymbol{x}), +1\big) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \le \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}), g(\boldsymbol{x}+\boldsymbol{\eta})\big)\right] \\ &+ \pi_{\text{N}}\mathbb{E}_{\text{N}}\left[\ell\big(g(\boldsymbol{x}), -1\big) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \le \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}), g(\boldsymbol{x}+\boldsymbol{\eta})\big)\right] \end{aligned} \tag{20}$$

as above. The corresponding empirical risk is

$$\begin{aligned} \widehat{R}_{\text{PN-TR}}(g) := \frac{\pi_{\text{P}}}{n_{\text{P}}}\sum_{i=1}^{n_{\text{P}}}&\left[\ell\big(g(\boldsymbol{x}_i^{\text{P}}), +1\big) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \le \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}_i^{\text{P}}), g(\boldsymbol{x}_i^{\text{P}}+\boldsymbol{\eta})\big)\right] \\ &+ \frac{\pi_{\text{N}}}{n_{\text{N}}}\sum_{i=1}^{n_{\text{N}}}\left[\ell\big(g(\boldsymbol{x}_i^{\text{N}}), -1\big) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \le \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}_i^{\text{N}}), g(\boldsymbol{x}_i^{\text{N}}+\boldsymbol{\eta})\big)\right] \end{aligned} \tag{21}$$

. Define the empirical risk minimizer as

$$\widehat{g}_{\text{PN-TR}} := \arg\min_{g\in\mathscr{G}} \widehat{R}_{\text{PN-TR}}(g)$$

Also, let

$$g^* \in \arg\min_{g\in\mathscr{G}} R_{\text{PN-TR}}(g)$$

be a population risk minimizer. With these definitions, we obtain the following upper bound on the estimation error of supervised TRADES.

Theorem 5.2 (Upper Bound on the Estimation Error of Supervised TRADES)**.** *Let a function class $\mathscr{G}$ be given. Assume the following:*

- ***(Boundedness of the loss)*** *One of the following holds:*
  - *There exists a constant $C_\ell > 0$ such that for any $\widehat{y} \in \mathbb{R}$ and $y \in \mathcal{Y}$, $\ell(\widehat{y}, y) \leq C_\ell$ holds; or*
  - *there exists a constant $C_g > 0$ such that $\|g\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} |g(\boldsymbol{x})| \leq C_g$ holds for any $g \in \mathscr{G}$, and for $|\widehat{y}| \leq C_g$, $\ell(\widehat{y}, y) \leq C_\ell$ holds.*
- ***(Lipschitz continuity of the loss)*** *$\ell(\widehat{y}, y)$ is L_ℓ-Lipschitz continuous with respect to $\widehat{y}$.*
- ***(Regularity of the KL term)*** *The TRADES regularization term $\ell_{\mathrm{KL}}\left(g(\boldsymbol{x}), g\left(\boldsymbol{x}'\right)\right)$ is L_{KL}-Lipschitz continuous in each argument, and is uniformly bounded: $\ell_{\mathrm{KL}}(u, v) \leq C_{\mathrm{KL}}$ holds.*

Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\begin{aligned} R_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}}) - R_{\text{PN-TR}}(g^*) \ &\leq 4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\Big(\pi_{\mathrm{P}}\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G}) + \pi_{\mathrm{N}}\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G})\Big) \\ &\quad + 8\beta L_{\mathrm{KL}}\, \varepsilon W d^{1/q}\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\right) \\ &\quad + \sqrt{2\ln\frac{2}{\delta}}\,(C_\ell + \beta C_{\mathrm{KL}})\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\right). \end{aligned} \tag{22}$$

Interpretation and implications (statistical convergence rate)

Under the linear-in-parameters model ($\|\boldsymbol{w}\|_p \leq W$) and bounded inputs, standard bounds in Appendix A.1 yield

$$\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G}) = \mathcal{O}\!\left(\frac{W}{\sqrt{n_{\mathrm{P}}}}\right), \qquad \mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G}) = \mathcal{O}\!\left(\frac{W}{\sqrt{n_{\mathrm{N}}}}\right)$$

Substituting this bound into Theorem 5.2 and absorbing constants such as β and ε, the estimation-error upper bound becomes

$$R_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}}) - R_{\text{PN-TR}}(g^*) = \mathcal{O}_p\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\right)$$

This implies an $\mathcal{O}_p\big(\pi_{\mathrm{P}}/\sqrt{n_{\mathrm{P}}} + \pi_{\mathrm{N}}/\sqrt{n_{\mathrm{N}}}\big)$ rate, and in particular the bound converges to 0 in probability as $n_{\mathrm{P}}, n_{\mathrm{N}} \to \infty$.

As preparation for Theorem 5.2, we first bound the uniform deviation for supervised TRADES.

Auxiliary lemmas

In what follows, to bound the Rademacher complexity terms arising from the adversarial component of TRADES, we use two auxiliary lemmas: Lemma A.4 and Lemma A.2. (Hereafter, q denotes the conjugate exponent of p (i.e., $1/p + 1/q = 1$).)

Lemma 5.3 (Upper Bound on the Uniform Deviation for Supervised TRADES)**.** *Under the above assumptions, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\begin{aligned}\sup_{g \in \mathscr{G}} \left| \widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g) \right| \leq 2\big(L_\ell + 4\beta L_{\text{KL}}\big)\Big(\pi_{\text{P}} \mathfrak{R}_{n_{\text{P}}, p_{\text{P}}}(\mathscr{G}) + \pi_{\text{N}} \mathfrak{R}_{n_{\text{N}}, p_{\text{N}}}(\mathscr{G})\Big) \\ + 4\beta L_{\text{KL}}\, \varepsilon W d^{1/q} \left(\frac{\pi_{\text{P}}}{\sqrt{n_{\text{P}}}} + \frac{\pi_{\text{N}}}{\sqrt{n_{\text{N}}}} \right) \\ + \sqrt{\frac{1}{2} \ln \frac{2}{\delta}}\, (C_\ell + \beta C_{\text{KL}}) \left(\frac{\pi_{\text{P}}}{\sqrt{n_{\text{P}}}} + \frac{\pi_{\text{N}}}{\sqrt{n_{\text{N}}}} \right).\end{aligned} \tag{23}$$

holds.

(Proof is given in Appendix A.2.)
(Proof is given in Appendix A.3.)

5.4 Upper Bound on the Estimation Error of uPU+TRADES

In this section, we consider the uPU+TRADES objective based on positive (P) and unlabeled (U) data, and derive an upper bound on its estimation error.

uPU+TRADES risk (population risk and empirical risk)

Using the composite loss $\tilde{\ell}(g(\boldsymbol{x}), y) = \ell(g(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), -y)$ we define the population risk of uPU+TRADES as

$$\begin{aligned} R_{\text{uPU-TR}}(g) := \pi_{\text{P}} \mathbb{E}_{\text{P}} \left[\tilde{\ell}(g(\boldsymbol{x}), +1) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \leq \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x} + \boldsymbol{\eta}), g(\boldsymbol{x})\big) \right] \\ + \mathbb{E}_{\text{U}} \left[\ell\big(g(\boldsymbol{x}), -1\big) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \leq \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x} + \boldsymbol{\eta}), g(\boldsymbol{x})\big) \right] \end{aligned} \tag{24}$$

as above. The corresponding empirical risk is

$$\begin{aligned} \widehat{R}_{\text{uPU-TR}}(g) := \frac{\pi_{\text{P}}}{n_{\text{P}}} \sum_{i=1}^{n_{\text{P}}} \left[\tilde{\ell}(g(\boldsymbol{x}_i^{\text{P}}), +1) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \leq \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}_i^{\text{P}} + \boldsymbol{\eta}), g(\boldsymbol{x}_i^{\text{P}})\big) \right] \\ + \frac{1}{n_{\text{U}}} \sum_{i=1}^{n_{\text{U}}} \left[\ell\big(g(\boldsymbol{x}_i^{\text{U}}), -1\big) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \leq \epsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}_i^{\text{U}} + \boldsymbol{\eta}), g(\boldsymbol{x}_i^{\text{U}})\big) \right] \end{aligned} \tag{25}$$

. Define the empirical risk minimizer as

$$\widehat{g}_{\text{uPU-TR}} := \arg\min_{g \in \mathscr{G}} \widehat{R}_{\text{uPU-TR}}(g)$$

Also, let

$$g^* \in \arg\min_{g \in \mathscr{G}} R_{\text{uPU-TR}}(g)$$

be a population risk minimizer.

Theorem 5.4 (Upper Bound on the Estimation Error of uPU+TRADES)**.** *Let a function class $\mathscr{G}$ be given. Assume the following (as in the previous subsection):*

- ***(Boundedness of the loss)*** *There exist constants $C_\ell, C_{\text{KL}} > 0$ such that for any $y \in \mathcal{Y}$ and any input, the following holds:*
  - *(Classification loss) One of the following holds:*
    - *There exists a constant $C_\ell > 0$ such that for any $\widehat{y} \in \mathbb{R}$, $\ell(\widehat{y}, y) \leq C_\ell$, or*
    - *there exists a constant $C_g > 0$ such that $\|g\|_\infty \leq C_g$ (for $g \in \mathscr{G}$) and for $|\widehat{y}| \leq C_g$, $\ell(\widehat{y}, y) \leq C_\ell$.*
  - *(TRADES term) For any u, v, $\ell_{\text{KL}}(u, v) \leq C_{\text{KL}}$*
- ***(Lipschitz continuity)*** *There exist constants $L_\ell, L_{\text{KL}} > 0$ such that*
  - *(Classification loss) $\ell(\widehat{y}, y)$ is L_ℓ-Lipschitz continuous with respect to $\widehat{y}$.*
  - *(TRADES term) $\ell_{\text{KL}}(u, v)$ is L_{KL}-Lipschitz continuous in each argument*

Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\begin{aligned} R_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}}) - R_{\text{uPU-TR}}(g^*) \;\leq\; & 8\pi_{\text{P}}\big(L_\ell + 2\beta L_{\text{KL}}\big)\mathfrak{R}_{n_{\text{P}},p_{\text{P}}}(\mathscr{G}) \\ & + 4\big(L_\ell + 4\beta L_{\text{KL}}\big)\mathfrak{R}_{n_{\text{U}},p_{\text{U}}}(\mathscr{G}) \\ & + 8\beta L_{\text{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\text{P}}}{\sqrt{n_{\text{P}}}} + \frac{1}{\sqrt{n_{\text{U}}}}\right) \\ & + \sqrt{2\ln\frac{2}{\delta}}\left(\frac{\pi_{\text{P}}(2C_\ell + \beta C_{\text{KL}})}{\sqrt{n_{\text{P}}}} + \frac{C_\ell + \beta C_{\text{KL}}}{\sqrt{n_{\text{U}}}}\right). \end{aligned} \tag{26}$$

Interpretation and implications (statistical convergence rate)

Under the linear-in-parameters model ($\|\boldsymbol{w}\|_p \leq W$) and bounded inputs, standard bounds in Appendix A.1 yield

$$\mathfrak{R}_{n_{\text{P}},p_{\text{P}}}(\mathscr{G}) = \mathcal{O}\left(\frac{W}{\sqrt{n_{\text{P}}}}\right), \qquad \mathfrak{R}_{n_{\text{U}},p_{\text{U}}}(\mathscr{G}) = \mathcal{O}\left(\frac{W}{\sqrt{n_{\text{U}}}}\right)$$

Substituting this bound into (26) and absorbing constants such as β and ε, the estimation-error upper bound becomes

$$R_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}}) - R_{\text{uPU-TR}}(g^*) = \mathcal{O}_p\left(\frac{\pi_{\text{P}}}{\sqrt{n_{\text{P}}}} + \frac{1}{\sqrt{n_{\text{U}}}}\right)$$

This yields an upper bound that converges to 0 in probability as $n_{\text{P}}, n_{\text{U}} \to \infty$.

As preparation for Theorem 5.4, we present a lemma bounding the uniform deviation for uPU+TRADES.

Auxiliary lemmas

Below we use the auxiliary Lemma A.4 (Rademacher increase under adversarial inputs) and Lemma A.2 (vector contraction).

Lemma 5.5 (Upper Bound on the Uniform Deviation for uPU+TRADES)**.** *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\begin{aligned}\sup_{g\in\mathscr{G}}\left|\widehat{R}_{\text{uPU-TR}}(g) - R_{\text{uPU-TR}}(g)\right| &\le 4\pi_{\text{P}}\left(L_\ell + 2\beta L_{\text{KL}}\right)\mathfrak{R}_{n_{\text{P}},p_{\text{P}}}(\mathscr{G}) \\ &\quad + 2\left(L_\ell + 4\beta L_{\text{KL}}\right)\mathfrak{R}_{n_{\text{U}},p_{\text{U}}}(\mathscr{G}) \\ &\quad + 4\beta L_{\text{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\text{P}}}{\sqrt{n_{\text{P}}}} + \frac{1}{\sqrt{n_{\text{U}}}}\right) \\ &\quad + \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\left(\frac{\pi_{\text{P}}(2C_\ell + \beta C_{\text{KL}})}{\sqrt{n_{\text{P}}}} + \frac{C_\ell + \beta C_{\text{KL}}}{\sqrt{n_{\text{U}}}}\right)\end{aligned} \tag{27}$$

(Proof is given in Appendix A.4.)
(Proof is given in Appendix A.5.)

5.5 Upper Bound on the Estimation Error of nnPU+TRADES

In this section, we derive an upper bound on the estimation error of nnPU+TRADES.

nnPU+TRADES risk (population risk and empirical risk)

First, define the adversarial regularization term in TRADES as

$$\psi(g, \boldsymbol{x}) := \max_{\|\boldsymbol{\eta}\|_\infty \le \varepsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}+\boldsymbol{\eta}),\, g(\boldsymbol{x})\big)$$

. We define the population risk of nnPU+TRADES by adding this regularization term to the nnPU (Kiryo et al., 2017) risk estimator:

$$\begin{aligned}R_{\text{nnPU-TR}}(g) := \pi_{\text{P}}\,\mathbb{E}_{\text{P}}\left[\ell\big(g(\boldsymbol{x}), +1\big) + \beta\,\psi(g, \boldsymbol{x})\right] \\ + \max\Big\{0,\, -\pi_{\text{P}}\,\mathbb{E}_{\text{P}}\left[\ell\big(g(\boldsymbol{x}), -1\big)\right] + \mathbb{E}_{\text{U}}\left[\ell\big(g(\boldsymbol{x}), -1\big)\right]\Big\} \\ + \beta\,\mathbb{E}_{\text{U}}[\psi(g, \boldsymbol{x})]\end{aligned} \tag{28}$$

as above. The corresponding empirical risk is

$$\begin{aligned}\widehat{R}_{\text{nnPU-TR}}(g) := \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\Big[\ell\big(g(\boldsymbol{x}_i^{\mathrm{P}}),+1\big)+\beta\,\psi(g,\boldsymbol{x}_i^{\mathrm{P}})\Big]& \\ +\max\left\{0,\,-\frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\ell\big(g(\boldsymbol{x}_i^{\mathrm{P}}),-1\big)+\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\ell\big(g(\boldsymbol{x}_i^{\mathrm{U}}),-1\big)\right\}& \\ +\frac{\beta}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\psi(g,\boldsymbol{x}_i^{\mathrm{U}})& \qquad (29)\end{aligned}$$

. Define the empirical risk minimizer as

$$\widehat{g}_{\text{nnPU-TR}} := \arg\min_{g\in\mathscr{G}} \widehat{R}_{\text{nnPU-TR}}(g)$$

Also, let

$$g^* \in \arg\min_{g\in\mathscr{G}} R_{\text{nnPU-TR}}(g)$$

be a population risk minimizer.

Theorem 5.6 (Upper Bound on the Estimation Error of nnPU+TRADES)**.** *Let a function class $\mathscr{G}$ be given. Assume the following (as in the previous subsection):*

- ***(Boundedness of the loss)*** *There exist constants $C_\ell, C_{\mathrm{KL}} > 0$ such that for any $y \in \mathcal{Y}$ and any input, the following holds:*
  - *(Classification loss) One of the following holds:*
    - *There exists a constant $C_\ell > 0$ such that for any $\widehat{y} \in \mathbb{R}$, $\ell(\widehat{y}, y) \le C_\ell$, or*
    - *there exists a constant $C_g > 0$ such that $\|g\|_\infty \le C_g$ (for $g \in \mathscr{G}$) and for $|\widehat{y}| \le C_g$, $\ell(\widehat{y}, y) \le C_\ell$.*
  - *(TRADES term) For any u, v, $\ell_{\mathrm{KL}}(u, v) \le C_{\mathrm{KL}}$*
- ***(Lipschitz continuity)*** *There exist constants $L_\ell, L_{\mathrm{KL}} > 0$ such that*
  - *(Classification loss) $\ell(\widehat{y}, y)$ is L_ℓ-Lipschitz continuous with respect to $\widehat{y}$.*
  - *(TRADES term) $\ell_{\mathrm{KL}}(u, v)$ is L_{KL}-Lipschitz continuous in each argument*

Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\begin{aligned}R_{\text{nnPU-TR}}(\widehat{g}_{\text{nnPU-TR}}) - R_{\text{nnPU-TR}}(g^*) \;&\le 8\pi_{\mathrm{P}}\big(L_\ell + 2\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G}) \\ &\quad + 4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{U}},p_{\mathrm{U}}}(\mathscr{G}) \\ &\quad + 8\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right) \\ &\quad + \sqrt{2\ln\frac{2}{\delta}}\left(\frac{\pi_{\mathrm{P}}(2C_\ell + \beta C_{\mathrm{KL}})}{\sqrt{n_{\mathrm{P}}}} + \frac{C_\ell + \beta C_{\mathrm{KL}}}{\sqrt{n_{\mathrm{U}}}}\right). \qquad (30)\end{aligned}$$

Interpretation and implications (statistical convergence rate)

Under the linear-in-parameters model ($\|\boldsymbol{w}\|_p \leq W$) and bounded inputs, standard bounds in Appendix A.1 yield

$$\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\mathscr{G}) = \mathcal{O}\left(\frac{W}{\sqrt{n_\mathrm{P}}}\right), \qquad \mathfrak{R}_{n_\mathrm{U},p_\mathrm{U}}(\mathscr{G}) = \mathcal{O}\left(\frac{W}{\sqrt{n_\mathrm{U}}}\right)$$

Substituting this bound into (30) and absorbing constants such as β and ε, the estimation-error upper bound becomes

$$R_{\text{nnPU-TR}}(\widehat{g}_{\text{nnPU-TR}}) - R_{\text{nnPU-TR}}(g^*) = \mathcal{O}_p\left(\frac{\pi_\mathrm{P}}{\sqrt{n_\mathrm{P}}} + \frac{1}{\sqrt{n_\mathrm{U}}}\right)$$

This yields an upper bound that converges to 0 in probability as $n_\mathrm{P}, n_\mathrm{U} \to \infty$.

As preparation for Theorem 5.6, we provide a lemma bounding the uniform deviation for nnPU+TRADES.

Auxiliary lemmas

Below we use the auxiliaryLemma A.4 (Rademacher increase under adversarial inputs) and Lemma A.2 (vector contraction).

Lemma 5.7 (Upper Bound on the Uniform Deviation for nnPU+TRADES)**.** *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\begin{aligned}\sup_{g\in\mathscr{G}} \left|\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\right| \leq\ & 4\pi_\mathrm{P}(L_\ell + 2\beta L_\mathrm{KL})\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\mathscr{G}) \\ & + 2(L_\ell + 4\beta L_\mathrm{KL})\mathfrak{R}_{n_\mathrm{U},p_\mathrm{U}}(\mathscr{G}) \\ & + 4\beta L_\mathrm{KL}\,\varepsilon W d^{1/q}\left(\frac{\pi_\mathrm{P}}{\sqrt{n_\mathrm{P}}} + \frac{1}{\sqrt{n_\mathrm{U}}}\right) \\ & + \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\left(\frac{\pi_\mathrm{P}(2C_\ell + \beta C_\mathrm{KL})}{\sqrt{n_\mathrm{P}}} + \frac{C_\ell + \beta C_\mathrm{KL}}{\sqrt{n_\mathrm{U}}}\right). \qquad (31)\end{aligned}$$

(Proof is given in Appendix A.6.)
(Proof is given in Appendix A.7.)

5.6 Sufficient Unlabeled Sample Size for PU+TRADES to Outperform Supervised TRADES

In this section, we compare the estimation-error upper bounds obtained in the previous subsections (Theorems 5.2, 5.4, and 5.6) and derive sufficient conditions on the unlabeled sample size n_{U} under which the bound for supervised TRADES (PN+TRADES) becomes larger than that for PU+TRADES (uPU+TRADES / nnPU+TRADES), i.e., conditions under which PU+TRADES can theoretically outperform supervised TRADES.

Rademacher complexity of the linear hypothesis class

Under the assumptions in this chapter ($\|\boldsymbol{x}\|_\infty \le C_x$ and $\|\boldsymbol{w}\|_p \le W$), for any distribution ν,

$$\mathfrak{R}_{n,\nu}(\mathscr{G}) = \mathbb{E}_{\boldsymbol{x}_{1:n}\sim\nu}\,\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\|\boldsymbol{w}\|_p\le W}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{w}^\top\boldsymbol{x}_i\Big] \;\le\; \frac{W\,\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|_q}{\sqrt{n}} \;\le\; \frac{WC_x\,d^{1/q}}{\sqrt{n}} \tag{32}$$

holds (where $1/p+1/q=1$). In the following, for notational clarity,

$$\kappa_\delta := \sqrt{2\ln\frac{2}{\delta}} \tag{33}$$

and we rewrite the estimation-error bounds into a $1/\sqrt{n}$ form for comparison. Moreover, to collect constant factors,

$$\Gamma_\delta := 4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\,WC_x d^{1/q} + 8\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q} + \kappa_\delta\,(C_\ell + \beta C_{\mathrm{KL}}) \tag{34}$$

.

(1) Comparing PN+TRADES and uPU+TRADES

First, applying (32) to Theorem 5.2 gives, with probability at least $1-\delta$,

$$R_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}}) - R_{\text{PN-TR}}(g^*) \;\le\; \Gamma_\delta\Big(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\Big) \tag{35}$$

is obtained.

Similarly, applying (32) to Theorem 5.4 yields, with probability at least $1-\delta$,

$$R_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}}) - R_{\text{uPU-TR}}(g^*) \;\le\; \frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}\Big(\Gamma_\delta + 4L_\ell\,WC_x d^{1/q} + \kappa_\delta C_\ell\Big) + \frac{\Gamma_\delta}{\sqrt{n_{\mathrm{U}}}} \tag{36}$$

follows.

Main result: sufficient unlabeled sample size

The following theorem, based on the comparison of the estimation-error upper bounds, provides a sufficient condition on the unlabeled sample size n_{U} under which PU+TRADES can outperform supervised TRADES.

Theorem 5.8 (Sufficient Unlabeled Sample Size for PU+TRADES to Outperform Supervised TRADES)**.** *We compare the estimation-error upper bounds in Theorems 5.2, 5.4, and 5.6. Using (32), we reduce them to $1/\sqrt{n}$-type forms, and we use Γ_δ and κ_δ defined in (34). Then*

$$\Gamma_\delta \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}} > \frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}\Big(4L_\ell W C_x d^{1/q} + \kappa_\delta C_\ell\Big) \tag{37}$$

assume it holds.

(i) PN+TRADES vs uPU+TRADES If

$$\boxed{n_{\mathrm{U}} > \left(\frac{\Gamma_\delta}{\Gamma_\delta \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}} - \frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}\big(4L_\ell W C_x d^{1/q} + \kappa_\delta C_\ell\big)}\right)^2} \tag{38}$$

as a comparison of the estimation-error upper bounds, the bound for PN+TRADES ((35)) is larger than that for uPU+TRADES ((36)).

(ii) PN+TRADES vs nnPU+TRADES Since the right-hand side of Theorem 5.6 is identical to that of Theorem 5.4, the same sufficient condition

$$\boxed{n_{\mathrm{U}} > \left(\frac{\Gamma_\delta}{\Gamma_\delta \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}} - \frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}\big(4L_\ell W C_x d^{1/q} + \kappa_\delta C_\ell\big)}\right)^2} \tag{39}$$

implies that the PN+TRADES estimation-error upper bound is larger than the nnPU+TRADES one.

(Proof is given in Appendix A.8.)

If condition (37) does not hold, the denominator on the right-hand side becomes non-positive, and thus there is no n_{U} satisfying the comparison inequality in (i); within the scope of comparisons based on the bounds derived in this chapter, we cannot claim that PU+TRADES outperforms supervised TRADES. On the other hand, when (37) holds, by (38) and (39), in the regime where the unlabeled sample size is sufficiently large, the superiority of PU+TRADES is theoretically guaranteed.

Experiments

In this chapter, we experimentally evaluate the effectiveness of PU+TRADES (uPU+TRADES and nnPU+TRADES), proposed in Chapter 4, on multiple benchmark datasets. Specifically, we compare our methods with existing approaches using clean accuracy (Clean Accuracy) and adversarial accuracy (Adversarial Accuracy) as evaluation metrics, and also analyze the effect of the TRADES coefficient β. Furthermore, we conduct additional experiments to examine how the theoretical threshold on the number of unlabeled samples, derived in Chapter 5.1, corresponds to empirical observations.

6 Experiments

6.1 Experimental Setup

Datasets.

We conduct evaluation experiments on four benchmark datasets: FashionMNIST (F-MNIST [?]), CIFAR-10 [?], CIFAR-100 [?], and the Alzheimer dataset[1]. For F-MNIST, we consider clothing-image classification; for CIFAR-10, we focus on identifying vehicle classes; for CIFAR-100, we target `organics`; and for Alzheimer, we aim at recognizing AD patients. The definition of the positive class, the number of positive samples n_{P}, the number of unlabeled samples n_{U}, the class prior π_{P}, and the model architectures for each dataset are summarized in Table **??**.

Evaluation metrics.

To compare the performance of each method, we report clean accuracy (Clean Accuracy) and adversarial accuracy (Adversarial Accuracy) on the test set. Accuracy is defined as

$$\text{Accuracy} = \frac{\#\ \text{correct predictions}}{\#\ \text{total samples}}.$$

Adversarial accuracy is computed as the classification accuracy on adversarial examples generated by PGD; the perturbation budget ϵ, step size α, and the number of iterations (10 steps) follow Table **??**.

Training details.

As listed in Table **??**, we use a 6-layer MLP [?] for F-MNIST, ResNet-18 [?] for CIFAR-10/100, and ResNet-50 [?] for Alzheimer. As baselines, for F-MNIST and CIFAR-10 we use uPU and nnPU, as well as PGD-based adversarial training (uPU-PGD and nnPU-PGD) and TRADES-based adversarial training (uPU+TRADES and nnPU+TRADES). For CIFAR-100 and Alzheimer, training is unstable due to uPU-specific overfitting; thus, we restrict evaluation to nnPU and nnPU+TRADES. We set the TRADES coefficient to $\beta \in \{6, 12, 18\}$, and other adversarial-perturbation settings (ϵ, α, and the number of PGD iterations) follow Table **??**.

Table 2 Main results on FashionMNIST (6-layer MLP) and CIFAR-10 (ResNet-18). We compare uPU/nnPU and their adversarially trained variants (PGD training and TRADES training), and report clean accuracy (Clean) and adversarial accuracy (Adv.) under PGD attacks (ϵ follows Table **??**, 10 steps) on the test set. For each setting, we select the epoch that achieves the highest Clean Accuracy. While uPU/nnPU without adversarial training achieve almost zero Adv., TRADES improves robustness without severely degrading clean accuracy.

Method	F-MNIST		CIFAR-10	
2-5	Clean	Adv.	Clean	Adv.
uPU	0.937	0.190	0.850	0.114
nnPU	**0.948**	0.001	**0.887**	0.001
uPU-PGD	0.935	0.843	0.802	0.714
nnPU-PGD	0.930	0.860	0.732	0.686
uPU+TRADES	0.934	0.914	0.845	0.711
nnPU+TRADES	0.944	**0.928**	0.850	**0.723**

Table 3 Main results on CIFAR-100 (ResNet-18) and Alzheimer MRI (ResNet-50) (nnPU variants only). We compare clean accuracy (Clean) and adversarial accuracy (Adv.) under PGD attacks (ϵ follows Table **??**, 10 steps) on the test set. For each setting, we select the epoch that achieves the highest Clean Accuracy. While nnPU alone attains almost zero Adv., nnPU+TRADES substantially improves robustness.

Method	CIFAR-100		Alzheimer	
2-5	Clean	Adv.	Clean	Adv.
nnPU	**0.680**	0.001	**0.683**	0.000
nnPU+TRADES	0.645	**0.450**	0.649	**0.409**

7 Main Results

Tables 2 and 3 report clean accuracy (Clean) and adversarial accuracy (Adv.) on each dataset. First, while uPU and nnPU achieve high clean accuracy, their adversarial accuracy is extremely low, indicating vulnerability to perturbations. In contrast, nnPU-PGD and (uPU/nnPU)+TRADES markedly improve adversarial accuracy, confirming gains in robustness.

Moreover, on F-MNIST and CIFAR-10, nnPU+TRADES consistently achieves higher adversarial accuracy than uPU+TRADES. On the other hand, methods with

[1] Dubey, S. *Alzheimer's Dataset* (Kaggle). https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images

Table 4 Ablation on the TRADES coefficient β (F-MNIST / CIFAR-10). We vary $\beta \in \{6, 12, 18\}$ and compare clean accuracy (Clean Acc.) and adversarial accuracy (Adv. Acc.) for uPU+TRADES and nnPU+TRADES. On FashionMNIST, performance changes only mildly with β, whereas on CIFAR-10 increasing β tends to improve adversarial accuracy at the cost of decreased clean accuracy.

Method	F-MNIST		CIFAR-10	
2-5	Clean Acc.	Adv. Acc.	Clean Acc.	Adv. Acc.
uPU+TRADES ($\beta = 6$)	0.939	0.861	0.846	0.711
uPU+TRADES ($\beta = 12$)	0.935	0.848	0.831	0.735
uPU+TRADES ($\beta = 18$)	0.937	0.870	0.809	0.735
nnPU+TRADES ($\beta = 6$)	**0.944**	**0.877**	**0.861**	0.723
nnPU+TRADES ($\beta = 12$)	0.933	0.875	0.845	0.740
nnPU+TRADES ($\beta = 18$)	0.930	0.871	0.836	**0.743**

adversarial training tend to suffer a drop in clean accuracy, revealing a trade-off between accuracy and robustness. For CIFAR-100 and Alzheimer (Table 3), nnPU+TRADES maintains non-trivial adversarial accuracy while retaining reasonable clean accuracy on both datasets.

Table 4 shows performance as we vary the TRADES coefficient β. On CIFAR-10, adversarial accuracy tends to increase as β becomes larger, while clean accuracy decreases step by step. For example, for uPU+TRADES, increasing β from 6 to 12 improves adversarial accuracy but reduces clean accuracy, and at $\beta = 18$ the adversarial accuracy appears to saturate. Similarly for nnPU+TRADES, increasing β improves adversarial accuracy but also induces a decrease in clean accuracy, indicating that β is a key factor controlling the accuracy–robustness trade-off.

On F-MNIST, the variation across β is smaller than on CIFAR-10, and in particular uPU+TRADES exhibits only limited changes in clean accuracy even as β increases. However, for nnPU+TRADES, setting β too large slightly degrades adversarial accuracy, suggesting that the optimal β may depend on the dataset and the learning scheme (uPU vs. nnPU).

8 Validation of Theoretical Analysis

Objective.

In Chapter 5.1, as a condition under which PU learning can become advantageous compared to supervised learning (PN), we compared the upper bounds on the estimation error of PN-TRADES and PU+TRADES, and derived a threshold on the number of unlabeled samples, $n_{\mathrm{U}}^{\star}$, which holds when the upper bound for PN-TRADES exceeds that of PU+TRADES (Theorem 5.8). The aim of this section is to examine to what extent this theoretical threshold aligns with experimental results as a qualitative guideline: "increasing the amount of unlabeled data can make PU methods preferable." Note that the theoretical threshold is a sufficient condition based on upper-bound comparison; therefore, we do not generally expect it to exactly match the empirical turning point.

Validation procedure.

We validate the theory in three steps: (i) substitute constants into the theoretical expression to obtain a numerical value of $n_{\mathrm{U}}^{\star}$; (ii) sweep n_{U} over discrete values and compare the test losses of PU+TRADES and PN-TRADES; (iii) compare the empirical turning point $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}}$ with $n_{\mathrm{U}}^{\star}$ and discuss reasons for the discrepancy.

Numerical instantiation of the theoretical threshold $n_{\mathrm{U}}^{\star}$.

The theoretical expression contains an upper bound on the weight norm W (i.e., $\|w\|_2 \leq W$), but the threshold used here is based on comparing estimation-error upper bounds. We define the norm upper bound of the perturbed input as

$$C_x^{\mathrm{adv}} := C_x + \varepsilon d^{1/q}, \qquad C_{\mathrm{KL}} := W C_x^{\mathrm{adv}} = W(C_x + \varepsilon d^{1/q}) \tag{40}$$

and approximate an upper bound of the logistic loss by $C_\ell \approx W C_x$. Then we can rewrite $\Gamma_\delta \approx W\,\overline{\Gamma}_\delta$, and factor out W from both sides of the comparison inequality. As a result, the numerical instantiation of the threshold simplifies to a form independent of W.

Below we show intermediate steps to obtain $\Gamma_\delta \approx W\,\overline{\Gamma}_\delta$. The quantity Γ_δ defined in Theorem 5.8 is

$$\begin{aligned}
\Gamma_\delta &= 4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\, W C_x d^{1/q} + 8\beta L_{\mathrm{KL}}\, \varepsilon W d^{1/q} + \kappa_\delta\,(C_\ell + \beta C_{\mathrm{KL}}) && (41)\\
&\approx 4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\, W C_x d^{1/q} + 8\beta L_{\mathrm{KL}}\, \varepsilon W d^{1/q} + \kappa_\delta\,(W C_x + \beta W C_x^{\mathrm{adv}}) && (42)\\
&= 4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\, W C_x d^{1/q} + 8\beta L_{\mathrm{KL}}\, \varepsilon W d^{1/q} + \kappa_\delta\, W\big(C_x + \beta C_x^{\mathrm{adv}}\big) && (43)\\
&= W\Big\{4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\, C_x d^{1/q} + 8\beta L_{\mathrm{KL}}\, \varepsilon d^{1/q} + \kappa_\delta\big(C_x + \beta C_x^{\mathrm{adv}}\big)\Big\} && (44)\\
&=: W\,\overline{\Gamma}_\delta. && (45)
\end{aligned}$$

Specifically, letting

$$\kappa_\delta := \sqrt{2\ln\frac{2}{\delta}}, \qquad \overline{\Gamma}_\delta := 4\big(L_\ell + 4\beta L_{\mathrm{KL}}\big) C_x d^{1/q} + 8\beta L_{\mathrm{KL}}\varepsilon d^{1/q} + \kappa_\delta\big(C_x + \beta C_x^{\mathrm{adv}}\big) \tag{46}$$

the feasibility condition (37) in Theorem 5.8 becomes

$$\overline{\Gamma}_\delta \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}} \;>\; \frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}\Big(4L_\ell\, C_x d^{1/q} + \kappa_\delta C_x\Big), \tag{47}$$

and under this condition we obtain

$$\boxed{\; n_{\mathrm{U}} \;>\; \left(\frac{\overline{\Gamma}_\delta}{\overline{\Gamma}_\delta \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}} - \frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}\big(4L_\ell\, C_x d^{1/q} + \kappa_\delta C_x\big)}\right)^2 \;} \tag{48}$$

When this condition is satisfied, comparing the estimation-error upper bounds of PN-TRADES and PU+TRADES shows that the upper bound for PN-TRADES exceeds

Table 5 Constants used to numerically instantiate the theoretical threshold $n_{\mathrm{U}}^{\star}$ derived in Theorem 5.8 (FashionMNIST / CIFAR-10). We follow Table **??** for the class priors and sample sizes, and summarize the input dimension d, norm bounds C_x, C_x^{adv}, Lipschitz constants L_ℓ, L_{KL}, confidence level δ, and so on.

Symbol	**Meaning**	**FashionMNIST**	**CIFAR-10**	**Remarks**
π_{P}	prior of the positive class	0.4	0.4	assumption
π_{N}	prior of the negative class	0.6	0.6	$\pi_{\mathrm{N}} = 1 - \pi_{\mathrm{P}}$
n_{P}	# positive samples	1000	1000	given setting
n_{N}	# negative samples	500	500	given setting
β	TRADES coefficient	6	6	representative value
ϵ	perturbation radius	0.3	0.031	experimental setting (attack radius)
q	norm index	2	2	dual norm of $p = 2$
d	input dimension	784	3072	$1 \times 28 \times 28$, $3 \times 32 \times 32$
C_x	input norm bound	$\sqrt{d}$	$\sqrt{d}$	$\|x\|_2 \le \sqrt{d}$
C_x^{adv}	perturbed-input bound	$C_x + \epsilon d^{1/q}$	$C_x + \epsilon d^{1/q}$	from (40)
L_ℓ	Lipschitz constant of loss	1	1	logistic loss
L_{KL}	Lipschitz constant of KL term	1	1	setting in Chapter 5.1
δ	confidence level	0.05	0.05	fixed

that of PU+TRADES. That is, in theory, beyond this threshold PU+TRADES (uPU/nnPU) can become preferable to PN-TRADES.

Constants and settings.

The main symbols and constants used for numerical instantiation are summarized in Table 5. In this section we use $q = 2$, and for the input norm bound we set $C_x = \sqrt{d}$ (i.e., $\|x\|_2 \le \sqrt{d}$). Since $d^{1/q} = \sqrt{d}$, we have

$$C_x d^{1/q} = \sqrt{d} \cdot \sqrt{d} = d. \tag{49}$$

Thus, we substitute $d = 784$ for F-MNIST and $d = 3072$ for CIFAR-10.

Substituting the above settings into (48), we obtain

$$n_{\mathrm{U}}^{\star}(\text{F-MNIST}) \approx 1443.37 \ (\Rightarrow \ n_{\mathrm{U}} \ge 1444),$$

$$n_{\mathrm{U}}^{\star}(\text{CIFAR-10}) \approx 1443.25 \ (\Rightarrow \ n_{\mathrm{U}} \ge 1444).$$

Experimental protocol for validation.

Since the theoretical comparison concerns estimation error (risk), in this section we use the loss on the test set as a proxy. Specifically, for each n_{U}, let $t^{\star}(n_{\mathrm{U}})$ denote the epoch achieving the highest Clean Accuracy, and we compare the sum of the clean loss and adversarial loss at that epoch:

$$\mathcal{L}_{\mathrm{sum}}(n_{\mathrm{U}}) := \mathcal{L}_{\mathrm{clean}}\big(t^{\star}(n_{\mathrm{U}}); n_{\mathrm{U}}\big) + \mathcal{L}_{\mathrm{adv}}\big(t^{\star}(n_{\mathrm{U}}); n_{\mathrm{U}}\big). \tag{50}$$

Table 6 Comparison between the theoretical threshold $n_{\mathrm{U}}^{\star}$ (Theorem 5.8) and the empirical turning point $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}}$. $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}}$ is estimated by linear interpolation from Fig. 7 as the value of n_{U} where the loss sums $\mathcal{L}_{\mathrm{sum}}$ of PU+TRADES and PN-TRADES coincide. "—" indicates that no intersection existed within the range of n_{U} evaluated in this study.

Setting	$n_{\mathrm{U}}^{\star}$ (theory)	$\widehat{n}_{\mathrm{U}}^{\mathrm{emp}}$ (emp.)
FMNIST / 6-layer MLP	1444	4788.14
FMNIST / Linear	1444	4247.13
CIFAR-10 / Linear	1444	2750.00
CIFAR-10 / ResNet-18	1444	—

Here, $\mathcal{L}_{\mathrm{clean}}$ is the clean loss on the test set, and $\mathcal{L}_{\mathrm{adv}}$ is the adversarial loss computed on adversarial examples generated from the same test set. We consider the following settings: (a) FMNIST: 6-layer MLP and a linear model; (b) CIFAR-10: a linear model and ResNet-18. For each setting, we train PU+TRADES and PN-TRADES and compare how $\mathcal{L}_{\mathrm{sum}}$ changes with n_{U}.

Empirical turning point.

To relate to the theoretical threshold $n_{\mathrm{U}}^{\star}$, we define the empirical turning point as the smallest n_{U} at which the sign of

$$\Delta(n_{\mathrm{U}}) := \mathcal{L}_{\mathrm{sum}}^{\mathrm{PU\text{-}TRADES}}(n_{\mathrm{U}}) - \mathcal{L}_{\mathrm{sum}}^{\mathrm{PN\text{-}TRADES}}(n_{\mathrm{U}})$$

becomes negative. Moreover, if the sign of Δ flips between $n_{\mathrm{U}} = U_k$ and U_{k+1}, we estimate the turning point by linear interpolation:

$$\widehat{n}_{\mathrm{U}}^{\mathrm{emp}} := U_k + \frac{-\Delta(U_k)}{\Delta(U_{k+1}) - \Delta(U_k)}(U_{k+1} - U_k). \tag{51}$$

Results.

Figure 7 shows the trajectories of $\mathcal{L}_{\mathrm{sum}}$ as a function of n_{U}. The orange curve represents $\mathcal{L}_{\mathrm{sum}}(n_{\mathrm{U}})$ for PU+TRADES, the dashed gray line is the baseline value for PN-TRADES, and the blue vertical line indicates the theoretical threshold $n_{\mathrm{U}}^{\star}$. When a crossing is observed within the sweep range, we mark $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}}$ estimated by (51) as a point in the figure.

Discussion.

From Fig. 7 and Table 6, for F-MNIST (both the linear model and the 6-layer MLP), $\mathcal{L}_{\mathrm{sum}}(n_{\mathrm{U}})$ decreases as n_{U} increases, and the intersection where the loss sums of PU+TRADES and PN-TRADES coincide is estimated as $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}} \approx 4247.13$ (linear model) and $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}} \approx 4788.14$ (6-layer MLP). These values are larger than the theoretical threshold $n_{\mathrm{U}}^{\star} = 1444$, which is consistent with the fact that $n_{\mathrm{U}}^{\star}$ is a sufficient condition derived from upper-bound comparison (and hence can be conservative).

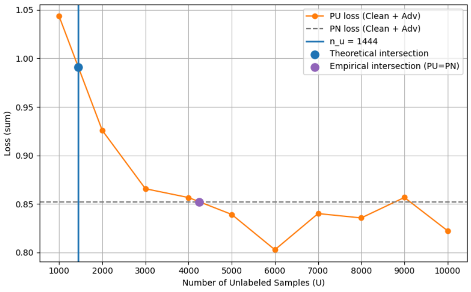


Fig. 3 *

(a) FMNIST / Linear

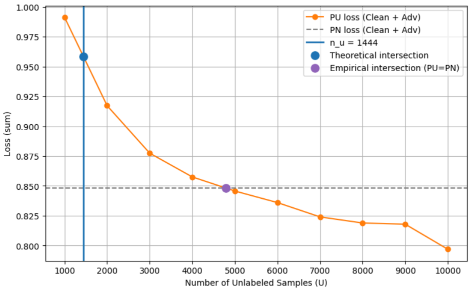


Fig. 4 *

(b) FMNIST / 6-layer MLP

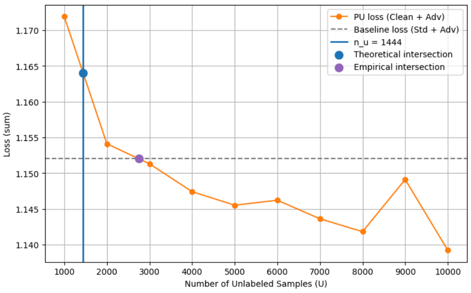


Fig. 5 *

(c) CIFAR-10 / Linear

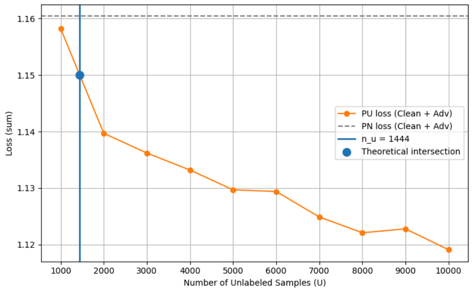


Fig. 6 *

(d) CIFAR-10 / ResNet-18

Fig. 7 Validation results for the theoretical threshold $n_{\mathrm{U}}^{\star}$ of the unlabeled sample size derived in our theoretical analysis. Each panel plots the test loss sum for PU+TRADES, $\mathcal{L}_{\mathrm{sum}}(n_{\mathrm{U}}) = \mathcal{L}_{\mathrm{clean}}(t^{\star}(n_{\mathrm{U}}); n_{\mathrm{U}}) + \mathcal{L}_{\mathrm{adv}}(t^{\star}(n_{\mathrm{U}}); n_{\mathrm{U}})$ (Eq. (50)), against n_{U}, and compares it with the PN-TRADES baseline value (which does not depend on n_{U}). Here, $t^{\star}(n_{\mathrm{U}})$ denotes the epoch achieving the highest Clean Accuracy. The intersection $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}}$ is estimated by linear interpolation.

For CIFAR-10, the results depend on the model: (i) **CIFAR-10 / Linear**: a crossing is observed at $\widehat{n}_{\mathrm{U}}^{\mathrm{emp}} \approx 2750.00$ (between $U = 2000$ and 3000). (ii) **CIFAR-10 / ResNet-18**: using the PN baseline value 1.1604, the PU loss sum remains smaller throughout the range $U = 1000$–10000, and thus there is no intersection within this range (i.e., PU is already better at the smallest observed value $U = 1000$). Overall, while the theoretical threshold $n_{\mathrm{U}}^{\star}$ is not an exact predictor of the intersection location, the experiments also indicate that with sufficiently many unlabeled samples, PU methods can outperform PN methods.

9 Conclusion

Conclusions may be used to restate your hypothesis or research question, restate your major findings, explain the relevance and the added value of your work, highlight any limitations of your study, describe future directions for research and recommendations.

In some disciplines use of Discussion or 'Conclusion' is interchangeable. It is not mandatory to use both. Please refer to Journal-level guidance for any specific requirements.

Supplementary information. If your article has accompanying supplementary file/s please state so here.

Authors reporting data from electrophoretic gels and blots should supply the full unprocessed scans for key as part of their Supplementary information. This may be requested by the editorial team/s if it is missing.

Please refer to Journal-level guidance for any specific requirements.

Acknowledgements. Acknowledgements are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

Appendix A Proofs of Theoretical Analysis

A.1 Background / Tool Box (Standard Results)

In this appendix, we summarize standard lemmas and inequalities used in the proofs of this chapter.

Lemma A.1 (Talagrand's Contraction Lemma). *Let $S_n = \{\boldsymbol{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \nu(\boldsymbol{x})$. Suppose $f : \mathbb{R} \to \mathbb{R}$ is L_f-Lipschitz. Then, for the Rademacher complexities of $\mathscr{G}$ and $f \circ \mathscr{G}$,*

$$\mathfrak{R}_{n,\nu}(f \circ \mathscr{G}) \leq L_f \, \mathfrak{R}_{n,\nu}(\mathscr{G}) \tag{A1}$$

holds.

Lemma A.2 (Vector Contraction). *If $\ell(u, v)$ is L_ℓ-Lipschitz in each argument, then for any pair of functions (f_h, g_h),*

$$\begin{aligned}\mathbb{E}_\sigma\Big[\sup_h \frac{1}{n}\sum_{i=1}^n \sigma_i \, \ell\big(f_h(\boldsymbol{x}_i),\, g_h(\boldsymbol{x}_i)\big)\Big] \leq 2L_\ell \Bigg\{ & \mathbb{E}_\sigma\Big[\sup_h \frac{1}{n}\sum_{i=1}^n \sigma_i f_h(\boldsymbol{x}_i)\Big] \\ & + \mathbb{E}_\sigma\Big[\sup_h \frac{1}{n}\sum_{i=1}^n \sigma_i g_h(\boldsymbol{x}_i)\Big] \Bigg\}.\end{aligned} \tag{A2}$$

Theorem A.3 (Upper Bound on Rademacher Complexity). *Assume the input satisfies $\|\boldsymbol{x}\|_\infty \leq C_x$, and let the class of linear classifiers be $\mathscr{G} = \{\boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} \mid \|\boldsymbol{w}\|_\infty \leq W\}$. Then,*

$$\mathfrak{R}_{n,\nu}(\mathscr{G}) \leq \frac{C_x W}{\sqrt{n}} \tag{A3}$$

holds.

Lemma A.4 (Adversarial Rademacher Additive Term)**.** *For the linear class* $\mathscr{G} = \{\boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} : \ \|\boldsymbol{w}\|_p \le W\}$*, for any distribution* ν *and any* $n \in \mathbb{N}$*,*

$$\mathfrak{R}_{n,\nu}\Big(\{\, \boldsymbol{x} \mapsto g(\boldsymbol{x}+\boldsymbol{\eta}) : \ \|\boldsymbol{\eta}\|_\infty \le \varepsilon, \ g \in \mathscr{G} \,\}\Big) \ \le \ \mathfrak{R}_{n,\nu}(\mathscr{G}) \ + \ \frac{\varepsilon W \, d^{1/q}}{\sqrt{n}}.$$

Theorem A.5 (McDiarmid's Inequality)**.** *Let* $X_1, \ldots, X_n$ *be independent random variables taking values in a set* $\mathcal{X}$*, and consider a function* $f : \mathcal{X}^n \to \mathbb{R}$*. Assume that there exist constants* $c_1, \ldots, c_n$ *such that, for each* $i = 1, \ldots, n$ *and any* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_i \in \mathcal{X}$*,*

$$|f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n) - f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}'_i, \ldots, \boldsymbol{x}_n)| \le c_i \tag{A4}$$

holds. Then, for any $t > 0$*,*

$$\Pr(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{A5}$$

holds.

A.2 Proof of the Lemma "Uniform Deviation Bound for Supervised TRADES" (Lemma 5.3)

Proof We divide the proof into the following steps (i)–(iv). We evaluate $\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\}$ and $\sup_{g\in\mathscr{G}}\{R_{\text{PN-TR}}(g) - \widehat{R}_{\text{PN-TR}}(g)\}$ with probability at least $1-\delta/2$, respectively, and then combine them via a union bound to obtain an upper bound on the absolute value. Below, we only show $\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\}$ (the reverse direction can be proved analogously).

(i) McDiarmid's inequality

Since $\ell(\cdot) \le C_\ell$ and $\ell_{\text{KL}}(\cdot) \le C_{\text{KL}}$, replacing one sample on the P side changes $\widehat{R}_{\text{PN-TR}}(g)$ by at most $\frac{\pi_{\text{P}}}{n_{\text{P}}}(C_\ell + \beta C_{\text{KL}})$, and replacing one sample on the N side changes it by at most $\frac{\pi_{\text{N}}}{n_{\text{N}}}(C_\ell + \beta C_{\text{KL}})$. Therefore, by McDiarmid's inequality (A.5), for any $t > 0$,

$$\begin{aligned}\Pr\Big(\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\} - \mathbb{E}\big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\}\big] \ge t\Big)& \\ \le \exp\left(-\frac{2t^2}{(C_\ell + \beta C_{\text{KL}})^2(\pi_{\text{P}}^2/n_{\text{P}} + \pi_{\text{N}}^2/n_{\text{N}})}\right)&.\end{aligned} \tag{A6}$$

Setting $\delta/2 = \exp\left(-2t^2/\left((C_\ell + \beta C_{\text{KL}})^2(\pi_{\text{P}}^2/n_{\text{P}} + \pi_{\text{N}}^2/n_{\text{N}})\right)\right)$ and solving for t, and then using the subadditivity of the square root, we obtain, with probability at least $1-\delta/2$,

$$\begin{aligned}\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\} \le \mathbb{E}\big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\}\big]& \\ + \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\,(C_\ell + \beta C_{\text{KL}})\left(\frac{\pi_{\text{P}}}{\sqrt{n_{\text{P}}}} + \frac{\pi_{\text{N}}}{\sqrt{n_{\text{N}}}}\right)&\end{aligned} \tag{A7}$$

as desired.

(ii) Ghost sampling and symmetrization

We use the symmetrization technique in statistical learning theory [**?**]. First, in

$$\mathbb{E}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\Big],$$

the expectation $\mathbb{E}$ is taken over repeated sampling of $(\mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}})$ used to evaluate $\widehat{R}_{\text{PN-TR}}(g)$. To make this explicit, we write

$$\widehat{R}_{\text{PN-TR}}(g) = \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}).$$

Let $(\mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})$ be a ghost sample (an independent copy with the same distribution) independent of $(\mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}})$. Since the true risk can be written as the expectation over the ghost sample,

$$R_{\text{PN-TR}}(g) = \mathbb{E}_{(\mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})}\big[\widehat{R}_{\text{PN-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})\big],$$

we obtain

$$\begin{aligned}
&\mathbb{E}_{(\mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}) - R_{\text{PN-TR}}(g)\Big]\\
&= \mathbb{E}_{(\mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}) - \mathbb{E}_{(\mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})}\widehat{R}_{\text{PN-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})\Big]\\
&\le \mathbb{E}_{(\mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}),(\mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}) - \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})\Big],
\end{aligned}\tag{A8}$$

where we used Jensen's inequality, since sup is a convex function.

Next, we decompose the difference into the P and N parts. For notational simplicity, define

$$\phi_y(g, \boldsymbol{x}) := \ell\big(g(\boldsymbol{x}), y\big) + \beta \max_{\|\boldsymbol{\eta}\|_\infty \le \varepsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}), g(\boldsymbol{x}+\boldsymbol{\eta})\big).$$

Then,

$$\begin{aligned}
&\widehat{R}_{\text{PN-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}) - \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})\\
&= \frac{\pi_{\text{P}}}{n_{\text{P}}}\sum_{i=1}^{n_{\text{P}}}\Big(\phi_{+1}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{+1}(g, \boldsymbol{x}_i^{\text{P}\prime})\Big) + \frac{\pi_{\text{N}}}{n_{\text{N}}}\sum_{i=1}^{n_{\text{N}}}\Big(\phi_{-1}(g, \boldsymbol{x}_i^{\text{N}}) - \phi_{-1}(g, \boldsymbol{x}_i^{\text{N}\prime})\Big).
\end{aligned}\tag{A9}$$

By the subadditivity of sup,

$$\begin{aligned}
&\mathbb{E}_{(\mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}),(\mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{N}}) - \widehat{R}_{\text{PN-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{N}})\Big]\\
&\le \frac{\pi_{\text{P}}}{n_{\text{P}}}\mathbb{E}_{\mathscr{X}_{\text{P}}, \mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\text{P}}}\big(\phi_{+1}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{+1}(g, \boldsymbol{x}_i^{\text{P}\prime})\big)\Big]\\
&\quad + \frac{\pi_{\text{N}}}{n_{\text{N}}}\mathbb{E}_{\mathscr{X}_{\text{N}}, \mathscr{X}'_{\text{N}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\text{N}}}\big(\phi_{-1}(g, \boldsymbol{x}_i^{\text{N}}) - \phi_{-1}(g, \boldsymbol{x}_i^{\text{N}\prime})\big)\Big].
\end{aligned}\tag{A10}$$

Now consider, for example, the P side. Since $\boldsymbol{x}_i^{\text{P}}$ and $\boldsymbol{x}_i^{\text{P}\prime}$ are independent and identically distributed (both from p_{P}), the two differences $\phi_{+1}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{+1}(g, \boldsymbol{x}_i^{\text{P}\prime})$ and $\phi_{+1}(g, \boldsymbol{x}_i^{\text{P}\prime}) - \phi_{+1}(g, \boldsymbol{x}_i^{\text{P}})$ have the same distribution. Therefore, letting $L_{2,n_{\text{P}}} := \sum_{i=2}^{n_{\text{P}}}\big(\phi_{+1}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{+1}(g, \boldsymbol{x}_i^{\text{P}\prime})\big)$, we have

$$\mathbb{E}_{\mathscr{X}_{\text{P}}, \mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\text{P}}}\big(\phi_{+1}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{+1}(g, \boldsymbol{x}_i^{\text{P}\prime})\big)\Big]$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathscr{X}_{\mathrm{P}},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\big(\phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}}) - \phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}\prime})\big) + L_{2,n_{\mathrm{P}}}\Big] \\
&= \frac{1}{2}\mathbb{E}_{\mathscr{X}_{\mathrm{P}},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\big(\phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}}) - \phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}\prime})\big) + L_{2,n_{\mathrm{P}}}\Big] \\
&\quad + \frac{1}{2}\mathbb{E}_{\mathscr{X}_{\mathrm{P}},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\big(\phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}\prime}) - \phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}})\big) + L_{2,n_{\mathrm{P}}}\Big] \\
&= \mathbb{E}_{\sigma_1,\mathscr{X}_{\mathrm{P}},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sigma_1\big(\phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}}) - \phi_{+1}(g,\boldsymbol{x}_1^{\mathrm{P}\prime})\big) + L_{2,n_{\mathrm{P}}}\Big]. \qquad \text{(A11)}
\end{aligned}
$$

Repeating the same argument n_{P} times yields

$$
\begin{aligned}
&\mathbb{E}_{\mathscr{X}_{\mathrm{P}},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\big(\phi_{+1}(g,\boldsymbol{x}_i^{\mathrm{P}}) - \phi_{+1}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\big)\Big] \\
&= \mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\big(\phi_{+1}(g,\boldsymbol{x}_i^{\mathrm{P}}) - \phi_{+1}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\big)\Big] \\
&\le \mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\,\phi_{+1}(g,\boldsymbol{x}_i^{\mathrm{P}})\Big] + \mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}(-\sigma_i)\,\phi_{+1}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\Big] \\
&= 2\,\mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\,\phi_{+1}(g,\boldsymbol{x}_i^{\mathrm{P}})\Big] = 2n_{\mathrm{P}}\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\phi_{+1}\circ\mathscr{G}). \qquad \text{(A12)}
\end{aligned}
$$

Similarly, for the N side,

$$
\mathbb{E}_{\mathscr{X}_{\mathrm{N}},\mathscr{X}'_{\mathrm{N}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{N}}}\big(\phi_{-1}(g,\boldsymbol{x}_i^{\mathrm{N}}) - \phi_{-1}(g,\boldsymbol{x}_i^{\mathrm{N}\prime})\big)\Big] \le 2n_{\mathrm{N}}\,\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\phi_{-1}\circ\mathscr{G})
$$

is obtained.

Finally, we upper bound $\mathfrak{R}(\phi_y\circ\mathscr{G})$ by $\mathfrak{R}(\mathscr{G})$. Write $\phi_y = \ell(\cdot,y) + \beta\,\psi$, where

$$
\psi(g,\boldsymbol{x}) := \max_{\|\boldsymbol{\eta}\|_\infty\le\varepsilon}\ \ell_{\mathrm{KL}}\big(g(\boldsymbol{x}),g(\boldsymbol{x}+\boldsymbol{\eta})\big).
$$

Then, by the subadditivity of sup,

$$
\mathfrak{R}_{n,q}(\phi_y\circ\mathscr{G}) \le \mathfrak{R}_{n,q}(\ell(\cdot,y)\circ\mathscr{G}) + \beta\,\mathfrak{R}_{n,q}(\psi\circ\mathscr{G}).
$$

For the classification loss term, by the contraction lemma,

$$
\mathfrak{R}_{n,q}(\ell(\cdot,y)\circ\mathscr{G}) \le L_\ell\,\mathfrak{R}_{n,q}(\mathscr{G}).
$$

For the KL term, applying Lemma A.2 to $f_g(\boldsymbol{x}) = g(\boldsymbol{x})$ and $g_{g,\boldsymbol{\eta}}(\boldsymbol{x}) = g(\boldsymbol{x}+\boldsymbol{\eta})$, and then using Lemma A.4, we obtain

$$
\begin{aligned}
\mathfrak{R}_{n,q}(\psi\circ\mathscr{G}) &\le 2L_{\mathrm{KL}}\Big(\mathfrak{R}_{n,q}(\mathscr{G}) + \mathfrak{R}_{n,q}(\{\,\boldsymbol{x}\mapsto g(\boldsymbol{x}+\boldsymbol{\eta}) :\ \|\boldsymbol{\eta}\|_\infty\le\varepsilon,\ g\in\mathscr{G}\,\})\Big) \\
&\le 2L_{\mathrm{KL}}\Big(2\mathfrak{R}_{n,q}(\mathscr{G}) + \frac{\varepsilon W d^{1/q}}{\sqrt{n}}\Big). \qquad \text{(A13)}
\end{aligned}
$$

Substituting these bounds into (A8)–(A10), we get

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{g\in\mathscr{G}}\ \widehat{R}_{\text{PN-TR}}(g) - R_{\text{PN-TR}}(g)\Big] \\
&\le \frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}}\cdot 2n_{\mathrm{P}}\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\phi_{+1}\circ\mathscr{G})\ +\ \frac{\pi_{\mathrm{N}}}{n_{\mathrm{N}}}\cdot 2n_{\mathrm{N}}\,\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\phi_{-1}\circ\mathscr{G}) \\
&= 2\pi_{\mathrm{P}}\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\phi_{+1}\circ\mathscr{G}) + 2\pi_{\mathrm{N}}\,\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\phi_{-1}\circ\mathscr{G})
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2\pi_{\mathrm{P}}\Big\{L_\ell\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G})+\beta\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\psi\circ\mathscr{G})\Big\}+2\pi_{\mathrm{N}}\Big\{L_\ell\,\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G})+\beta\,\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\psi\circ\mathscr{G})\Big\}\\
&\leq 2\pi_{\mathrm{P}}\Big\{L_\ell\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G})+\beta\cdot 2L_{\mathrm{KL}}\Big(2\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G})+\frac{\varepsilon W d^{1/q}}{\sqrt{n_{\mathrm{P}}}}\Big)\Big\}\\
&\quad+2\pi_{\mathrm{N}}\Big\{L_\ell\,\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G})+\beta\cdot 2L_{\mathrm{KL}}\Big(2\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G})+\frac{\varepsilon W d^{1/q}}{\sqrt{n_{\mathrm{N}}}}\Big)\Big\}\\
&=2\big(L_\ell+4\beta L_{\mathrm{KL}}\big)\Big(\pi_{\mathrm{P}}\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G})+\pi_{\mathrm{N}}\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G})\Big)+4\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q}\Big(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}+\frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\Big),
\end{aligned}
\tag{A14}
$$

where $\phi_y(g,\boldsymbol{x}) := \ell(g(\boldsymbol{x}),y)+\beta\max_{\|\boldsymbol{\eta}\|_\infty\leq\varepsilon}\ell_{\mathrm{KL}}(g(\boldsymbol{x}),g(\boldsymbol{x}+\boldsymbol{\eta}))$ and $\psi(g,\boldsymbol{x}) := \max_{\|\boldsymbol{\eta}\|_\infty\leq\varepsilon}\ell_{\mathrm{KL}}(g(\boldsymbol{x}),g(\boldsymbol{x}+\boldsymbol{\eta}))$.

(iii) Combining the results of (i) and (ii)

Substituting the expectation bound in (ii), i.e., (A14), into McDiarmid's inequality result in (i), i.e., (A7), we obtain, with probability at least $1-\delta/2$,

$$
\begin{aligned}
\sup_{g\in\mathscr{G}}\Big\{\widehat{R}_{\text{PN-TR}}(g)-R_{\text{PN-TR}}(g)\Big\}\ \leq\ &2\big(L_\ell+4\beta L_{\mathrm{KL}}\big)\Big(\pi_{\mathrm{P}}\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G})+\pi_{\mathrm{N}}\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G})\Big)\\
&+4\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}+\frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\right)\\
&+\sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\,(C_\ell+\beta C_{\mathrm{KL}})\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}+\frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\right)
\end{aligned}
\tag{A15}
$$

as claimed.

(iv) Reverse direction and union bound

By the same argument, $\sup_{g\in\mathscr{G}}\{R_{\text{PN-TR}}(g)-\widehat{R}_{\text{PN-TR}}(g)\}$ is also bounded by the same right-hand side with probability at least $1-\delta/2$. Therefore, by the union bound, (23) holds with probability at least $1-\delta$. □

A.3 Proof of Theorem 5.2

Proof Since $\widehat{g}_{\text{PN-TR}}$ is an empirical risk minimizer, we have $\widehat{R}_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}})\leq\widehat{R}_{\text{PN-TR}}(g^*)$. Moreover, by Lemma 5.3, the following holds with probability at least $1-\delta$:

$$
\begin{aligned}
&R_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}})-R_{\text{PN-TR}}(g^*)\\
&=\Big(R_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}})-\widehat{R}_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}})\Big)+\Big(\widehat{R}_{\text{PN-TR}}(\widehat{g}_{\text{PN-TR}})-\widehat{R}_{\text{PN-TR}}(g^*)\Big)\\
&+\Big(\widehat{R}_{\text{PN-TR}}(g^*)-R_{\text{PN-TR}}(g^*)\Big)\\
&\leq\sup_{g\in\mathscr{G}}\big(R_{\text{PN-TR}}(g)-\widehat{R}_{\text{PN-TR}}(g)\big)+0+\sup_{g\in\mathscr{G}}\big(\widehat{R}_{\text{PN-TR}}(g)-R_{\text{PN-TR}}(g)\big)\\
&\leq 2\sup_{g\in\mathscr{G}}\Big|\widehat{R}_{\text{PN-TR}}(g)-R_{\text{PN-TR}}(g)\Big|\\
&\leq\ 4\big(L_\ell+4\beta L_{\mathrm{KL}}\big)\Big(\pi_{\mathrm{P}}\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G})+\pi_{\mathrm{N}}\mathfrak{R}_{n_{\mathrm{N}},p_{\mathrm{N}}}(\mathscr{G})\Big)\\
&+8\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}+\frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\right)
\end{aligned}
$$

$$+ \sqrt{2\ln\frac{2}{\delta}}\,(C_\ell + \beta C_{\mathrm{KL}})\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\right).$$

Thus, the claim of Theorem 5.2 follows. □

A.4 Proof of the Lemma "Uniform Deviation Bound for uPU+TRADES" (Lemma 5.5)

Proof As in the previous section, (i) we first show concentration of the uniform deviation using McDiarmid's inequality, (ii) evaluate the expectation via ghost sampling and Rademacher complexity, (iii) combine the results of (i) and (ii) to obtain a one-sided upper bound, and (iv) combine the reverse direction via a union bound to obtain the uniform deviation bound (the lemma). Finally, in (v), we derive the estimation error bound (the theorem) by the standard decomposition for an empirical risk minimizer.

(i) McDiarmid's inequality

Since $\ell(\cdot) \le C_\ell$ and $\ell_{\mathrm{KL}}(\cdot) \le C_{\mathrm{KL}}$, replacing one sample on the P side changes $\widehat{R}_{\mathrm{uPU\text{-}TR}}(g)$ by at most $\frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}}(2C_\ell + \beta C_{\mathrm{KL}})$. Similarly, replacing one sample on the U side changes $\widehat{R}_{\mathrm{uPU\text{-}TR}}(g)$ by at most $\frac{1}{n_{\mathrm{U}}}(C_\ell + \beta C_{\mathrm{KL}})$. Therefore, by McDiarmid's inequality, for any $t > 0$,

$$\Pr\left(\sup_{g\in\mathscr{G}}\left\{\widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)\right\} - \mathbb{E}\Big[\sup_{g\in\mathscr{G}}\left\{\widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)\right\}\Big] \ge t\right)$$
$$\le \exp\left(-\frac{2t^2}{\frac{\pi_{\mathrm{P}}^2(2C_\ell+\beta C_{\mathrm{KL}})^2}{n_{\mathrm{P}}} + \frac{(C_\ell+\beta C_{\mathrm{KL}})^2}{n_{\mathrm{U}}}}\right). \tag{A16}$$

Setting the right-hand side equal to $\delta/2$, we obtain, with probability at least $1 - \delta/2$,

$$\sup_{g\in\mathscr{G}}\{\widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)\} \le \mathbb{E}\Big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)\}\Big]$$
$$+ \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\sqrt{\frac{\pi_{\mathrm{P}}^2(2C_\ell + \beta C_{\mathrm{KL}})^2}{n_{\mathrm{P}}} + \frac{(C_\ell + \beta C_{\mathrm{KL}})^2}{n_{\mathrm{U}}}}. \tag{A17}$$

Furthermore, by using the subadditivity of the square root, we obtain

$$\sup_{g\in\mathscr{G}}\{\widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)\} \le \mathbb{E}\Big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)\}\Big]$$
$$+ \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\left(\frac{\pi_{\mathrm{P}}(2C_\ell + \beta C_{\mathrm{KL}})}{\sqrt{n_{\mathrm{P}}}} + \frac{C_\ell + \beta C_{\mathrm{KL}}}{\sqrt{n_{\mathrm{U}}}}\right) \tag{A18}$$

as desired.

(ii) Ghost sampling

Here, we upper bound $\mathbb{E}\big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)\}\big]$ by the Rademacher complexity. The expectation in $\mathbb{E}[\sup_{g\in\mathscr{G}} \widehat{R}_{\mathrm{uPU\text{-}TR}}(g) - R_{\mathrm{uPU\text{-}TR}}(g)]$ is taken over repeated sampling of $(\mathscr{X}_{\mathrm{P}}, \mathscr{X}_{\mathrm{U}})$ used to evaluate $\widehat{R}_{\mathrm{uPU\text{-}TR}}(g)$. That is, $\mathbb{E} = \mathbb{E}_{(\mathscr{X}_{\mathrm{P}},\mathscr{X}_{\mathrm{U}})}$, and $\widehat{R}_{\mathrm{uPU\text{-}TR}}(g)$ can also be written as $\widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{U}})$. Introducing an independent ghost sample $(\mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{U}})$, we obtain

$$\begin{aligned}
&\mathbb{E}_{(\mathscr{X}_{\text{P}},\mathscr{X}_{\text{U}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{uPU-TR}}(g) - R_{\text{uPU-TR}}(g)\Big] \\
&= \mathbb{E}_{(\mathscr{X}_{\text{P}},\mathscr{X}_{\text{U}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{uPU-TR}}(g) - \mathbb{E}_{(\mathscr{X}'_{\text{P}},\mathscr{X}'_{\text{U}})}\widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{U}})\Big] \\
&\le \mathbb{E}_{(\mathscr{X}_{\text{P}},\mathscr{X}_{\text{U}}),(\mathscr{X}'_{\text{P}},\mathscr{X}'_{\text{U}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{U}}) - \widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{U}})\Big] \qquad \text{(A19)}
\end{aligned}$$

The last inequality follows from Jensen's inequality, since sup is convex.

Next, we decompose the difference:

$$\phi_{\text{P}}(g, \boldsymbol{x}) := \ell\big(g(\boldsymbol{x}), +1\big) - \ell\big(g(\boldsymbol{x}), -1\big) + \beta\psi(g, \boldsymbol{x}), \qquad \phi_{\text{U}}(g, \boldsymbol{x}) := \ell\big(g(\boldsymbol{x}), -1\big) + \beta\psi(g, \boldsymbol{x}),$$

where

$$\psi(g, \boldsymbol{x}) := \max_{\|\boldsymbol{\eta}\|_\infty \le \varepsilon} \ell_{\text{KL}}\big(g(\boldsymbol{x}+\boldsymbol{\eta}), g(\boldsymbol{x})\big).$$

Then,

$$\begin{aligned}
&\widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{U}}) - \widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{U}}) \\
&= \frac{\pi_{\text{P}}}{n_{\text{P}}}\sum_{i=1}^{n_{\text{P}}}\big(\phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}\prime})\big) + \frac{1}{n_{\text{U}}}\sum_{i=1}^{n_{\text{U}}}\big(\phi_{\text{U}}(g, \boldsymbol{x}_i^{\text{U}}) - \phi_{\text{U}}(g, \boldsymbol{x}_i^{\text{U}\prime})\big). \qquad \text{(A20)}
\end{aligned}$$

Therefore, by the subadditivity of sup,

$$\begin{aligned}
&\mathbb{E}_{(\mathscr{X}_{\text{P}},\mathscr{X}_{\text{U}}),(\mathscr{X}'_{\text{P}},\mathscr{X}'_{\text{U}})}\Big[\sup_{g\in\mathscr{G}} \widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}_{\text{P}}, \mathscr{X}_{\text{U}}) - \widehat{R}_{\text{uPU-TR}}(g; \mathscr{X}'_{\text{P}}, \mathscr{X}'_{\text{U}})\Big] \\
&\le \frac{\pi_{\text{P}}}{n_{\text{P}}}\mathbb{E}_{\mathscr{X}_{\text{P}},\mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\text{P}}}\big(\phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}\prime})\big)\Big] \\
&\quad + \frac{1}{n_{\text{U}}}\mathbb{E}_{\mathscr{X}_{\text{U}},\mathscr{X}'_{\text{U}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\text{U}}}\big(\phi_{\text{U}}(g, \boldsymbol{x}_i^{\text{U}}) - \phi_{\text{U}}(g, \boldsymbol{x}_i^{\text{U}\prime})\big)\Big]. \qquad \text{(A21)}
\end{aligned}$$

Next, we symmetrize the difference terms appearing on the right-hand side of (A21) and bound them by the Rademacher complexity. For example, consider the P side. Since $\boldsymbol{x}_i^{\text{P}}$ and $\boldsymbol{x}_i^{\text{P}\prime}$ are independent and identically distributed (both from p_{P}), the two differences $\phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}\prime})$ and $\phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}\prime}) - \phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}})$ have the same distribution. Therefore, letting $L_{2,n_{\text{P}}} := \sum_{i=2}^{n_{\text{P}}}\big(\phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}\prime})\big)$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathscr{X}_{\text{P}},\mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\text{P}}}\big(\phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}\prime})\big)\Big] \\
&= \frac{1}{2}\mathbb{E}_{\mathscr{X}_{\text{P}},\mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\big(\phi_{\text{P}}(g, \boldsymbol{x}_1^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_1^{\text{P}\prime})\big) + L_{2,n_{\text{P}}}\Big] \\
&\quad + \frac{1}{2}\mathbb{E}_{\mathscr{X}_{\text{P}},\mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\big(\phi_{\text{P}}(g, \boldsymbol{x}_1^{\text{P}\prime}) - \phi_{\text{P}}(g, \boldsymbol{x}_1^{\text{P}})\big) + L_{2,n_{\text{P}}}\Big] \\
&= \mathbb{E}_{\sigma_1,\mathscr{X}_{\text{P}},\mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\sigma_1\big(\phi_{\text{P}}(g, \boldsymbol{x}_1^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_1^{\text{P}\prime})\big) + L_{2,n_{\text{P}}}\Big]. \qquad \text{(A22)}
\end{aligned}$$

Repeating the same argument n_{P} times, and using independent Rademacher variables $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{n_{\text{P}}})$, we obtain

$$\mathbb{E}_{\mathscr{X}_{\text{P}},\mathscr{X}'_{\text{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\text{P}}}\big(\phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}}) - \phi_{\text{P}}(g, \boldsymbol{x}_i^{\text{P}\prime})\big)\Big]$$

$$\begin{aligned}
&= \mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\big(\phi_{\mathrm{P}}(g,\boldsymbol{x}_i^{\mathrm{P}}) - \phi_{\mathrm{P}}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\big)\Big] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\,\phi_{\mathrm{P}}(g,\boldsymbol{x}_i^{\mathrm{P}})\Big] + \mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}'_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}(-\sigma_i)\,\phi_{\mathrm{P}}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\Big] \\
&= 2\,\mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\,\phi_{\mathrm{P}}(g,\boldsymbol{x}_i^{\mathrm{P}})\Big] = 2n_{\mathrm{P}}\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\phi_{\mathrm{P}}\circ\mathscr{G}). \qquad \text{(A23)}
\end{aligned}$$

Similarly, for the U side,

$$\mathbb{E}_{\mathscr{X}_{\mathrm{U}},\mathscr{X}'_{\mathrm{U}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{U}}}\big(\phi_{\mathrm{U}}(g,\boldsymbol{x}_i^{\mathrm{U}}) - \phi_{\mathrm{U}}(g,\boldsymbol{x}_i^{\mathrm{U}\prime})\big)\Big] \leq 2n_{\mathrm{U}}\,\mathfrak{R}_{n_{\mathrm{U}},p_{\mathrm{U}}}(\phi_{\mathrm{U}}\circ\mathscr{G}) \qquad \text{(A24)}$$

is obtained.

Finally, we upper bound $\mathfrak{R}(\phi_{\mathrm{P}}\circ\mathscr{G})$ and $\mathfrak{R}(\phi_{\mathrm{U}}\circ\mathscr{G})$ by $\mathfrak{R}(\mathscr{G})$. By the subadditivity of sup,

$$\begin{aligned}
\mathfrak{R}_{n,q}(\phi_{\mathrm{P}}\circ\mathscr{G}) &\leq \mathfrak{R}_{n,q}(\ell(\cdot,+1)\circ\mathscr{G}) + \mathfrak{R}_{n,q}(\ell(\cdot,-1)\circ\mathscr{G}), \\
\mathfrak{R}_{n,q}(\phi_{\mathrm{U}}\circ\mathscr{G}) &\leq \mathfrak{R}_{n,q}(\ell(\cdot,-1)\circ\mathscr{G}) + \beta\,\mathfrak{R}_{n,q}(\psi\circ\mathscr{G}).
\end{aligned}$$

For the classification loss term, the contraction lemma gives

$$\mathfrak{R}_{n,q}(\ell(\cdot,y)\circ\mathscr{G}) \leq L_\ell\,\mathfrak{R}_{n,q}(\mathscr{G}) \qquad (y\in\mathcal{Y}).$$

For the KL term, applying Lemma A.2 to $f_g(\boldsymbol{x}) = g(\boldsymbol{x})$ and $g_{g,\boldsymbol{\eta}}(\boldsymbol{x}) = g(\boldsymbol{x}+\boldsymbol{\eta})$, and then using Lemma A.4, we obtain

$$\begin{aligned}
\mathfrak{R}_{n,q}(\psi\circ\mathscr{G}) &\leq 2L_{\mathrm{KL}}\Big(\mathfrak{R}_{n,q}(\mathscr{G}) + \mathfrak{R}_{n,q}(\{\,\boldsymbol{x}\mapsto g(\boldsymbol{x}+\boldsymbol{\eta}) :\ \|\boldsymbol{\eta}\|_\infty \leq \varepsilon,\ g\in\mathscr{G}\,\})\Big) \\
&\leq 2L_{\mathrm{KL}}\Big(2\mathfrak{R}_{n,q}(\mathscr{G}) + \frac{\varepsilon W d^{1/q}}{\sqrt{n}}\Big). \qquad \text{(A25)}
\end{aligned}$$

Substituting these bounds into (A19)–(A21), we obtain

$$\begin{aligned}
\mathbb{E}\Bigg[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{uPU-TR}}(g) - R_{\text{uPU-TR}}(g)\}\Bigg] &\leq 2\pi_{\mathrm{P}}\,\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}\big(\phi_{\mathrm{P}}\circ\mathscr{G}\big) + 2\,\mathfrak{R}_{n_{\mathrm{U}},p_{\mathrm{U}}}\big(\phi_{\mathrm{U}}\circ\mathscr{G}\big) \\
&\leq 4\pi_{\mathrm{P}}\big(L_\ell + 2\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G}) \\
&\quad + 2\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{U}},p_{\mathrm{U}}}(\mathscr{G}) \\
&\quad + 4\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right). \qquad \text{(A26)}
\end{aligned}$$

This completes the bound.

(iii) Combining the results of (i) and (ii)

Substituting the expectation bound in (ii), i.e., (A26), into the McDiarmid inequality result in (i), i.e., (A18), we obtain, with probability at least $1-\delta/2$,

$$\begin{aligned}
\sup_{g\in\mathscr{G}}\Big\{\widehat{R}_{\text{uPU-TR}}(g) - R_{\text{uPU-TR}}(g)\Big\} &\leq 4\pi_{\mathrm{P}}\big(L_\ell + 2\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G}) + 2\big(L_\ell + 4\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{U}},p_{\mathrm{U}}}(\mathscr{G}) \\
&\quad + 4\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right) \\
&\quad + \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\left(\frac{\pi_{\mathrm{P}}(2C_\ell + \beta C_{\mathrm{KL}})}{\sqrt{n_{\mathrm{P}}}} + \frac{C_\ell + \beta C_{\mathrm{KL}}}{\sqrt{n_{\mathrm{U}}}}\right). \qquad \text{(A27)}
\end{aligned}$$

This completes the one-sided bound. □

A.5 Proof of Theorem 5.4

Proof Since $\widehat{g}_{\text{uPU-TR}}$ is an empirical risk minimizer, we have

$$\widehat{R}_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}}) \le \widehat{R}_{\text{uPU-TR}}(g^*)$$

Moreover, on the event where Lemma 5.5 holds (which occurs with probability at least $1-\delta$),

$$\begin{aligned}
&R_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}}) - R_{\text{uPU-TR}}(g^*)\\
&= \Big(R_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}}) - \widehat{R}_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}})\Big)\\
&+ \Big(\widehat{R}_{\text{uPU-TR}}(\widehat{g}_{\text{uPU-TR}}) - \widehat{R}_{\text{uPU-TR}}(g^*)\Big)\\
&+ \Big(\widehat{R}_{\text{uPU-TR}}(g^*) - R_{\text{uPU-TR}}(g^*)\Big)\\
&\le \sup_{g\in\mathscr{G}} \big(R_{\text{uPU-TR}}(g) - \widehat{R}_{\text{uPU-TR}}(g)\big) + 0 + \sup_{g\in\mathscr{G}} \big(\widehat{R}_{\text{uPU-TR}}(g) - R_{\text{uPU-TR}}(g)\big)\\
&\le 2\sup_{g\in\mathscr{G}} \Big|\widehat{R}_{\text{uPU-TR}}(g) - R_{\text{uPU-TR}}(g)\Big|.
\end{aligned}$$

Therefore, multiplying (27) by 2 yields (26). □

A.6 Proof of the Lemma "Uniform Deviation Bound for nnPU+TRADES" (Lemma 5.7)

(i) McDiarmid's inequality

Since $\ell(\cdot) \le C_\ell$ and $\ell_{\text{KL}}(\cdot) \le C_{\text{KL}}$, replacing one sample on the P side changes the sum in the first line by at most $\frac{\pi_{\text{P}}}{n_{\text{P}}}(C_\ell + \beta C_{\text{KL}})$. In addition, since the truncation term $\max\{0, \cdot\}$ in nnPU is 1-Lipschitz, it can be upper bounded by the change in its argument. Inside the truncation term, $\ell(g(\boldsymbol{x}), -1)$ on the P side is replaced at only one point, so the change is at most $\frac{\pi_{\text{P}}}{n_{\text{P}}}C_\ell$. Therefore, replacing one P-side sample changes $\widehat{R}_{\text{nnPU-TR}}(g)$ by at most $\frac{\pi_{\text{P}}}{n_{\text{P}}}(2C_\ell + \beta C_{\text{KL}})$ in total. Similarly, replacing one U-side sample affects $\ell(g(\boldsymbol{x}), -1)$ inside the truncation term and ψ in the third line at only one point, and hence changes $\widehat{R}_{\text{nnPU-TR}}(g)$ by at most $\frac{1}{n_{\text{U}}}(C_\ell + \beta C_{\text{KL}})$. Therefore, by McDiarmid's inequality, for any $t > 0$,

$$\begin{aligned}
&\Pr\left(\sup_{g\in\mathscr{G}} \left\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\right\} - \mathbb{E}\Big[\sup_{g\in\mathscr{G}} \left\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\right\}\Big] \ge t\right)\\
&\quad\le \exp\left(-\frac{2t^2}{\frac{\pi_{\text{P}}^2(2C_\ell+\beta C_{\text{KL}})^2}{n_{\text{P}}} + \frac{(C_\ell+\beta C_{\text{KL}})^2}{n_{\text{U}}}}\right). \qquad \text{(A28)}
\end{aligned}$$

Setting the right-hand side equal to $\delta/2$, we obtain, with probability at least $1-\delta/2$,

$$\begin{aligned}
\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\} &\le \mathbb{E}\Big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\}\Big]\\
&\quad + \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\sqrt{\frac{\pi_{\text{P}}^2(2C_\ell + \beta C_{\text{KL}})^2}{n_{\text{P}}} + \frac{(C_\ell + \beta C_{\text{KL}})^2}{n_{\text{U}}}}. \qquad \text{(A29)}
\end{aligned}$$

Furthermore, by using the subadditivity of the square root, we obtain

$$\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\} \le \mathbb{E}\Big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\}\Big]$$

$$+\sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\left(\frac{\pi_{\mathrm{P}}(2C_\ell+\beta C_{\mathrm{KL}})}{\sqrt{n_{\mathrm{P}}}}+\frac{C_\ell+\beta C_{\mathrm{KL}}}{\sqrt{n_{\mathrm{U}}}}\right) \tag{A30}$$

as desired.

(ii) Ghost sampling

Here, we upper bound $\mathbb{E}\big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\mathrm{nnPU\text{-}TR}}(g)-R_{\mathrm{nnPU\text{-}TR}}(g)\}\big]$ by the Rademacher complexity. Unlike the uPU case, $\widehat{R}_{\mathrm{nnPU\text{-}TR}}(g)$ contains the nnPU truncation term $\max\{0,\cdot\}$, and therefore, in general, $R_{\mathrm{nnPU\text{-}TR}}(g)=\mathbb{E}[\widehat{R}_{\mathrm{nnPU\text{-}TR}}(g)]$ does not hold. Thus, using the 1-Lipschitz property of $\max\{0,\cdot\}$,

$$|\max\{0,a\}-\max\{0,b\}|\le|a-b|,$$

we first reduce the problem to a sum of differences between empirical and population averages, and then evaluate each term by ghost sampling and symmetrization, as in uPU+TRADES. First, define

$$\phi_+(g,\boldsymbol{x}):=\ell\big(g(\boldsymbol{x}),+1\big)+\beta\,\psi(g,\boldsymbol{x}),\qquad \phi_-(g,\boldsymbol{x}):=\ell\big(g(\boldsymbol{x}),-1\big),\qquad \phi_\psi(g,\boldsymbol{x}):=\psi(g,\boldsymbol{x}),$$

where

$$\psi(g,\boldsymbol{x}):=\max_{\|\boldsymbol{\eta}\|_\infty\le\varepsilon}\ell_{\mathrm{KL}}\big(g(\boldsymbol{x}+\boldsymbol{\eta}),g(\boldsymbol{x})\big).$$

Then, writing the inside of the truncation term as

$$\begin{aligned}\widehat{s}(g)&:=-\frac{\pi_{\mathrm{P}}}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_-\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)+\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\phi_-\big(g,\boldsymbol{x}_i^{\mathrm{U}}\big),\\ s(g)&:=-\pi_{\mathrm{P}}\,\mathbb{E}_{\mathrm{P}}[\phi_-(g,\boldsymbol{x})]+\mathbb{E}_{\mathrm{U}}[\phi_-(g,\boldsymbol{x})]\,,\end{aligned}$$

we have

$$\begin{aligned}\widehat{R}_{\mathrm{nnPU\text{-}TR}}(g)-R_{\mathrm{nnPU\text{-}TR}}(g)&=\pi_{\mathrm{P}}\left\{\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_+\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathrm{P}}[\phi_+(g,\boldsymbol{x})]\right\}\\&\quad+\beta\left\{\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\phi_\psi\big(g,\boldsymbol{x}_i^{\mathrm{U}}\big)-\mathbb{E}_{\mathrm{U}}\left[\phi_\psi(g,\boldsymbol{x})\right]\right\}\\&\quad+\max\{0,\widehat{s}(g)\}-\max\{0,s(g)\}\\&\le\pi_{\mathrm{P}}\left\{\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_+\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathrm{P}}[\phi_+(g,\boldsymbol{x})]\right\}\\&\quad+\beta\left\{\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\phi_\psi\big(g,\boldsymbol{x}_i^{\mathrm{U}}\big)-\mathbb{E}_{\mathrm{U}}\left[\phi_\psi(g,\boldsymbol{x})\right]\right\}\\&\quad+|\widehat{s}(g)-s(g)|\\&\le\pi_{\mathrm{P}}\left\{\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_+\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathrm{P}}[\phi_+(g,\boldsymbol{x})]\right\}\\&\quad+\beta\left\{\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\phi_\psi\big(g,\boldsymbol{x}_i^{\mathrm{U}}\big)-\mathbb{E}_{\mathrm{U}}\left[\phi_\psi(g,\boldsymbol{x})\right]\right\}\\&\quad+\pi_{\mathrm{P}}\left|\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_-\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathrm{P}}[\phi_-(g,\boldsymbol{x})]\right|\end{aligned}$$

$$+\left|\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\phi_{-}\big(g,\boldsymbol{x}_i^{\mathrm{U}}\big)-\mathbb{E}_{\mathrm{U}}[\phi_{-}(g,\boldsymbol{x})]\right|. \tag{A31}$$

Therefore, by the subadditivity of sup and the triangle inequality,

$$\begin{aligned}\mathbb{E}\Big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{nnPU-TR}}(g)-R_{\text{nnPU-TR}}(g)\}\Big] \leq\ & \pi_{\mathrm{P}}\,\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left\{\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_{+}\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathrm{P}}[\phi_{+}(g,\boldsymbol{x})]\right\}\right]\\ &+\beta\,\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left\{\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\phi_{\psi}\big(g,\boldsymbol{x}_i^{\mathrm{U}}\big)-\mathbb{E}_{\mathrm{U}}\left[\phi_{\psi}(g,\boldsymbol{x})\right]\right\}\right]\\ &+\pi_{\mathrm{P}}\,\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left|\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_{-}\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathrm{P}}[\phi_{-}(g,\boldsymbol{x})]\right|\right]\\ &+\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left|\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\phi_{-}\big(g,\boldsymbol{x}_i^{\mathrm{U}}\big)-\mathbb{E}_{\mathrm{U}}[\phi_{-}(g,\boldsymbol{x})]\right|\right].\end{aligned} \tag{A32}$$

Next, we evaluate each term by ghost sampling and symmetrization (as in uPU+TRADES). For illustration, consider the first term on the P side. Let $\mathscr{X}_{\mathrm{P}}'=\{\boldsymbol{x}_i^{\mathrm{P}\prime}\}_{i=1}^{n_{\mathrm{P}}}$ be an independent ghost sample on the P side. Then,

$$\begin{aligned}&\mathbb{E}_{\mathscr{X}_{\mathrm{P}}}\left[\sup_{g\in\mathscr{G}}\left\{\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_{+}\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathrm{P}}[\phi_{+}(g,\boldsymbol{x})]\right\}\right]\\ &=\mathbb{E}_{\mathscr{X}_{\mathrm{P}}}\left[\sup_{g\in\mathscr{G}}\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_{+}\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\mathbb{E}_{\mathscr{X}_{\mathrm{P}}'}\Big\{\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\phi_{+}\big(g,\boldsymbol{x}_i^{\mathrm{P}\prime}\big)\Big\}\right]\\ &\leq\mathbb{E}_{\mathscr{X}_{\mathrm{P}},\mathscr{X}_{\mathrm{P}}'}\left[\sup_{g\in\mathscr{G}}\frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\Big(\phi_{+}\big(g,\boldsymbol{x}_i^{\mathrm{P}}\big)-\phi_{+}\big(g,\boldsymbol{x}_i^{\mathrm{P}\prime}\big)\Big)\right]\end{aligned} \tag{A33}$$

where the last inequality follows from Jensen's inequality, since sup is convex. Here, since $\boldsymbol{x}_i^{\mathrm{P}}$ and $\boldsymbol{x}_i^{\mathrm{P}\prime}$ are independent and identically distributed (both from p_{P}), the two differences $\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}})-\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})$ and $\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})-\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}})$ have the same distribution. Therefore, letting $L_{2,n_{\mathrm{P}}}:=\sum_{i=2}^{n_{\mathrm{P}}}\big(\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}})-\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\big)$, we have

$$\begin{aligned}&\mathbb{E}_{\mathscr{X}_{\mathrm{P}},\mathscr{X}_{\mathrm{P}}'}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\big(\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}})-\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\big)\Big]\\ &=\mathbb{E}_{\sigma_1,\mathscr{X}_{\mathrm{P}},\mathscr{X}_{\mathrm{P}}'}\Big[\sup_{g\in\mathscr{G}}\sigma_1\big(\phi_{+}(g,\boldsymbol{x}_1^{\mathrm{P}})-\phi_{+}(g,\boldsymbol{x}_1^{\mathrm{P}\prime})\big)+L_{2,n_{\mathrm{P}}}\Big].\end{aligned} \tag{A34}$$

Repeating the same argument n_{P} times, and using independent Rademacher variables $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_{n_{\mathrm{P}}})$, we obtain

$$\begin{aligned}&\mathbb{E}_{\mathscr{X}_{\mathrm{P}},\mathscr{X}_{\mathrm{P}}'}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\big(\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}})-\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\big)\Big]\\ &=\mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}},\mathscr{X}_{\mathrm{P}}'}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\big(\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}})-\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\big)\Big]\\ &\leq\mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}}}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}\sigma_i\,\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}})\Big]+\mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_{\mathrm{P}}'}\Big[\sup_{g\in\mathscr{G}}\sum_{i=1}^{n_{\mathrm{P}}}(-\sigma_i)\,\phi_{+}(g,\boldsymbol{x}_i^{\mathrm{P}\prime})\Big]\end{aligned}$$

$$= 2\,\mathbb{E}_{\boldsymbol{\sigma},\mathscr{X}_\mathrm{P}}\Big[\sup_{g\in\mathscr{G}} \sum_{i=1}^{n_\mathrm{P}} \sigma_i\,\phi_+(g,\boldsymbol{x}_i^\mathrm{P})\Big] = 2n_\mathrm{P}\,\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\phi_+\circ\mathscr{G}). \tag{A35}$$

Hence, combining this with (A33), we obtain

$$\mathbb{E}_{\mathscr{X}_\mathrm{P}}\left[\sup_{g\in\mathscr{G}}\left\{\frac{1}{n_\mathrm{P}}\sum_{i=1}^{n_\mathrm{P}}\phi_+\big(g,\boldsymbol{x}_i^\mathrm{P}\big) - \mathbb{E}_\mathrm{P}[\phi_+(g,\boldsymbol{x})]\right\}\right] \le 2\,\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\phi_+\circ\mathscr{G}).$$

Similarly, the second term on the U side and the third and fourth terms with absolute values can also be symmetrized, yielding

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left\{\frac{1}{n_\mathrm{U}}\sum_{i=1}^{n_\mathrm{U}}\phi_\psi\big(g,\boldsymbol{x}_i^\mathrm{U}\big) - \mathbb{E}_\mathrm{U}\left[\phi_\psi(g,\boldsymbol{x})\right]\right\}\right] &\le 2\,\mathfrak{R}_{n_\mathrm{U},p_\mathrm{U}}(\phi_\psi\circ\mathscr{G}),\\
\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left|\frac{1}{n_\mathrm{P}}\sum_{i=1}^{n_\mathrm{P}}\phi_-\big(g,\boldsymbol{x}_i^\mathrm{P}\big) - \mathbb{E}_\mathrm{P}[\phi_-(g,\boldsymbol{x})]\right|\right] &\le 2\,\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\phi_-\circ\mathscr{G}),\\
\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left|\frac{1}{n_\mathrm{U}}\sum_{i=1}^{n_\mathrm{U}}\phi_-\big(g,\boldsymbol{x}_i^\mathrm{U}\big) - \mathbb{E}_\mathrm{U}[\phi_-(g,\boldsymbol{x})]\right|\right] &\le 2\,\mathfrak{R}_{n_\mathrm{U},p_\mathrm{U}}(\phi_-\circ\mathscr{G}) \qquad \text{(A36)}
\end{aligned}$$

Substituting these bounds into (A32), we obtain

$$\begin{aligned}
\mathbb{E}\Big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\}\Big] \le\ & 2\pi_\mathrm{P}\,\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\phi_+\circ\mathscr{G})\\
&+ 2\beta\,\mathfrak{R}_{n_\mathrm{U},p_\mathrm{U}}(\phi_\psi\circ\mathscr{G})\\
&+ 2\pi_\mathrm{P}\,\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\phi_-\circ\mathscr{G})\\
&+ 2\,\mathfrak{R}_{n_\mathrm{U},p_\mathrm{U}}(\phi_-\circ\mathscr{G}). \qquad \text{(A37)}
\end{aligned}$$

Finally, we upper bound each Rademacher complexity term by $\mathfrak{R}(\mathscr{G})$. By the subadditivity of sup,

$$\mathfrak{R}_{n,p}(\phi_+\circ\mathscr{G}) \le \mathfrak{R}_{n,p}(\ell(\cdot,+1)\circ\mathscr{G}) + \beta\,\mathfrak{R}_{n,p}(\psi\circ\mathscr{G}), \qquad \mathfrak{R}_{n,p}(\phi_-\circ\mathscr{G}) = \mathfrak{R}_{n,p}(\ell(\cdot,-1)\circ\mathscr{G}).$$

For the classification loss, the contraction lemma implies

$$\mathfrak{R}_{n,p}(\ell(\cdot,y)\circ\mathscr{G}) \le L_\ell\,\mathfrak{R}_{n,p}(\mathscr{G}) \qquad (y\in\mathcal{Y}).$$

For ψ, applying Lemma A.2 and Lemma A.4, we obtain

$$\begin{aligned}
\mathfrak{R}_{n,p}(\psi\circ\mathscr{G}) &\le 2L_\mathrm{KL}\Big(\mathfrak{R}_{n,p}(\mathscr{G}) + \mathfrak{R}_{n,p}(\{\,\boldsymbol{x}\mapsto g(\boldsymbol{x}+\boldsymbol{\eta}):\ \|\boldsymbol{\eta}\|_\infty\le\varepsilon,\ g\in\mathscr{G}\,\})\Big)\\
&\le 2L_\mathrm{KL}\Big(2\mathfrak{R}_{n,p}(\mathscr{G}) + \frac{\varepsilon W d^{1/q}}{\sqrt{n}}\Big). \qquad \text{(A38)}
\end{aligned}$$

Substituting these bounds into (A37) and simplifying yields

$$\begin{aligned}
\mathbb{E}\Big[\sup_{g\in\mathscr{G}}\{\widehat{R}_{\text{nnPU-TR}}(g) - R_{\text{nnPU-TR}}(g)\}\Big] &\le 4\pi_\mathrm{P}\big(L_\ell + 2\beta L_\mathrm{KL}\big)\mathfrak{R}_{n_\mathrm{P},p_\mathrm{P}}(\mathscr{G})\\
&\quad + 2\big(L_\ell + 4\beta L_\mathrm{KL}\big)\mathfrak{R}_{n_\mathrm{U},p_\mathrm{U}}(\mathscr{G})\\
&\quad + 4\beta L_\mathrm{KL}\,\varepsilon W d^{1/q}\left(\frac{\pi_\mathrm{P}}{\sqrt{n_\mathrm{P}}} + \frac{1}{\sqrt{n_\mathrm{U}}}\right). \qquad \text{(A39)}
\end{aligned}$$

as desired.

(iii) Combining the results of (i) and (ii)

Substituting the expectation bound in (ii), i.e., (A39), into the McDiarmid inequality result in (i), i.e., (A30), we obtain, with probability at least $1-\delta/2$,

$$\begin{aligned}\sup_{g\in\mathscr{G}}\Big\{\widehat{R}_{\text{nnPU-TR}}(g)-R_{\text{nnPU-TR}}(g)\Big\} \;&\le 4\pi_{\mathrm{P}}\big(L_\ell+2\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{P}},p_{\mathrm{P}}}(\mathscr{G})+2\big(L_\ell+4\beta L_{\mathrm{KL}}\big)\mathfrak{R}_{n_{\mathrm{U}},p_{\mathrm{U}}}(\mathscr{G})\\ &\quad+4\beta L_{\mathrm{KL}}\,\varepsilon W d^{1/q}\left(\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}+\frac{1}{\sqrt{n_{\mathrm{U}}}}\right)\\ &\quad+\sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\left(\frac{\pi_{\mathrm{P}}(2C_\ell+\beta C_{\mathrm{KL}})}{\sqrt{n_{\mathrm{P}}}}+\frac{C_\ell+\beta C_{\mathrm{KL}}}{\sqrt{n_{\mathrm{U}}}}\right).\end{aligned}\tag{A40}$$

This establishes the one-sided bound.

(iv) Opposite direction and union bound Similarly, $\sup_{g\in\mathscr{G}}\{R_{\text{nnPU-TR}}(g)-\widehat{R}_{\text{nnPU-TR}}(g)\}$ is also bounded by the right-hand side of (A40) with probability at least $1-\delta/2$. Therefore, by the union bound, (31) holds with probability at least $1-\delta$. □

A.7 Proof of Theorem 5.6

Proof Proof Since $\widehat{g}_{\text{nnPU-TR}}$ is an empirical risk minimizer, we have

$$\widehat{R}_{\text{nnPU-TR}}(\widehat{g}_{\text{nnPU-TR}})\le\widehat{R}_{\text{nnPU-TR}}(g^*)$$

Moreover, on the event where Lemma 5.7 holds (which occurs with probability at least $1-\delta$),

$$\begin{aligned}&R_{\text{nnPU-TR}}(\widehat{g}_{\text{nnPU-TR}})-R_{\text{nnPU-TR}}(g^*)\\ &=\Big(R_{\text{nnPU-TR}}(\widehat{g}_{\text{nnPU-TR}})-\widehat{R}_{\text{nnPU-TR}}(\widehat{g}_{\text{nnPU-TR}})\Big)\\ &+\Big(\widehat{R}_{\text{nnPU-TR}}(\widehat{g}_{\text{nnPU-TR}})-\widehat{R}_{\text{nnPU-TR}}(g^*)\Big)\\ &+\Big(\widehat{R}_{\text{nnPU-TR}}(g^*)-R_{\text{nnPU-TR}}(g^*)\Big)\\ &\le\sup_{g\in\mathscr{G}}\big(R_{\text{nnPU-TR}}(g)-\widehat{R}_{\text{nnPU-TR}}(g)\big)+0+\sup_{g\in\mathscr{G}}\big(\widehat{R}_{\text{nnPU-TR}}(g)-R_{\text{nnPU-TR}}(g)\big)\\ &\le 2\sup_{g\in\mathscr{G}}\Big|\widehat{R}_{\text{nnPU-TR}}(g)-R_{\text{nnPU-TR}}(g)\Big|.\end{aligned}$$

Therefore, multiplying (31) by 2 yields (30). □

A.8 Proof of the Theorem "Condition on the Number of Unlabeled Samples for PU+TRADES to Outperform Supervised TRADES" (Theorem 5.8)

Proof (i) Comparing (35) and (36), and canceling $\Gamma_\delta(\pi_{\mathrm{P}}/\sqrt{n_{\mathrm{P}}})$ from both sides, we obtain

$$\Gamma_\delta\frac{\pi_{\mathrm{N}}}{\sqrt{n_{\mathrm{N}}}}\;>\;\frac{\pi_{\mathrm{P}}}{\sqrt{n_{\mathrm{P}}}}\Big(4L_\ell WC_x d^{1/q}+\kappa_\delta C_\ell\Big)+\frac{\Gamma_\delta}{\sqrt{n_{\mathrm{U}}}}.$$

Rearranging by moving the first term on the right-hand side, and solving for $\Gamma_\delta/\sqrt{n_{\mathrm{U}}}$, we obtain (38) under condition (37). Part (ii) follows immediately from the fact that the bound for nnPU+TRADES has the same form as that for uPU+TRADES. □