Semi-Supervised Ordinal Regression Based on Empirical Risk Minimization

Taira Tsuchiya*†1,2, Nontawat Charoenphakdee1,2, Issei Sato1, and Masashi Sugiyama2,1

[1]The University of Tokyo
[2]RIKEN AIP

Abstract

Ordinal regression is aimed at predicting an ordinal class label. In this paper, we consider its semi-supervised formulation, in which we have unlabeled data along with ordinal-labeled data to train an ordinal regressor. There are several metrics to evaluate the performance of ordinal regression, such as the mean absolute error, mean zero-one error, and mean squared error. However, the existing studies do not take the evaluation metric into account, have a restriction on the model choice, and have no theoretical guarantee. To overcome these problems, we propose a novel generic framework for semi-supervised ordinal regression based on the empirical risk minimization principle that is applicable to optimizing all of the metrics mentioned above. Besides, our framework has flexible choices of models, surrogate losses, and optimization algorithms without the common geometric assumption on unlabeled data such as the cluster assumption or manifold assumption. We further provide an estimation error bound to show that our risk estimator is consistent. Finally, we conduct experiments to show the usefulness of our framework.

1 Introduction

The goal of ordinal regression is to learn an ordinal regressor to predict a label from a discrete and ordered label set (Chu and Keerthi, 2005; Chu and Ghahramani, 2005; Gutierrez et al., 2016; Pedregosa et al., 2017). For example, consider the problem of predicting the diabetes stage of a patient. The progress of diabetes consists of five stages ranging from mild to severe conditions (Weir and Bonner-Weir, 2004). The stage number is discrete, and the possible stages are total-ordered. Ordinal regression has been employed in a variety of fields such as medical research (Bender and Grouven, 1997, 1998), credit rating (Kim and Ahn, 2012; Dikkers and Rothkrantz, 2005), and social sciences (Fullerton and Xu, 2012). In the real-world, the labeling process can be costly and time-consuming. Hence, it is desirable to make use of unlabeled data to improve the prediction accuracy. Although semi-supervised ordinal regression has great benefits to practical applications, it has not been extensively explored yet (Hua-fu, 2010; Liu et al., 2011; Seah et al., 2012; Srijith et al., 2013; Pérez-Ortiz et al., 2016).

The main challenge of semi-supervised learning is how to incorporate unlabeled data to improve the prediction performance (Chapelle et al., 2006). To make use of unlabeled data, many assumptions on unlabeled data have been proposed, which often be violated in real-world problems. The representative one is the manifold assumption (Belkin et al., 2006), in which input data is assumed to be distributed in a lower-dimensional manifold. The manifold assumption has also been utilized in the context of ordinal regression (Liu et al., 2011; Pérez-Ortiz et al., 2016). Another popular assumption is the cluster

*tsuchiya@sys.i.kyoto-u.ac.jp
†Currently at Kyoto University

assumption (Seeger, 2000), in which examples belonging to the same cluster in the input space should have the same label. Seah et al. (2012) proposed a transductive ordinal regression method based on the cluster assumption. However, such a cluster assumption may no be satisfied in practice. From a different perspective, a semi-supervised ordinal regression method based on Gaussian processes has been proposed (Srijith et al., 2013), which relies on the low-density separation assumption (Chapelle et al., 2006). However, the low-density separation assumption may rarely be satisfied in reality, and Gaussian processes have high computation costs and also restrictions on the choice of models. It is known that the performance of semi-supervised learning methods based on such a geometric assumption is significantly degraded if the assumption on unlabeled data is violated (Li and Zhou, 2015; Sakai et al., 2017). Moreover, all the existing semi-supervised ordinal regression methods mentioned above do not have a theoretical guarantee, do not take the target evaluation metric into account, and have limitations in the choice of models.

Our goal is to establish a novel generic framework that allows flexible choices of models, evaluation metrics, and optimization algorithms and does not require any geometric assumption on unlabeled data. From the recent theoretical advances of ordinal regression in Pedregosa et al. (2017) and semi-supervised binary classification based on positive-unlabeled classification in Sakai et al. (2017), we propose an empirical risk minimization (ERM) based framework that achieves this ambitious goal. To theoretically justify our framework, we show that our proposed unbiased risk estimator is consistent by establishing an estimation error bound. In addition, we conduct the analysis of variance reduction to illustrate that unlabeled data can help improve the performance. Finally, we demonstrate the usefulness of our framework by conducting experiments for the three evaluation metrics using benchmark datasets.

2 Preliminaries

In this section, we formulate the standard supervised ordinal regression problem. Besides, we discuss its loss function, which is specific to the ordinal regression problem and largely different from the standard loss functions for regression and classification.

2.1 Supervised Ordinal Regression

We formulate the risk used in supervised ordinal regression. Let $\mathcal{X} \subset \mathbb{R}^d$ be a d-dimensional input space and $\mathcal{Y} = \{1, \ldots, K\}$ be an ordered label space, where K is the number of classes. We assume that labeled data $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is drawn from the joint probability distribution with density p. In ordinal regression, a function $g : \mathcal{X} \to \mathcal{Y}$, which is called a *prediction function*, is used to predict an ordinal label y from input $\boldsymbol{x}$ as

$$g(\boldsymbol{x}; f, \boldsymbol{\theta}) \coloneqq 1 + \sum_{i=1}^{K-1} \mathbb{1}[f(\boldsymbol{x}) > \theta_i], \tag{1}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{K-1}]^\top \in \mathbb{R}^{K-1}$ is the *threshold parameters* that should be ordered $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{K-1}$, $f : \mathcal{X} \to \mathbb{R}$ is the *decision function*, and $\mathbb{1}[\cdot]$ is the indicator function which is 1 if the inner condition holds and otherwise is 0 (Pedregosa et al., 2017). Note that f is a function parameterized by some parameters that is unrelated to $\boldsymbol{\theta}$, and that f and $\boldsymbol{\theta}$ are together the parameters of g. The goal of ordinal regression is to obtain a prediction function that minimizes the *task risk* defined as

$$\mathcal{R}(g) \coloneqq \mathbb{E}_{X,Y}\left[\mathcal{L}(g(X), Y)\right], \tag{2}$$

where $\mathbb{E}_{X,Y}[\cdot]$ denotes the expectation over the joint distribution over $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is called a *task loss function*. The task loss function $\mathcal{L}(g(\boldsymbol{x}), y)$ measures the performance of a prediction function based on a difference between the output of the prediction function $g(\boldsymbol{x}) \in \mathcal{Y}$ and the true class label $y \in \mathcal{Y}$.

There are a variety of task losses used in the ordinal regression problem (Pedregosa et al., 2017). One of the most often used task losses is the *absolute loss*, which is

$$\mathcal{L}(g(\boldsymbol{x}), y) = |y - g(\boldsymbol{x})|. \tag{3}$$

The task risk (2) with the absolute loss is expressed as

$$\mathcal{R}(g) = \mathbb{E}_{X,Y}\left[|Y - g(X)|\right]. \tag{4}$$

With the property of the prediction function, the following proposition holds for the absolute loss.

Proposition 1 (Pedregosa et al. (2017))**.** *The absolute loss is equivalently expressed as*

$$|y - g(\boldsymbol{x}; f, \boldsymbol{\theta})| = \sum_{i=1}^{y-1} \mathbb{1}[\alpha_i(\boldsymbol{x}) \geq 0] + \sum_{i=y}^{K-1} \mathbb{1}[\alpha_i(\boldsymbol{x}) < 0], \tag{5}$$

where

$$\boldsymbol{\alpha} : \mathcal{X} \to \mathbb{R}^{K-1},\ \alpha_i(\boldsymbol{x}) \coloneqq \theta_i - f(\boldsymbol{x})\ (i = 1, \ldots, K-1), \tag{6}$$

and $\alpha_i(\boldsymbol{x})$ denotes the i-th element of $\boldsymbol{\alpha}(\boldsymbol{x})$.

Another task loss that has been popularly used is the *zero-one loss* (Chu and Keerthi, 2005), which is

$$\mathcal{L}(g(\boldsymbol{x}), y) = \mathbb{1}[y \neq g(\boldsymbol{x})]. \tag{7}$$

Using the definition of the prediction function, the following proposition holds for the zero-one loss.

Proposition 2 (Pedregosa et al. (2017))**.** *The zero-one loss is equivalently expressed as*

$$\mathbb{1}[y \neq g(\boldsymbol{x})] = \begin{cases} \mathbb{1}[\alpha_1(\boldsymbol{x}) < 0] & (y = 1), \\ \mathbb{1}[\alpha_{y-1}(\boldsymbol{x}) \geq 0] + \mathbb{1}[\alpha_y(\boldsymbol{x}) < 0] & (y \in \{2, \ldots, K-1\}), \\ \mathbb{1}[\alpha_{K-1}(\boldsymbol{x}) \geq 0] & (y = K). \end{cases} \tag{8}$$

The other task loss function used that is frequently used is the *squared loss* (Pedregosa et al., 2017), which is defined as

$$\mathcal{L}(g(\boldsymbol{x}), y) = (y - g(\boldsymbol{x}))^2. \tag{9}$$

The above definitions of each task loss function imply that the task losses in ordinal regression are non-continuous functions, due to the discrete nature of the prediction function. It is known that the optimization over such a non-continuous function are computationally infeasible (Feldman et al., 2012; Ben-David et al., 2003). In practice, we do not directly minimize the task risk but relax the task risk to a *surrogate* risk, which will be explained in Section 2.2.

2.2 Surrogate Losses for Ordinal Regression

We first discuss one of the *task surrogate losses* to the absolute loss called the *all threshold* (AT) loss (Pedregosa et al., 2017). Table 1 shows the notations used throughout this paper. The AT loss can be obtained by replacing the 0-1 loss $\mathbb{1}[m < 0]$ in (5) with a *binary surrogate loss* $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$. The main motivation to use a surrogate loss instead of the 0-1 loss is to make optimization easier. The binary surrogate

Table 1: Notations of losses that are used in this paper.

Task loss	$\mathcal{L}(g(\boldsymbol{x}), y)$	Task surrogate loss	$\psi(\boldsymbol{\alpha}(\boldsymbol{x}), y)$
Task risk	$\mathcal{R}(g)$	Task surrogate risk	$\mathcal{S}(g)$
0-1 loss	$\mathbb{1}[m < 0]$	Binary surrogate loss	$\ell(z)$

loss should be an optimization-friendly function and also satisfy the minimum requirement to be *Fisher-consistent*, as described in Pedregosa et al. (2017). For example, the well-known squared loss and logistic loss are valid to be employed in the AT losses in ordinal regression. Specifically, the task surrogate loss $\psi_{\mathrm{AT}} : \mathbb{R}^{K-1} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ based on (5) can be expressed as

$$\psi_{\mathrm{AT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) \coloneqq \sum_{i=1}^{y-1} \ell(-\alpha_i(\boldsymbol{x})) + \sum_{i=y}^{K-1} \ell(\alpha_i(\boldsymbol{x})). \tag{10}$$

It is known that the AT loss has good empirical performance compared with other surrogate losses in most cases (Rennie, 2005). In a similar manner, we can define the task surrogate loss to the zero-one loss called the *immediate threshold* (IT) loss as

$$\psi_{\mathrm{IT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) \coloneqq \begin{cases} \ell(\alpha_1(\boldsymbol{x})) & (y = 1), \\ \ell(-\alpha_{y-1}(\boldsymbol{x})) + \ell(\alpha_y(\boldsymbol{x})) & (y \in \{2, \ldots, K-1\}), \\ \ell(-\alpha_{K-1}(\boldsymbol{x})) & (y = K). \end{cases} \tag{11}$$

Defining $\alpha_0(\boldsymbol{x}) = \alpha_K(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X}$, we may write $\psi_{\mathrm{IT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) = \ell(-\alpha_{y-1}(\boldsymbol{x})) + \ell(\alpha_y(\boldsymbol{x}))$ for the notational simplicity. The task surrogate loss to the squared loss called the *least-squares* (LS) loss does not rely on the binary surrogate loss, and is defined as

$$\psi_{\mathrm{LS}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) \coloneqq \left(y + \alpha_1(\boldsymbol{x}) - \frac{3}{2}\right)^2. \tag{12}$$

In general, the task surrogate risk guided by a task surrogate loss $\psi : \mathbb{R}^{K-1} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ can be written as

$$\mathcal{S}(g) \coloneqq \mathbb{E}_{X,Y}\left[\psi(\boldsymbol{\alpha}(X), Y)\right]. \tag{13}$$

At glance, the notation of (13) can be confusing since g does not appear in the right-hand side. However, g contains f and $\boldsymbol{\theta}$ as parameters, and $\boldsymbol{\alpha}$ is defined by f and $\boldsymbol{\theta}$. Thus, we adopt this notation in this paper. In supervised ordinal regression, we are given labeled data drawn independently from the joint density p. Then, the risk (4) can be naively minimized by the ERM framework. Note that the minimization is run over the decision function f and the thresholds $\boldsymbol{\theta}$ for the AT and IT losses, while it is run only over f for the LS loss. Note also that regularization schemes can be applied if necessary.

In the real world, collecting labeled training data can be costly. Thus, it is preferable to incorporate unlabeled data to train an ordinal regressor. We can see that direct empirical estimation of the risk term in (4) cannot utilize unlabeled data. In this paper, we mitigate this problem by extending the ERM framework to semi-supervised ordinal regression.

3 Proposed Framework

In this section, we introduce our new formulation for semi-supervised ordinal regression.

Unbiased Risk Estimator. We first derive a risk estimator within the proposed ERM framework. Then, we theoretically investigate the behavior of our risk estimator. Here, we assume that we are given the following data:

$$\mathcal{X}_{\mathrm{L}} := \{(\boldsymbol{x}_j^{\mathrm{L}}, y_j)\}_{j=1}^{n_{\mathrm{L}}} \overset{\text{i.i.d.}}{\sim} p(X,Y), \quad \mathcal{X}_{\mathrm{U}} := \{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p(X) = \sum_{y=1}^{K} \pi_y \, p(X|Y=y), \tag{14}$$

where n_{L} and n_{U} denote the number of labeled data and unlabeled data, respectively. Let n_y denote the number of labeled data in class y, i.e., $n_{\mathrm{L}} = \sum_{y=1}^{K} n_y$, and $\pi_y := p(Y = y)$ denote the class-prior probability of the class y such that $\sum_{y=1}^{K} \pi_y = 1$.

As discussed in Section 2.2, it is not straightforward to incorporate unlabeled data into the task surrogate risk (13). To handle this problem, we propose to find an equivalent expression of the task surrogate risk (13) so that we can obtain an unbiased risk estimator that uses both unlabeled and labeled data. Our following lemma states that we can rewrite the task surrogate risk to contain the expectations over $K-1$ classes of labeled data and the expectation over the marginal density $p(X)$.

Lemma 3. *For any $k \in \{1,\dots,K\}$, the task surrogate risk (13) is equivalently expressed as*

$$\mathcal{S}_{\mathrm{LU}}^{\backslash k}(g) = \sum_{y \in \mathcal{Y}^{\backslash k}} \pi_y \mathbb{E}_{X|Y=y}\left[\psi(\boldsymbol{\alpha}(X), y)\right] + \mathbb{E}_{\mathrm{U}}\left[\psi(\boldsymbol{\alpha}(X), k)\right] - \sum_{y \in \mathcal{Y}^{\backslash k}} \pi_y \mathbb{E}_{X|Y=y}\left[\psi(\boldsymbol{\alpha}(X), k)\right], \tag{15}$$

where $\mathbb{E}_{\mathrm{U}}[\cdot]$ denotes the expectation over unlabeled data, $\mathcal{Y}^{\backslash k} := \mathcal{Y} \backslash \{k\}$, and LU *stands for "*Labeled-Unlabeled*".*

The proof of Lemma 3 is given in Appendix A. With the risk obtained from Lemma 3, we can derive the following unbiased risk estimator for the task surrogate risk (13):

$$\widehat{\mathcal{S}}_{\mathrm{LU}}^{\backslash k}(g) := \underbrace{\sum_{y \in \mathcal{Y}^{\backslash k}} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), y)}_{=:\widehat{\mathcal{S}}_{\mathrm{L}}^{(1)}(g)} + \underbrace{\frac{1}{n_{\mathrm{U}}} \sum_{j=1}^{n_{\mathrm{U}}} \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^{\mathrm{U}}), k)}_{=:\widehat{\mathcal{S}}_{\mathrm{U}}(g)} - \underbrace{\sum_{y \in \mathcal{Y}^{\backslash k}} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), k)}_{=:\widehat{\mathcal{S}}_{\mathrm{L}}^{(2)}(g)}. \tag{16}$$

We can interpret the risk estimator in (16) as follows. The first term on the right-hand side of (16) indicates that labeled data that are not from class k should be predicted correctly. The second term indicates that unlabeled data should be predicted as k. The purpose of the third term is to cancel the bias from the second term by subtracting the risk of all labeled data that are not from class k to be predicted as class k. Lemma 3 is inspired by the technique used in weakly-supervised learning, such as binary classification from positive and unlabeled data and semi-supervised binary classification, which have been shown to be effective (du Plessis et al., 2015; Sakai et al., 2017).

Although we can obtain a risk estimator that can utilize unlabeled data, we still cannot make full use of all given data since our risk estimator in (16) ignores labeled data from class k. To mitigate this problem, inspired by the work on semi-supervised learning for binary classification (Sakai et al., 2017), we propose to combine the risk estimator of supervised ordinal regression with our risk estimator in (16) by the convex combination as follows.

Theorem 4. *For any $k \in \mathcal{Y}$ and $\gamma \in [0,1]$, the task surrogate risk (13) is equivalently expressed as*

$$\mathcal{S}_{\text{SEMI-}\gamma}^{\backslash k}(g) := \gamma \mathcal{S}_{\mathrm{LU}}^{\backslash k}(g) + (1-\gamma)\mathcal{S}(g). \tag{17}$$

Theorem 4 directly follows from Lemma 3. It is worth noting that our risk (17) is equivalent to the ordinary surrogate risk in (13). Therefore, the theory of the Fisher-consistency of surrogate losses and excess risk bounds in the ordinary ordinal regression (Pedregosa et al., 2017) are directly applicable to our framework, which will be discussed in the next section.

Roles of Unlabeled Data. The proposed risk does not use geometric information of unlabeled data; thus, it is applicable even when a specific geometric assumption does not hold. Then, an important question to be answered is, "*how does unlabeled data help us obtain a good ordinal regressor?*" In the training phase, by introducing unlabeled data, we can expect that the variance of the empirical risk is decreased, resulting in more accurate risk estimation. Also, in the validation phase, the variance of the validation risk estimator is reduced, helping us to select good hyper-parameters.

The removed class k can be determined arbitrarily as shown in Lemma 3, and the performance can be depending on the choice of k. We will investigate strategies to select the class k in Section 5.1 based on the theory discussed in the next section.

4 Theoretical Analysis

This section establishes the Fisher-consistency and estimation error bounds to elucidate that our risk estimator $\widehat{\mathcal{S}}^{\setminus k}_{\text{SEMI-}\gamma}(g)$ has theoretical guarantees. Besides, we provide a theoretical analysis of variance reduction, which shows that the variance of the empirical semi-supervised risk can be smaller than that of the empirical supervised risk.

4.1 Fisher-Consistency

To ensure that the minimizer of our task surrogate risk $\mathcal{S}^{\setminus k}_{\text{SEMI-}\gamma}$ can give the optimal solution to the task risk (2), we give the following proposition.

Proposition 5 (Fisher-Consistency)**.** *Let the task surrogate loss ψ be the AT loss, IT loss, or LS loss. Assume for the AT and IT losses that the binary surrogate loss ℓ is convex, differentiable at 0, and $\ell'(0) < 0$. Then, every minimizer g of $\mathcal{S}^{\setminus k}_{\text{SEMI-}\gamma}$ reaches Bayes optimal risk* $\inf\{\mathcal{R}(g) : \text{ *all measurable functions* } g\}$.

The proof of Proposition 5 is given in Appendix B. Although we avoid going into the details, Proposition 5 also holds for the cumulative link and least absolute deviation losses, we can prove Fisher-consistency for them in a similar manner (see Pedregosa et al. (2017) for the definitions of theses losses). With Proposition 5, we can clarify that the following estimation error bound is applicable to a wide range of task surrogate losses.

4.2 Estimation Error Bound

To ensure that our empirical risk estimator $\widehat{\mathcal{S}}^{\setminus k}_{\text{SEMI-}\gamma}(g)$ is consistent to the risk $\mathcal{S}^{\setminus k}_{\text{SEMI-}\gamma}(g)$ and the empirical risk minimizer approaches the true risk minimizer, we establish estimation error bounds here. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a class of specified decision functions. Then, we will consider a distribution-dependent complexity measure of functions, called the *expected Rademacher complexity* (Bartlett and Mendelson, 2002).

Definition 6 (Expected Rademacher complexity)**.** *Let n be a positive integer, $Z_1, \dots, Z_n$ be i.i.d. random variables drawn from a probability distribution with density p over a set $\mathcal{Z}$, $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ be a family of functions from $\mathcal{Z}$ to $\mathbb{R}$, and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ be random variables, which take $+1$ and -1 with equal probabilities. Then the expected Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathfrak{R}(\mathcal{F}; n, p) := \mathbb{E}_{Z_1,\dots,Z_n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right].$$

Intuitively, the Rademacher complexity quantifies how much our decision function class $\mathcal{F}$ can correlate to the random noise. Thus, a large Rademacher complexity indicates that a decision function is highly flexible to fit the noise. This complexity term is an important tool to derive an estimation error bound (more details about the measures of the complexity of a hypothesis class can be found in Bartlett and Mendelson (2002); Shalev-Shwartz and Ben-David (2014); Mohri et al. (2018)).

To formally define the Rademacher complexity in the context of ordinal regression, we introduce the following definitions.

$$\begin{aligned}\Theta &:= \{\boldsymbol{\theta} \in \mathbb{R}^{K-1} : \theta_1 \leq \theta_2 \leq \cdots \leq \theta_{K-1},\ \|\boldsymbol{\theta}\|_2 \leq C_{\boldsymbol{\theta}}\},\\ \mathcal{G} &:= \{(f, \boldsymbol{\theta}) : f \in \mathcal{F},\ \boldsymbol{\theta} \in \Theta\},\\ \mathcal{A}_i &:= \{\alpha_i(\boldsymbol{x}) = \theta_i - f(\boldsymbol{x}) : f \in \mathcal{F},\ \theta \in \Theta\} \text{ for } i \in \{0, 1, \ldots, K\}.\end{aligned}$$

Note that $\mathfrak{R}(\mathcal{A}_0) = \mathfrak{R}(\mathcal{A}_K) = 0$ from the extended definition of $\boldsymbol{\alpha}$. Let

$$g^* := \mathop{\arg\min}_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(g) (= \mathop{\arg\min}_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \mathcal{S}(g)) \tag{18}$$

be the true risk minimizer, and

$$\widehat{g}^{\backslash k} := \mathop{\arg\min}_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma} (\not\equiv \mathop{\arg\min}_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \widehat{\mathcal{S}}_{\psi}(g)) \tag{19}$$

be the empirical risk minimizer. Note that $g^*, \widehat{g}^{\backslash k} \in \mathcal{G}$ are pairs of some $f \in \mathcal{F}$ and $\boldsymbol{\theta} \in \Theta$, respectively. We also define the hypothesis classes guided by each task surrogate loss as

$$\begin{aligned}\mathcal{H}_{\text{AT}} &:= \{(\boldsymbol{x}, y) \mapsto \psi_{\text{AT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\},\\ \mathcal{H}_{\text{IT}} &:= \{(\boldsymbol{x}, y) \mapsto \psi_{\text{IT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\},\\ \mathcal{H}_{\text{LS}} &:= \{(\boldsymbol{x}, y) \mapsto \psi_{\text{LS}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}\}.\end{aligned} \tag{20}$$

In a similar manner, we define the the hypothesis classes guided by each task surrogate loss for fixed label $y \in \mathcal{Y}$ as

$$\begin{aligned}\mathcal{H}^{(y)}_{\text{AT}} &:= \{\boldsymbol{x} \mapsto \psi_{\text{AT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\},\\ \mathcal{H}^{(y)}_{\text{IT}} &:= \{\boldsymbol{x} \mapsto \psi_{\text{IT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\},\\ \mathcal{H}^{(y)}_{\text{LS}} &:= \{\boldsymbol{x} \mapsto \psi_{\text{LS}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}\}.\end{aligned} \tag{21}$$

We use $\mathcal{H}$ (reps. $\mathcal{H}^{(y)}$) instead of $\mathcal{H}_{\text{AT}}, \mathcal{H}_{\text{IT}}$, and $\mathcal{H}_{\text{LS}}$ (reps. $\mathcal{H}^{(y)}_{\text{AT}}, \mathcal{H}^{(y)}_{\text{IT}}$, and $\mathcal{H}^{(y)}_{\text{LS}}$), when a statement holds without depending on the used task surrogate loss. At a glance, some definitions may look redundant, but they are needed to investigate properties specific to each of the AT, IT, and LS losses.

To the best of our knowledge, the Rademacher complexity in the context of the ordinal regression problem has never been investigated. The next two theorems establish upper bounds for the guided hypothesis classes.

Theorem 7. *Fix $y \in \mathcal{Y}$. Let $n \in \mathbb{N}$ and assume for the AT and IT losses that the binary surrogate loss ℓ is ρ-Lipschitz. Then, the expected Rademacher complexities of $\mathcal{H}^{(y)}_{\text{AT}}, \mathcal{H}^{(y)}_{\text{IT}}$, and $\mathcal{H}^{(y)}_{\text{LS}}$ are bounded as*

$$\begin{aligned}\mathfrak{R}(\mathcal{H}^{(y)}_{\text{AT}}; n) &\leq \rho \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n),\\ \mathfrak{R}(\mathcal{H}^{(y)}_{\text{IT}}; n) &\leq \rho \left(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)\right),\\ \mathfrak{R}(\mathcal{H}^{(y)}_{\text{LS}}; n) &\leq 2|y + \theta_1 - 3/2| \mathfrak{R}(\mathcal{F}; n) + \mathfrak{R}(\text{sq} \circ \mathcal{F}; n),\end{aligned} \tag{22}$$

where $\text{sq} : \text{Im} f \ni z \mapsto z^2 \in \mathbb{R}$.

Theorem 8 (Upper bounds of Rademacher complexities of task surrogate losses)**.** *Let $n \in \mathbb{N}$ and assume for the AT and IT losses that the binary surrogate loss ℓ is ρ-Lipschitz. Then, the expected Rademacher complexities of $\mathcal{H}_{\mathrm{AT}}$, $\mathcal{H}_{\mathrm{IT}}$, and $\mathcal{H}_{\mathrm{LS}}$ are bounded as*

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}_{\mathrm{AT}}; n) &\leq \rho\, K \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n), \\
\mathfrak{R}(\mathcal{H}_{\mathrm{IT}}; n) &\leq \rho \sum_{y \in \mathcal{Y}} \left(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)\right), \\
\mathfrak{R}(\mathcal{H}_{\mathrm{LS}}; n) &\leq 2(K + |\theta_1 - 3/2|)\mathfrak{R}(\mathcal{F}; n) + \mathfrak{R}(\mathrm{sq} \circ \mathcal{F}; n),
\end{aligned} \tag{23}$$

where $\mathrm{sq} : \mathrm{Im}\, f \ni z \mapsto z^2 \in \mathbb{R}$.

Remark. The bound on $\mathfrak{R}(\mathcal{H}_{\mathrm{LS}}; n)$ is given using $\mathfrak{R}(\mathcal{F}; n)$ instead of $\mathfrak{R}(\mathcal{A}_j; n)$. This is because, in the LS loss, the threshold is fixed.

The proofs of Theorems 7 and 8 are given in Appendix E. Both theorems state that the Rademacher complexities with the guided hypothesis classes can be bounded by the Rademacher complexity of the underlying decision functions class $\mathcal{F}$, and show how $\mathfrak{R}(\mathcal{H}; n_y)$ of each task surrogate loss depends on the problem-dependent parameters. Then, the next theorem establishes estimation error bounds with a class of general decision functions based on Theorems 7 and 8.

Theorem 9 (Estimation error bounds with general decision functions)**.** *Assume for the AT and IT losses that the binary surrogate loss ℓ is ρ-Lipschitz, and that there exists a constant $C_\psi > 0$ such that $\psi(y, \boldsymbol{\alpha}) \leq C_\psi$ for any $y \in \mathcal{Y}$, $\boldsymbol{\alpha} \in \mathbb{R}^{K-1}$. Then, for any $k \in \mathcal{Y}$ and $\delta \in (0, 1]$, with probability at least $1 - \delta$,*

$$\begin{aligned}
&\mathcal{S}(\widehat{g}^{\setminus k}) - \mathcal{S}(g^*) \\
&\leq 4\left(\gamma\left(\sum_{y \in \mathcal{Y}^{\setminus k}} \pi_y \left(\mathfrak{R}_{\mathrm{UB}}^{(y)}(n_y) + \mathfrak{R}_{\mathrm{UB}}^{(k)}(n_y)\right) + \mathfrak{R}_{\mathrm{UB}}^{(k)}(n_{\mathrm{U}})\right) + (1-\gamma)\mathfrak{R}_{\mathrm{UB}}(n_{\mathrm{L}})\right) \\
&\quad + 2\sqrt{2}C_\psi \left(\gamma\left(2\sum_{y \in \mathcal{Y}^{\setminus k}} \frac{\pi_y}{\sqrt{n_y}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right) + (1-\gamma)\frac{1}{\sqrt{n_{\mathrm{L}}}}\right)\sqrt{\ln \frac{2(K+1)}{\delta}},
\end{aligned} \tag{24}$$

where $\mathfrak{R}_{\mathrm{UB}}^{(y)}(n)$ for $y \in \mathcal{Y}$ and $\mathfrak{R}_{\mathrm{UB}}(n)$ are defined as

$$\mathfrak{R}_{\mathrm{UB}}^{(y)}(n) := \begin{cases} \rho \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n) & (\textit{AT loss}), \\ \rho\, \left(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)\right) & (\textit{IT loss}), \\ 2\,|y + \theta_1 - 3/2|\,\mathfrak{R}(\mathcal{F}; n) + \mathfrak{R}(\mathrm{sq} \circ \mathcal{F}; n) & (\textit{LS loss}), \end{cases}$$

$$\mathfrak{R}_{\mathrm{UB}}(n) := \begin{cases} \rho K \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n) & (\textit{AT loss}), \\ \rho \sum_{y \in \mathcal{Y}} \left(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)\right) & (\textit{IT loss}), \\ 2(K + |\theta_1 - 3/2|)\mathfrak{R}(\mathcal{F}; n) + \mathfrak{R}(\mathrm{sq} \circ \mathcal{F}; n) & (\textit{LS loss}). \end{cases}$$

The proof of Theorem 9 is given in Appendix D. Note that the Rademacher complexity with a class of underlying decision functions $\mathcal{F}$ can be bounded in many decision function classes as discussed in Niu et al. (2016); Bao et al. (2018); Mohri et al. (2018), and will be given for a linear-in-parameter models (Lemma 21 in Appendix E).

Next, we restrict the decision function class $\mathcal{F}$ to a class of *linear-in-parameter* models and see more in detail the behavior of the Rademacher complexities with hypothesis classes guided by each task surrogate loss. The class of linear-in-parameter models is defined as

$$\mathcal{F} = \{f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) \colon \|\boldsymbol{w}\| \leq C_{\boldsymbol{w}},\ \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \leq C_{\boldsymbol{\phi}}\} \tag{25}$$

for positive constants $C_{\boldsymbol{w}}$ and $C_{\boldsymbol{\phi}}$, where $\boldsymbol{w} \in \mathbb{R}^b$ is a parameter and $\boldsymbol{\phi} : \mathbb{R}^{d+1} \to \mathbb{R}^b$ is a basis function. We assume that the bias parameter is included in $\boldsymbol{w}$. We can prove the following lemma using the theorem on the Rademacher complexity bound for ordinal regression (Theorem 8) and a well-known upper bound for the Rademacher complexity with linear-in-parameter models (Lemma 21 in Appendix E).

Lemma 10. *Let $n \in \mathbb{N}$ and assume for the AT and IT losses that the binary surrogate loss ℓ is ρ-Lipschitz. Then, the expected Rademacher complexities of $\mathcal{H}_{\mathrm{AT}}^{(y)}$, $\mathcal{H}_{\mathrm{IT}}^{(y)}$, $\mathcal{H}_{\mathrm{LS}}^{(y)}$, $\mathcal{H}_{\mathrm{AT}}$, $\mathcal{H}_{\mathrm{IT}}$, and $\mathcal{H}_{\mathrm{LS}}$ with the class of linear-in-parameter models are bounded as*

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}_{\mathrm{AT}}^{(y)}; n) &\leq \frac{(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho\, K}{\sqrt{n}}, \\
\mathfrak{R}(\mathcal{H}_{\mathrm{IT}}^{(y)}; n) &\leq \frac{2(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho}{\sqrt{n}}, \\
\mathfrak{R}(\mathcal{H}_{\mathrm{LS}}^{(y)}; n) &\leq \frac{2(|y + \theta_1 - 3/2| + C_{\boldsymbol{w}} C_{\boldsymbol{\phi}})(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}} K}{\sqrt{n}}, \\
\mathfrak{R}(\mathcal{H}_{\mathrm{AT}}; n) &\leq \frac{(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho\, K(K-1)}{\sqrt{n}}, \\
\mathfrak{R}(\mathcal{H}_{\mathrm{IT}}; n) &\leq \frac{2(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho\, K}{\sqrt{n}}, \\
\mathfrak{R}(\mathcal{H}_{\mathrm{LS}}; n) &\leq \frac{2(K + |\theta_1 - 3/2| + C_{\boldsymbol{w}} C_{\boldsymbol{\phi}})(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}} K}{\sqrt{n}}.
\end{aligned} \tag{26}$$

The proof of Lemma 10 is given in Appendix E. Combining Theorem 9 and Lemma 10 establishes the following estimation error bounds for the linear-in-parameter models.

Corollary 11 (Estimation error bounds with linear-in-parameter models)**.** *Assume for the AT and IT losses that the binary surrogate loss ℓ is ρ-Lipschitz and that there exists a constant $C_\psi > 0$ such that $\psi(y, \boldsymbol{\alpha}) \leq C_\psi$ for any $y \in \mathcal{Y}$, $\boldsymbol{\alpha} \in \mathbb{R}^{K-1}$. Then, for any $k \in \mathcal{Y}$ and $\delta \in (0, 1]$, with probability at least $1 - \delta$,*

$$\begin{aligned}
&\mathcal{S}(\widehat{g}^{\setminus k}) - \mathcal{S}(g^*) \\
&\leq 4\left(\gamma\left(\sum_{y \in \mathcal{Y}^{\setminus k}} \pi_y \left(\frac{C^{(y)}}{\sqrt{n_y}} + \frac{C^{(k)}}{\sqrt{n_y}}\right) + \frac{C^{(k)}}{\sqrt{n_{\mathrm{U}}}}\right) + (1-\gamma)\frac{C}{\sqrt{n_{\mathrm{L}}}}\right) \\
&\quad + 2\sqrt{2}C_\psi \left(\gamma\left(2\sum_{y \in \mathcal{Y}^{\setminus k}} \frac{\pi_y}{\sqrt{n_y}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right) + (1-\gamma)\frac{1}{\sqrt{n_{\mathrm{L}}}}\right)\sqrt{\ln \frac{2(K+1)}{\delta}},
\end{aligned} \tag{27}$$

where $C^{(y)}$ for $y \in \mathcal{Y}$ and C are constants defined as

$$C^{(y)} := \begin{cases} (C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho\, K & (\textit{AT loss}), \\ 2(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho & (\textit{IT loss}), \\ 2(|y + \theta_1 - 3/2| + C_{\boldsymbol{w}} C_{\boldsymbol{\phi}})(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}} K & (\textit{LS loss}), \end{cases} \tag{28}$$

$$C := \begin{cases} (C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho\, K(K-1) & (\textit{AT loss}), \\ 2(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}}\, \rho\, K & (\textit{IT loss}), \\ 2(K + |\theta_1 - 3/2| + C_{\boldsymbol{w}} C_{\boldsymbol{\phi}})(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}} K & (\textit{LS loss}). \end{cases} \tag{29}$$

Corollary 11 shows that our proposed risk estimator is consistent, i.e., $\mathcal{S}(\widehat{g}^{\setminus k}) \to \mathcal{S}(g^*)$ as $n_y \to \infty$ $(y = 1, \dots, K)$ and $n_{\mathrm{U}} \to \infty$. The convergence rate is

$$\mathcal{O}_p\left(\sum_{y \in \mathcal{Y}^{\setminus k}} \frac{1}{\sqrt{n_y}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right),$$

where $\mathcal{O}_p$ denotes the order in probability. This order is the optimal parametric rate for empirical risk minimization without any additional assumption (Mendelson, 2008).

4.3 Variance Reduction

Here, we conduct an analysis of variance reduction to emphasize why incorporating unlabeled data to our risk estimator improves the performance. The following theorem indicates that the variance of the empirical semi-supervised risk can be smaller than that of the empirical supervised risk.

Theorem 12 (Variance reduction)**.** *Fix any $g \in \mathcal{G}$ and assume that $\gamma \in [0, 1]$ satisfies*

$$0 < \gamma < \frac{2\left(\mathrm{Var}[\widehat{\mathcal{S}}(g)] - \mathrm{Cov}(\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g), \widehat{\mathcal{S}}(g))\right)}{\mathrm{Var}[\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g)] + \mathrm{Var}[\widehat{\mathcal{S}}(g)] - 2\,\mathrm{Cov}(\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g), \widehat{\mathcal{S}}(g))}.$$

Then,

$$\mathrm{Var}[\widehat{\mathcal{S}}_{\mathrm{SEMI}\text{-}\gamma}^{\setminus k}(g)] < \mathrm{Var}[\widehat{\mathcal{S}}(g)].$$

The proof of Theorem 12 is given in Appendix F. This theorem implies that if we select γ properly, the variance of the empirical semi-supervised risk is strictly smaller than that of the empirical supervised risk. Since the empirical semi-supervised risk is unbiased and has a smaller variance than the standard supervised risk, the former risk estimator is expected to be more accurate and stable, as we will see experimentally in Section 6.

5 Practical Implementation

With Theorem 4, we can obtain a risk estimator that can fully use both labeled and unlabeled data. It is also straightforward to see that our risk estimator based on Theorem 4 is unbiased. However, to use our risk estimator effectively in practice, one important question is: *how can we decide the class k to calculate $\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g)$?* We discuss strategies to handle this problem theoretically.

5.1 Strategies to Remove One Class for $\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g)$

If the choice of k is highly sensitive, it could be cumbersome to tune this hyperparameter as the number of classes K increases. Here, we consider two strategies for selecting a class k to removed in $\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g)$ motivated by the property of the estimator. We provide two strategies.

The first strategy is based on finite sample estimation error. More specifically, when we approximate the expectation term by a limited number of samples, the variance of the estimator can be large. Then a naive strategy would be to remove the class that contains the smallest number of labeled data as

$$k = \operatorname*{arg\,min}_{y \in \mathcal{Y}} n_y. \tag{30}$$

The other strategy is based on the estimation error bound. As discussed in Theorem 9, the convergence rate of the estimation error for $\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g)$ is $\mathcal{O}_p(\sum_{y \in \mathcal{Y}^{\setminus k}} \pi_y/\sqrt{n_y})$ under the assumptions that we have enough unlabeled data $(n_{\mathrm{U}} \to \infty)$. This implies the following proposition which provides a strategy to decide the removed class k.

Proposition 13. *Assume that labeled data are obtained under the assumption in (14), and the above assumptions are satisfied. Then,*

$$k = \arg\max_{y \in \mathcal{Y}} n_y \tag{31}$$

gives the lowest upper bound of the estimation error as $n_{\mathrm{L}} \to \infty$.

It is interesting to see that the above two strategies give completely opposite solutions as the removed class. We will experimentally compare these strategies in Section 6, where we find that both strategies performed similarly although they are completely different, indicating that the choice of k may not be critical to the performance.

5.2 Order Constraints

In ordinal regression problems, the threshold parameters should be ordered, i.e., $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{K-1}$ (Pedregosa et al., 2017). Here, we introduce a simple trick to constrain the threshold parameters $\boldsymbol{\theta}$ by adding the term

$$\Omega(\boldsymbol{\theta}) := \mu \sum_{i=1}^{K-2} \max\left\{0,\ -\log(\theta_{i+1} - \theta_i)\right\}, \tag{32}$$

where $\mu \geq 0$ is a regularization parameter for the order constraints. We show that this simple trick works well in the experiments. In fact, even without any regularization term on $\boldsymbol{\theta}$, we empirically observed that the values of threshold constraints $\boldsymbol{\theta}$ are usually ordered. This observation suggests that the order constraints are not difficult to satisfy when optimizing $\widehat{\mathbb{S}}^{\backslash k}_{\text{SEMI-}\gamma}$. This threshold regularization scheme is based on the idea of the log-barrier method (Boyd and Vandenberghe, 2004), in which we impose a high cost $-\log(-c(\boldsymbol{\theta}))$ to the objective function for an inequality constraint $c(\boldsymbol{\theta}) \leq 0$ for $c : \Theta \to \mathbb{R}$. However, only using $-\log(\theta_{i+1} - \theta_i)$ may lead to the following two problems. First, the algorithm may make $\theta_{i+1} - \theta_i$ large more than necessary while $\theta_{i+1} - \theta_i \geq 0$ is only required. Second, the function $-\log(\theta_{i+1} - \theta_i)$ becomes negative if $\theta_{i+1} - \theta_i > 1$. This may cause an overall risk to be arbitrarily negative by making $\theta_{i+1} - \theta_i$ larger, and cause instability in the optimization procedure when combining it with the empirical version of (15). As a result, we further introduce $\max\{0, \cdot\}$ in (32) to mitigate these problems.

Here, we investigate the objective function when the linear-in-parameter model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$ is employed as f. Our next theorem states a sufficient condition to guarantee that, for a certain task surrogate loss function, the optimization problem is convex with respect to both the model and order constraints parameters.

Theorem 14. *Let C_ℓ be a positive constant. If the AT or LS losses are adopted as the task surrogate loss, and the binary surrogate loss $\ell(z)$ is convex and satisfies*

$$\ell(z) - \ell(-z) = -C_\ell z, \tag{33}$$

then the objective function $\widehat{J}_\ell(\boldsymbol{w}, \boldsymbol{\theta}) := \widehat{\mathbb{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g) + \Omega(\boldsymbol{\theta})$ is convex with respect to $\boldsymbol{w}$ and $\boldsymbol{\theta}$.

The proof of Theorem 14 is given in Appendix G. The condition (33) is known as the linear-odd condition (Patrini et al., 2016). Examples of binary surrogate losses are shown in Figure 1 and Table 2. In our experiments, we used the logistic loss as the binary surrogate loss. Note that the objective function with the IT loss is not always convex.

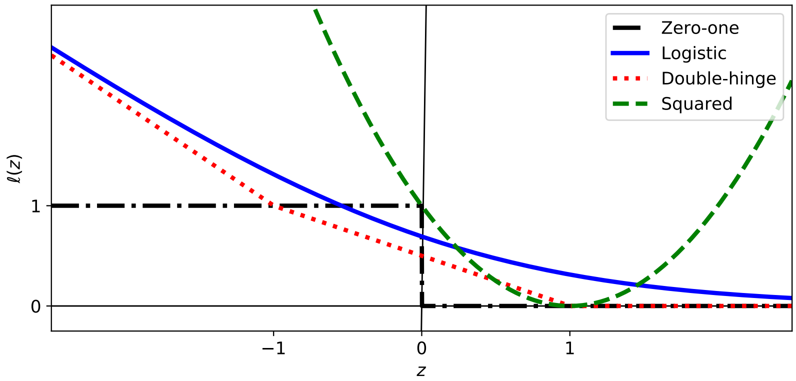


Figure 1: Examples of the binary surrogate losses satisfying (33). The zero-one function is depicted for reference.

Table 2: Examples of binary surrogate losses and their satisfiability of the sufficient condition in (33).

Loss name	$\ell(z)$	Eq. (33)
Hinge	$\max(0, 1-z)$	No
Exponential	$\exp(-z)$	No
Logistic	$\log(1+e^{-z})$	Yes
Double-hinge	$\max(-z, \max(0, \frac{1}{2}-\frac{1}{2}z))$	Yes
Squared	$(1-z)^2$	Yes

5.3 Non-Negative Risk Estimator

We discuss a risk modification technique called a *non-negative risk estimator*. This technique was originally proposed in Kiryo et al. (2017) for classification from positive and unlabeled data based on an unbiased risk estimator. We can see that the sum of the second and third terms of our empirical risk estimator (16), $\widehat{\mathbb{S}}_{\mathrm{U}}(g) - \widehat{\mathbb{S}}_{\mathrm{L}}^{(2)}(g)$, can be negative while the corresponding expected risk $\mathbb{S}_{\mathrm{U}}(g) - \mathbb{S}_{\mathrm{L}}^{(2)}(g) = \pi_k \mathbb{E}_{X|Y=k}\left[\psi(\boldsymbol{\alpha}(X), k)\right]$ is always non-negative. This observation suggests that the model can excessively reduce the empirical risk by maximizing $\widehat{\mathbb{S}}_{\mathrm{L}}^{(2)}(g)$. To prevent this problem, following Kiryo et al. (2017), we modify the empirical risk estimator (16) as

$$\widehat{\mathbb{S}}_{\mathrm{L}}^{(1)}(g) + \max\left\{0, \widehat{\mathbb{S}}_{\mathrm{U}}(g) - \widehat{\mathbb{S}}_{\mathrm{L}}^{(2)}(g)\right\}. \tag{34}$$

This approach is later generalized in Lu et al. (2020) in the context of unlabeled-unlabeled classification. Specifically, they considered the LeakyReLU-like regularization function instead of the max operator used in (34). They empirically showed that this regularization technique makes trained models generalize better. Following this observation, we proposed to modify our empirical risk estimator (16) as

$$\widehat{\mathbb{S}}_{\mathrm{L}}^{(1)}(g) + \mathrm{GLReLU}\left(\widehat{\mathbb{S}}_{\mathrm{U}}(g) - \widehat{\mathbb{S}}_{\mathrm{L}}^{(2)}(g)\right), \tag{35}$$

where the Generalized Leaky ReLU function $\mathrm{GLReLU} : \mathbb{R} \to \mathbb{R}$ is defined as $\mathrm{GLReLU}(x) := x\mathbb{1}[x \geq 0] + \lambda x \mathbb{1}[x < 0]$ for $\lambda \leq 0$. We will employ this regularization technique in experiments.

6 Experiments

In this section, we begin by numerically investigating that unlabeled data helps reduce the variance of the risk estimator as theoretically discussed in Section 4. Besides, we experimentally investigate which strategy is better to remove one class for $\widehat{\mathbb{S}}_{\mathrm{LU}}^{\setminus k}(g)$ as discussed in Section 5.1. Finally, we present the experimental comparison results of semi-supervised ordinal regression on benchmark datasets.

Common Setup. As evaluation metrics, we adopted the *mean absolute error*, *mean zero-one error*, and *mean squared error*. Note that each metric coincides with the task surrogate risk, which adopts the AT, IT, and LS losses as the task surrogate loss, respectively. We did experiments on the AT, IT, and LS losses as the task surrogate loss and the logistic loss as the binary surrogate loss. For validation of the hyperparameters, we used the hold-out method by splitting the training data set with the ratio of $2:1$. For both models, we fixed the hyper-parameters γ and μ to 0.5 and 10, respectively. Note that we empirically observed that if μ is large enough, μ is insensitive to the performance. We ran the experiments 20 times to calculate the mean and standard error of the performances. Also, we used the non-negative risk estimator described in Section 5.3 with $\lambda = -0.2$ to prevent over-fitting. We used Chainer (Tokui et al., 2019) to implement our models. We obtained datasets from a survey paper on ordinal regression (Gutierrez et al., 2016), and the website on ordinal regression benchmark data[1]. The detail of the dataset description is given in the following.

Baselines. We compared our proposed methods (SEMI) against the following baselines.

1. **Supervised Ordinal Regression (SV):** Supervised ordinal regression here is based on empirical risk minimization of the task surrogate risk, which is described in Pedregosa et al. (2017) and Section 2. Recall that the different task surrogate losses are used for each objective function. In other words, the AT, IT, and LS losses are used when the mean absolute error, mean zero-one error, and mean squared error are used as evaluation metrics, respectively.

2. **Transductive Ordinal Regression (TOR):** Transductive Ordinal Regression (Seah et al., 2012) is a method based on pseudo-labeling, and TOR assumes that the data satisfies the cluster assumption. TOR tries to minimize objective function by repeatedly minimizing the sum of risks on labeled data and pseudo-labeled data. Note that the classifier obtained by the training can be used as an inductive classifier. We noticed that the TOR is essentially a pseudo-labeling method based on supervised empirical risk minimization with AT loss, while it is not mentioned in the paper, implying that we can extend the use IT and LS loss instead of AT loss. Thus, we used IT and LS loss in our experiments when adopting mean zero-one error and mean squared error, respectively.

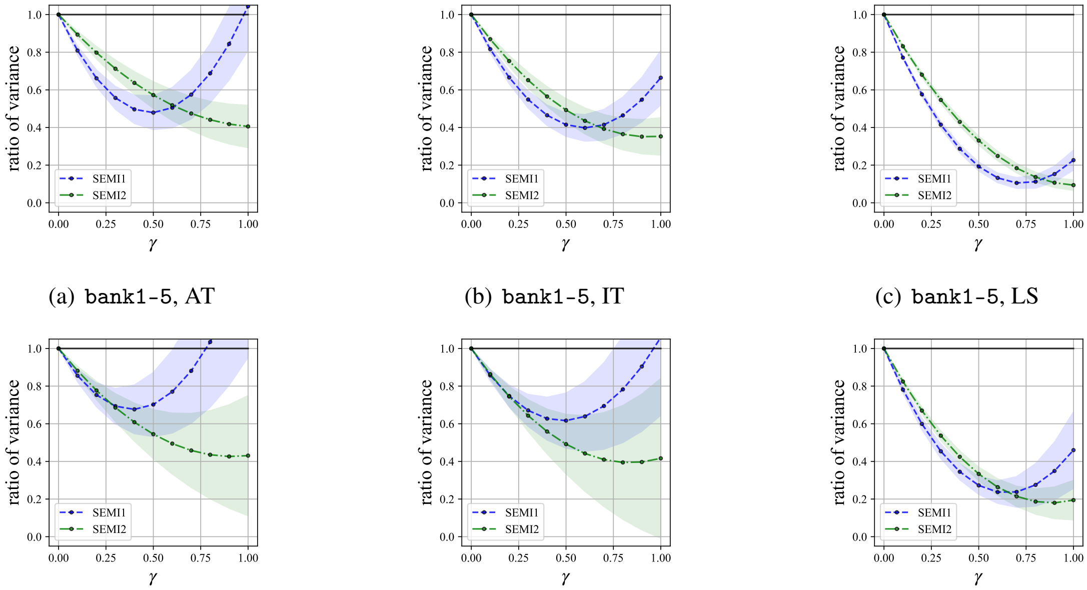
3. **Semi-Supervised Manifold Ordinal Regression (SSMOR):** Semi-Supervised Manifold Ordinal Regression (Liu et al., 2011) is a method based on the manifold assumption. SSMOR first minimizes the objective function, which is the sum of smoothness constraint (nearby points should have the same label) and fitting constraint (fitting on the given labels). After optimizing the objective function, SSMOR tries to obtain labels for new data points by calculating thresholds in a heuristic manner. However, we found that the threshold calculated via this approach can be not ordered, and its performance is poor. Thus, we first sorted the data points based on the prediction score and then allocated labels with the class prior.

6.1 Variance Reduction and Performance Comparison on Small Datasets

Here, we empirically investigate the effect of the variance reduction discussed in Section 4, and compare the empirical performance of proposed method on small datasets. The class size was fixed to $K = 3$ by merging some classes into one class for each dataset. We sub-sampled labeled data with $n_{\mathrm{L}} = 20$. while the number of unlabeled data depends on the dataset size. Full dataset statistics is given in Appendix H.1. We used a linear-in-input model $f(\boldsymbol{x}) = \boldsymbol{w}^{\top}\boldsymbol{x}$. For training our proposed method, we trained the model for 1000 epochs using Adam with $\alpha = 5.0 \times 10^{-2}$ (full batch size). The candidates of the weight decay parameter were $\{10^{-6}, 10^{-4}, 10^{-2}\}$.

Variance Reduction. Here, before going to the performance comparison, we investigate the effect of variance reduction by using unlabeled data when estimating the training risk. We adopted the randomly

[1] https://www.gagolewski.com/resources/data/ordinal-regression/



(a) `bank1-5`, AT

(b) `bank1-5`, IT

(c) `bank1-5`, LS

(d) `census2-5`, AT

(e) `census2-5`, IT

(f) `census2-5`, LS

Figure 2: The ratio between the variance of the empirical semi-supervised risk to the supervised risk when we adopted AT, IT, and LS as the task surrogate loss and used linear-in-input model. The ratio below 1.0 implies that the variance of the semi-supervised risk is lower than that of the supervised risk.

initialized ordinal regressor g_{rand}, which is the linear-in-input model, to compute the variance of the empirical supervised risk $\mathrm{Var}[\widehat{\mathbb{S}}(g_{\mathrm{rand}})]$ and the empirical semi-supervised risk $\mathrm{Var}[\widehat{\mathbb{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g_{\mathrm{rand}})]$. We selected the removed class k using the strategies described in Section 5.1. We denote SEMI1 as our proposed semi-supervised methods, where the strategy of removing a class is based on the finite sample approximation (30), while SEMI2 denotes the strategy based on the estimation error bound (31). Then, we calculated the ratio between the variance of the empirical semi-supervised risk and supervised risk

$$\mathrm{Var}[\widehat{\mathbb{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g_{\mathrm{rand}})] / \mathrm{Var}[\widehat{\mathbb{S}}(g_{\mathrm{rand}})].$$

Figure 2 shows the ratio between the variance of the empirical semi-supervised risk and supervised risk, which is evaluated on different $\gamma \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$. Additional results of experiments are given in Appendix H.2. It can be observed that the ratios were less than 1 in most cases. This indicates that the variance of our empirical semi-supervised risk is smaller than that of the empirical supervised risk, as suggested by Theorem 12.

Performance Comparison. Tables 3 shows the performances of the proposed framework against the supervised method. We can see that there is no clear difference between the simple strategy (SEMI1) and the strategy based on the estimation error bound (SEMI2). As for the performance comparison, We can see that the proposed methods (SEMI1, SEMI2) perform better than the standard supervised learning (SV). It is worth noting that our framework performs well in all task losses, indicating the broad applicability of our proposed framework.

The Effect of the Number of Unlabeled Data. Here, we investigate the effect of the number of unlabeled data on the performance improvement against the supervised empirical risk minimization approach. We first trained the ordinal regressors with the linear-in-input model for a different number of unlabeled data. Then, we calculated the ratio between the error of the semi-supervised method to the supervised method. Note again that the different evaluation metrics are used for the different task surrogate losses.

Figure 3 shows the ratio between the error of the empirical semi-supervised risk and supervised risk,

Table 3: The mean and standard error of the mean absolute error, mean zero-one error, and mean squared error on small benchmark datasets. The experiments were conducted 20 times. Outperforming methods in each evaluation metric an dataset are highlighted in boldface using the Wilcoxon signed-rank test with a significance level of 5%.

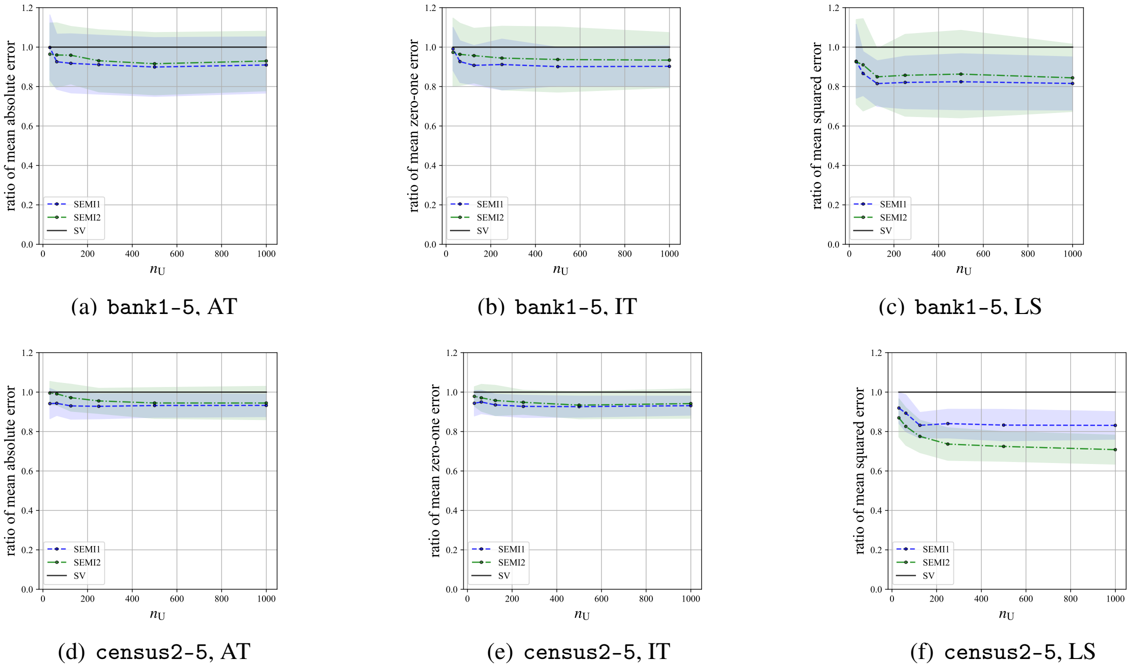
	Mean Absolute Error			Mean Zero-one Error			Mean Squared Error		
	SV	SEMI1	SEMI2	SV	SEMI1	SEMI2	SV	SEMI1	SEMI2
`abalone`	0.381 (0.02)	**0.370 (0.04)**	**0.356 (0.03)**	0.372 (0.02)	**0.368 (0.03)**	**0.353 (0.03)**	0.447 (0.06)	0.392 (0.05)	**0.365 (0.03)**
`bank1-5`	0.312 (0.06)	**0.280 (0.06)**	**0.285 (0.05)**	0.309 (0.06)	**0.277 (0.06)**	**0.283 (0.05)**	0.353 (0.09)	**0.281 (0.05)**	**0.290 (0.06)**
`bank2-5`	**0.518 (0.04)**	**0.518 (0.04)**	0.616 (0.04)	0.484 (0.03)	**0.475 (0.03)**	0.552 (0.03)	1.110 (0.08)	**0.701 (0.07)**	**0.710 (0.09)**
`census1-5`	0.453 (0.07)	**0.414 (0.04)**	**0.416 (0.04)**	0.425 (0.06)	**0.400 (0.03)**	**0.397 (0.03)**	0.544 (0.07)	0.468 (0.04)	**0.431 (0.03)**
`census2-5`	0.534 (0.07)	**0.496 (0.06)**	**0.500 (0.05)**	0.488 (0.06)	**0.453 (0.05)**	**0.457 (0.04)**	0.755 (0.11)	0.627 (0.10)	**0.530 (0.06)**
`computer1-5`	0.308 (0.05)	**0.295 (0.05)**	**0.302 (0.04)**	**0.299 (0.05)**	**0.291 (0.04)**	**0.300 (0.04)**	0.461 (0.10)	0.403 (0.06)	**0.329 (0.06)**
`computer2-5`	0.312 (0.04)	**0.286 (0.03)**	0.317 (0.04)	0.306 (0.04)	**0.283 (0.03)**	0.310 (0.04)	0.511 (0.08)	0.432 (0.06)	**0.356 (0.05)**
`fireman`	**0.556 (0.08)**	**0.556 (0.08)**	**0.554 (0.07)**	**0.471 (0.06)**	**0.464 (0.06)**	**0.461 (0.04)**	0.637 (0.12)	0.626 (0.11)	**0.591 (0.08)**
`kinematics`	0.635 (0.06)	**0.606 (0.05)**	**0.608 (0.06)**	**0.536 (0.03)**	**0.526 (0.04)**	**0.528 (0.03)**	0.779 (0.10)	**0.702 (0.07)**	**0.743 (0.12)**
`lev`	0.339 (0.07)	**0.308 (0.04)**	**0.301 (0.04)**	0.335 (0.06)	**0.308 (0.03)**	**0.299 (0.04)**	0.354 (0.07)	**0.345 (0.13)**	**0.304 (0.05)**
`swd`	0.701 (0.05)	0.700 (0.12)	**0.620 (0.05)**	0.579 (0.03)	0.566 (0.03)	**0.537 (0.04)**	0.873 (0.12)	0.765 (0.10)	**0.683 (0.08)**
`toy`	0.266 (0.07)	**0.215 (0.05)**	**0.226 (0.05)**	0.235 (0.05)	**0.212 (0.04)**	**0.215 (0.04)**	0.361 (0.10)	**0.250 (0.08)**	0.293 (0.08)

which is evaluated on the number of unlabeled data. Additional results of experiments are given in Appendix H.2. It can be observed that SEMI1 and SEMI2 outperform the supervised method. However, it is also observed that the performance gain is saturated at a certain point depending on the dataset, i.e., increasing more unlabeled data may no longer improve the performance of both SEMI1 and SEMI2. Note that this is also a common phenomenon observed in the literature of semi-supervised learning (Sakai et al., 2017; Oliver et al., 2018; Sakai et al., 2018). Nevertheless, it is a challenging future work to consider a way to improve the way to utilize unlabeled data, for example, by developing a new regularization technique for the unbiased risk estimator approach in the context of semi-supervised learning.

6.2 Performance Comparison on Large Datasets

Here, we compare the empirical performance of proposed framework on large datasets. We sub-sampled labeled data with $n_{\mathrm{L}} = 500$, and $n_{\mathrm{U}} = 2000$, and the number of test size was fixed to 2000. The statistics of datasets used in the experiments is given in Appendix H. We used linear-in-parameter models with Gaussian kernel $f(\boldsymbol{x}) = \sum_{i=1}^{n_{\mathrm{L}}} w_i \exp(-\left\|\boldsymbol{x} - \boldsymbol{x}_i\right\|^2 / \left(2\sigma^2\right))$, and neural networks with one hidden layer (of size 256) and ReLU as an activation function. For the Gaussian kernel, the bandwidth candidates were $\{0.5, 1.0, 2.0\} \cdot \mathrm{median}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{i,j=1}^{n_{\mathrm{L}}})$ where $\{\boldsymbol{x}_i\}_{i=1}^{n_{\mathrm{L}}} = \mathfrak{X}_{\mathrm{L}}$. For training our proposed method, we trained the model for 1000 epochs using Adam with $\alpha = 5.0 \times 10^{-3}$ (full batch size). The weight decay parameter was fixed to 10^{-4} for both models.

Performance Comparison. Tables 4 to 6 show the performances of the proposed methods against the baseline methods. We can see that the proposed methods (SEMI1-NN and SEMI2-NN) perform best than the baseline methods on most benchmark datasets. Specifically, we can observe that the method based on the manifold assumption (SSMOR) performs poorly while the performance method based on the cluster assumption (TOR) is somewhat comparable with the proposed methods. This is because the datasets used in the experiments do not satisfy the manifold assumption as the SSMOR is original proposed in the context of image ranking, while TOR does not heavily rely on the geometric assumption. As well as the case for the small datasets, our framework performs well in all task losses, indicating the broad applicability of our proposed framework. However, unlike neural networks, when a linear-in-parameter model with Gaussian kernel is used (SEMI1-Ker and SEMI2-Ker), we found that our method failed to improve the performance over the supervised method. This could be due to the risk modification in (35) might not be suitable for a kernel model since it was originally designed for using with neural networks (Lu et al., 2020). We hypothesize that the poor performance of the kernel method when using (35) can be attributed to the fact that the modification makes the risk formulation to be no longer convex, which causes difficulty in optimization. Note that the supervised baseline does not use the risk modification in (35). Hence, the supervised baseline has a convex formulation and therefore is easier to train. We note that in



(a) bank1-5, AT

(b) bank1-5, IT

(c) bank1-5, LS

(d) census2-5, AT

(e) census2-5, IT

(f) census2-5, LS

Figure 3: The ratio between the error of the empirical semi-supervised risk to the supervised risk when we adopted AT, IT, and LS as the task surrogate loss and used a linear-in-input model. The ratio below 1.0 implies that the error of the semi-supervised method is better than that of the supervised risk.

Table 4: The mean and standard error of the mean absolute error on benchmark datasets. Outperforming methods in each dataset are highlighted in boldface using the Wilcoxon signed-rank test with a significance level of 5%. The experiments were conducted 20 times.

	SSMOR	TOR	SV-Ker	SEMI1-Ker	SEMI2-Ker	SV-NN	SEMI1-NN	SEMI2-NN
bank1-5	0.935 (0.16)	**0.250 (0.01)**	**0.249 (0.01)**	0.294 (0.05)	0.339 (0.06)	0.267 (0.01)	0.251 (0.01)	**0.248 (0.01)**
bank2-5	1.184 (0.07)	0.907 (0.03)	0.883 (0.02)	0.887 (0.03)	0.915 (0.03)	0.834 (0.02)	0.839 (0.02)	**0.828 (0.02)**
census1-5	1.190 (0.11)	0.803 (0.03)	0.785 (0.03)	0.792 (0.02)	0.803 (0.02)	0.785 (0.03)	**0.686 (0.02)**	0.697 (0.02)
census2-5	1.206 (0.14)	0.784 (0.02)	0.768 (0.02)	0.779 (0.02)	0.798 (0.03)	0.745 (0.02)	**0.693 (0.02)**	**0.689 (0.02)**
computer1-5	0.671 (0.05)	0.444 (0.02)	0.441 (0.02)	0.445 (0.02)	0.459 (0.02)	0.474 (0.03)	**0.413 (0.02)**	**0.409 (0.02)**
computer2-5	0.668 (0.07)	0.379 (0.01)	0.376 (0.01)	0.387 (0.01)	0.397 (0.02)	0.457 (0.01)	**0.367 (0.01)**	**0.370 (0.02)**
fireman	2.999 (0.44)	3.690 (0.07)	3.649 (0.06)	3.732 (0.13)	3.798 (0.18)	**1.298 (0.05)**	1.602 (0.08)	1.593 (0.09)
kinematics	1.797 (0.13)	1.442 (0.02)	1.454 (0.02)	1.474 (0.02)	1.474 (0.02)	**1.155 (0.03)**	1.175 (0.04)	1.175 (0.04)

the case of neural networks, the optimization problem is non-convex regardless of whether we use a risk modification or not. Thus, the modification does not affect the convexity of the problem To improve the performance of the kernel model with our proposed method, it is an important future work to explore a risk modification that can retain the convexity of the formulation.

Table 5: The mean and standard error of the zero-one error on benchmark datasets. Outperforming methods in each dataset are highlighted in boldface using the Wilcoxon signed-rank test with a significance level of 5%. The experiments were conducted 20 times.

	SSMOR	TOR	SV-Ker	SEMI1-Ker	SEMI2-Ker	SV-NN	SEMI1-NN	SEMI2-NN
`bank1-5`	0.608 (0.04)	**0.238 (0.01)**	**0.238 (0.01)**	0.244 (0.01)	0.246 (0.01)	0.273 (0.01)	0.255 (0.01)	0.259 (0.01)
`bank2-5`	0.706 (0.02)	0.646 (0.01)	0.651 (0.02)	0.641 (0.01)	0.647 (0.01)	**0.602 (0.01)**	**0.601 (0.01)**	0.605 (0.01)
`census1-5`	0.697 (0.02)	0.569 (0.01)	0.579 (0.01)	0.584 (0.02)	0.585 (0.01)	0.574 (0.01)	**0.556 (0.01)**	**0.560 (0.01)**
`census2-5`	0.689 (0.03)	0.568 (0.01)	0.572 (0.01)	0.579 (0.01)	0.576 (0.01)	0.549 (0.01)	0.541 (0.01)	**0.538 (0.01)**
`computer1-5`	0.544 (0.04)	0.386 (0.01)	0.386 (0.01)	0.386 (0.01)	0.386 (0.01)	0.416 (0.02)	**0.376 (0.02)**	0.383 (0.01)
`computer2-5`	0.537 (0.04)	**0.339 (0.01)**	**0.339 (0.01)**	**0.341 (0.01)**	0.344 (0.02)	0.398 (0.01)	**0.343 (0.01)**	0.357 (0.02)
`fireman`	0.863 (0.02)	0.927 (0.01)	0.926 (0.00)	0.927 (0.01)	0.926 (0.01)	**0.682 (0.01)**	0.718 (0.01)	0.718 (0.02)
`kinematics`	0.785 (0.01)	0.768 (0.01)	0.779 (0.01)	0.781 (0.01)	0.781 (0.01)	**0.705 (0.01)**	0.711 (0.01)	0.711 (0.01)

Table 6: The mean and the standard error of the mean squared error on benchmark datasets. Outperforming methods in each dataset are highlighted in boldface using the Wilcoxon signed-rank test with a significance level of 5%. The experiments were conducted 20 times.

	SSMOR	TOR	SV-Ker	SEMI1-Ker	SEMI2-Ker	SV-NN	SEMI1-NN	SEMI2-NN
`bank1-5`	1.719 (0.47)	0.364 (0.02)	0.363 (0.02)	0.415 (0.08)	0.432 (0.06)	0.317 (0.02)	0.313 (0.02)	**0.306 (0.02)**
`bank2-5`	2.448 (0.23)	1.282 (0.05)	**1.256 (0.04)**	1.595 (0.07)	1.380 (0.08)	1.464 (0.07)	1.481 (0.07)	1.476 (0.07)
`census1-5`	2.524 (0.44)	1.233 (0.08)	1.183 (0.06)	1.188 (0.07)	1.222 (0.06)	1.264 (0.08)	**1.053 (0.07)**	**1.068 (0.06)**
`census2-5`	2.493 (0.51)	1.205 (0.07)	1.172 (0.06)	1.140 (0.04)	1.198 (0.06)	1.189 (0.06)	1.100 (0.05)	**1.079 (0.04)**
`computer1-5`	0.958 (0.09)	0.550 (0.03)	0.538 (0.03)	0.542 (0.03)	0.569 (0.05)	0.618 (0.05)	0.536 (0.03)	**0.515 (0.03)**
`computer2-5`	0.944 (0.16)	0.457 (0.02)	**0.453 (0.02)**	**0.456 (0.02)**	0.484 (0.02)	0.604 (0.05)	0.502 (0.05)	0.480 (0.03)
`fireman`	15.753 (4.34)	6.472 (0.24)	6.519 (0.24)	7.553 (1.32)	7.226 (1.01)	**2.766 (0.19)**	3.319 (0.33)	3.168 (0.28)
`kinematics`	5.633 (0.74)	2.878 (0.14)	2.751 (0.08)	2.994 (0.14)	2.994 (0.14)	2.597 (0.11)	**2.362 (0.13)**	**2.362 (0.13)**

7 Conclusions

We presented a novel framework to incorporate unlabeled data in ordinal regression based on empirical risk minimization. We proposed an unbiased risk estimator that is applicable to all well-known task losses, such as the absolute loss, squared loss, and zero-one loss. We also elucidated the property of the proposed unbiased risk estimator through the analysis of the estimation error bound to guarantee that the proposed risk estimator is consistent. Experimental results showed that our proposed framework could effectively make use of unlabeled data, resulting in better scores compared to the standard supervised learning method in three task losses in terms of the mean absolute error, mean zero-one error, and mean squared error. In the future work, we plan to consider a new regularization technique for the unbiased risk estimator approach in the context of semi-supervised learning so that the method can much more utilize the information of unlabeled data.

Acknowledgements

The authors would like to thank Han Bao, Yusuke Konno, and Tomoya Sakai for helpful discussions, and Kento Nozawa, Ikko Yamane, and Soma Yokoi for maintaining servers for our experiments. TT was supported by JST AIP Challenge Program, UTokyo Toyota-Dwango AI Scholarship, and Softbank AI Scholarship. NC was supported by MEXT Scholarship, JST AIP Challenge Program, and Google PhD Fellowship Program. IS was supported by JST CREST Grant Number JPMJCR17A1, and MS was supported by JST CREST JPMJCR18A2.

References

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 452–461, 2018.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Shai Ben-David, Nadav Eiron, and Philip M Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

Ralf Bender and Ulrich Grouven. Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London*, 31(5):546–551, 1997.

Ralf Bender and Ulrich Grouven. Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, 51(10):809–816, 1998.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

Wei Chu and S Sathiya Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 145–152, 2005.

Harmen Dikkers and Leon Rothkrantz. Support vector machines in ordinal classification: An application to corporate credit scoring. *Neural Network World*, 15(6):491, 2005.

Marthinus du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1386–1394, 2015.

Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

Andrew S Fullerton and Jun Xu. The proportional odds with partial proportionality constraints model for ordinal response variables. *Social Science Research*, 41(1):182–198, 2012.

Pedro Antonio Gutierrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016.

Liu Hua-fu. Towards semi-supervised ordinal regression with nearest neighbor. *Journal of Computer Applications*, 2010.

Kyoung-jae Kim and Hyunchul Ahn. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research*, 39(8): 1800–1811, 2012.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems 30*, pages 1675–1685, 2017.

Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2015.

Yang Liu, Yan Liu, Shenghua Zhong, and Keith CC Chan. Semi-supervised manifold ordinal regression for image ranking. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1393–1396, 2011.

Nan Lu, Tianyi Zhang, Gang Niu, and Masashi Sugiyama. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 1115–1125, 2020.

Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems 29*, pages 1199–1207, 2016.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems 31*, volume 31, pages 3235–3246, 2018.

Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 708–717, 2016.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18:1 – 35, 2017.

María Pérez-Ortiz, Pedro Antonio Gutiérrez, Mariano Carbonero-Ruz, and César Hervás-Martínez. Semi-supervised learning for ordinal kernel discriminant analysis. *Neural Networks*, 84:57–66, 2016.

Jason D. M. Rennie. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, pages 180–186, 2005.

Tomoya Sakai, Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2998–3006, 2017.

Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Semi-supervised AUC optimization based on positive-unlabeled learning. *Machine Learning*, 107(4):767–794, 2018.

Chun-Wei Seah, Ivor W Tsang, and Yew-Soon Ong. Transductive ordinal regression. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1074–1086, 2012.

Matthias Seeger. Learning with labeled and unlabeled data. 2000.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

PK Srijith, Shirish Shevade, and S Sundararajan. Semi-supervised Gaussian process ordinal regression. In *Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 144–159, 2013.

Seiya Tokui, Ryosuke Okuta, Takuya Akiba, Yusuke Niitani, Toru Ogawa, Shunta Saito, Shuji Suzuki, Kota Uenishi, Brian Vogel, and Hiroyuki Yamazaki Vincent. Chainer: A deep learning framework for accelerating the research cycle. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2002–2011, 2019.

Gordon C Weir and Susan Bonner-Weir. Five stages of evolving beta-cell dysfunction during progression to diabetes. *Diabetes*, 53(suppl 3):S16–S21, 2004.

A Proof of Lemma 3

Proof. We can rewrite the task surrogate risk as follows:

$$\begin{aligned}\mathcal{S}(g) &= \mathbb{E}_{X,Y}\left[\psi(\boldsymbol{\alpha}(X), Y)\right] \\ &= \sum_{y=1}^{K} \pi_y \mathbb{E}_{X|Y=y}\left[\psi(\boldsymbol{\alpha}(X), y)\right] \\ &= \sum_{y\in\mathcal{Y}\setminus k} \pi_y \mathbb{E}_{X|Y=y}\left[\psi(\boldsymbol{\alpha}(X), y)\right] + \pi_k \mathbb{E}_{X|Y=k}\left[\psi(\boldsymbol{\alpha}(X), k)\right]. \qquad (36)\end{aligned}$$

By expanding the marginal distribution, we have

$$\mathbb{E}_{\mathrm{U}}\left[\psi(\boldsymbol{\alpha}(X), k)\right] = \sum_{y=1}^{K} \pi_y \mathbb{E}_{X|Y=y}\left[\psi(\boldsymbol{\alpha}(X), k)\right], \qquad (37)$$

Then, we can express $\pi_k \mathbb{E}_{X|Y=k}\left[\psi(\boldsymbol{\alpha}(X), k)\right]$ in terms of the expectation of unlabeled data and the class-conditional expectations of all classes except class k as

$$\pi_k \mathbb{E}_{X|Y=k}\left[\psi(\boldsymbol{\alpha}(X), k)\right] = \mathbb{E}_{\mathrm{U}}\left[\psi(\boldsymbol{\alpha}(X), k)\right] - \sum_{y\in\mathcal{Y}\setminus k} \pi_y \mathbb{E}_{X|Y=y}\left[\psi(\boldsymbol{\alpha}(X), k)\right]. \qquad (38)$$

Replacing the last term of the right-hand side of (36) with the right-hand side of (38), we complete the proof. □

B Proof of Proposition 5

Proof. From Theorem 4, we have $\mathcal{S}^{\setminus k}_{\text{SEMI-}\gamma}(g) = \mathcal{S}(g)$. Also it is known that $\mathcal{S}(g)$ is Fisher-consistent to $\mathcal{R}(g)$, if we adopt the all threshold, cumulative link, least absolute deviation, immediate threshold, or least squares as the task surrogate loss ψ (Pedregosa et al., 2017). By combining all of the results mentioned above, the proof is completed. □

C Technical Lemmas

The following lemma is used when deriving the estimation error bounds in Appendix D.

Lemma 15 (McDiarmid's inequality (Theorem D.8 in Mohri et al. (2018)))**.** *Let $X_1, \ldots, X_n$ be a set of independent random variables taking values in $\mathcal{X}$ and assume that there exists $c_1, \ldots, c_n > 0$ such that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies*

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \le c_i$$

for all $i \in \{1, \ldots, n\}$ and any $x_1, \ldots, x_n, x_i' \in \mathcal{X}$. Then,

$$\begin{aligned}\mathbb{P}\left\{f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \ge \epsilon\right\} &\le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i}\right) \\ \mathbb{P}\left\{f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \le -\epsilon\right\} &\le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i}\right).\end{aligned}$$

The following lemma is used when bounding the Rademacher complexities of hypothesis classes in Appendix E.

Lemma 16 (Talagrand's lemma (Lemma 5.7 in Mohri et al. (2018)))**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be ρ-Lipschitz function. Then, for any set $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$,*

$$\mathfrak{R}(\Phi \circ \mathcal{H}) \le \rho \mathfrak{R}(\mathcal{H}).$$

D Analysis of Estimation Error Bounds (Proof of Theorem 9)

Before proving Theorem 9, we define and recall notations and give some lemmas. Defining

$$\mathcal{S}_{\mathrm{L}}(g) := \mathcal{S}_{\mathrm{L}}^{(1)}(g) - \mathcal{S}_{\mathrm{L}}^{(2)}(g) = \sum_{y \in \mathcal{Y} \setminus k} \pi_y \underbrace{\mathbb{E}_{X|Y=y}\left(\psi(\boldsymbol{\alpha}(X), y) - \psi(\boldsymbol{\alpha}(X), k)\right)}_{=:\mathcal{S}_{\mathrm{L},y}(g)},$$

we have

$$\mathcal{S}_{\mathrm{LU}}^{\setminus k}(g) = \mathcal{S}_{\mathrm{L}}(g) + \mathcal{S}_{\mathrm{U}}(g),$$

where we recall that

$$\mathcal{S}_{\mathrm{U}}(g) = \mathbb{E}_{\mathrm{U}}\left[\psi(\boldsymbol{\alpha}(X), k)\right].$$

In a similar manner, we define the empirical version of each risk. Defining

$$\widehat{\mathcal{S}}_{\mathrm{L}}(g) := \widehat{\mathcal{S}}_{\mathrm{L}}^{(1)}(g) - \widehat{\mathcal{S}}_{\mathrm{L}}^{(2)}(g) = \sum_{y \in \mathcal{Y} \setminus k} \pi_y \underbrace{\frac{1}{n_y} \sum_{j=1}^{n_y} (\psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), y) - \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), k))}_{=:\widehat{\mathcal{S}}_{\mathrm{L},y}(g)},$$

we have

$$\widehat{\mathcal{S}}_{\mathrm{LU}}^{\setminus k}(g) = \widehat{\mathcal{S}}_{\mathrm{L}}(g) + \widehat{\mathcal{S}}_{\mathrm{U}}(g),$$

where we recall that

$$\widehat{\mathcal{S}}_{\mathrm{U}}(g) = \frac{1}{n_{\mathrm{U}}} \sum_{j=1}^{n_{\mathrm{U}}} \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^{\mathrm{U}}), k).$$

First, we prove three lemmas in the following.

Lemma 17. *Assume that there exists a constant* $C_\psi > 0$ *such that* $\psi(y, \boldsymbol{\alpha}) \leq C_\psi$ *for any* $y \in \mathbb{R}$, $\boldsymbol{\alpha} \in \mathbb{R}^{K-1}$. *Then, for any* $\delta > 0$, *with probability at least* $1 - \frac{K-1}{K+1}\delta$,

$$\sup_{g \in \mathcal{G}} \left| \mathcal{S}_{\mathrm{L}}(g) - \widehat{\mathcal{S}}_{\mathrm{L}}(g) \right| \\ \leq 2 \sum_{y \in \mathcal{Y} \setminus k} \pi_y \left(\mathfrak{R}(\mathcal{H}^{(y)}; n_y) + \mathfrak{R}(\mathcal{H}^{(k)}; n_y) \right) + 2\sqrt{2}(K-1) C_\psi \sqrt{\ln \frac{2(K+1)}{\delta}} \sum_{y \in \mathcal{Y} \setminus k} \frac{\pi_y}{\sqrt{n_y}}. \tag{39}$$

Proof. Recalling that

$$\widehat{\mathcal{S}}_{\mathrm{L},y}(g) = \frac{1}{n_y} \sum_{j=1}^{n_y} (\psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), y) - \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), k)),$$

we have

$$\begin{aligned} &\left| \sup_{g \in \mathcal{G}} \left(\mathcal{S}_{\mathrm{L},y}(g) - \widehat{\mathcal{S}}_{\mathrm{L},y}(g) \right) - \sup_{g \in \mathcal{G}} \left(\mathcal{S}_{\mathrm{L},y}(g) - \widehat{\mathcal{S}}_{\mathrm{L},y}^m(g) \right) \right| \\ &\leq \sup_{g \in \mathcal{G}} \left| \widehat{\mathcal{S}}_{\mathrm{L},y}^m(g) - \widehat{\mathcal{S}}_{\mathrm{L},y}(g) \right| \\ &\leq \sup_{g \in \mathcal{G}} \left| \frac{1}{n_y} \left[(\psi(\boldsymbol{\alpha}(\boldsymbol{x}_m^y), y) - \psi(\boldsymbol{\alpha}(\boldsymbol{x}_m^y), k)) - \left(\psi(\boldsymbol{\alpha}(\boldsymbol{x}_{m'}^y), y) - \psi(\boldsymbol{\alpha}(\boldsymbol{x}_{m'}^y), k) \right) \right] \right| \\ &\leq \frac{4C_\psi}{n_y}, \end{aligned} \tag{40}$$

where $\widehat{\mathcal{S}}^m_{\mathrm{L},y}(g)$ is the slight modification of $\widehat{\mathcal{S}}_{\mathrm{L},y}(g)$ where $\boldsymbol{x}^y_m$ is changed to $\boldsymbol{x}^y_{m'}$. Thus, by McDiarmid's inequality (Lemma 15 in Appendix C), we have

$$\mathbb{P}\left\{\sup_g(\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g))-\mathbb{E}_{S_y|Y=y}\left[\sup_g(\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g))\right]\geq\epsilon\right\}\leq\exp\left(-\frac{\epsilon^2 n_y}{8C_\psi^2}\right), \tag{41}$$

where $S_y|Y=y$ denotes the distribution over $\mathcal{X}^{n_y}$ when conditioned as $Y=y$. Therefore, with probability at least $1-\frac{\delta}{2(K+1)}$, it holds that

$$\sup_g(\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g))\leq\mathbb{E}_{S_y|Y=y}\left[\sup_g(\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g))\right]+2\sqrt{2}C_\psi\sqrt{\ln\frac{2(K+1)}{\delta}}\frac{1}{\sqrt{n_y}}. \tag{42}$$

The first term of the right-hand side of the above inequality can be bounded as

$$\mathbb{E}_{S_y|Y=y}\left[\sup_g(\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g))\right] \tag{43}$$

$$=\mathbb{E}_{S_y|Y=y}\left[\sup_g\left(\mathbb{E}_{S'_y|Y=y}\left[\widehat{\mathcal{S}}'_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g)\right]\right)\right]$$

$$\leq\mathbb{E}_{S_y,S'_y|Y=y}\left[\sup_g\left(\widehat{\mathcal{S}}'_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g)\right)\right] \tag{44}$$

$$=\mathbb{E}_{S_y,S'_y|Y=y}\left[\sup_g\frac{1}{n_y}\sum_{j=1}^{n_y}\left((\psi(\boldsymbol{\alpha}(\boldsymbol{x}'^y_j),y)-\psi(\boldsymbol{\alpha}(\boldsymbol{x}'^y_j),k))-(\psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j),y)-\psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j),k))\right)\right]$$

$$=\mathbb{E}_{S_y,S'_y|Y=y}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_g\frac{1}{n_y}\sum_{j=1}^{n_y}\sigma_j\left((\psi(\boldsymbol{\alpha}(\boldsymbol{x}'^y_j),y)-\psi(\boldsymbol{\alpha}(\boldsymbol{x}'^y_j),k))-(\psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j),y)-\psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j),k))\right)\right] \tag{45}$$

$$\leq\mathbb{E}_{S'_y|Y=y}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_g\frac{1}{n_y}\sum_{j=1}^{n_y}\sigma_j\psi(\boldsymbol{\alpha}(\boldsymbol{x}'^y_j),y)\right]+\mathbb{E}_{S'_y|Y=y}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_g\frac{1}{n_y}\sum_{j=1}^{n_y}(-\sigma_j)\psi(\boldsymbol{\alpha}(\boldsymbol{x}'^y_j),k)\right]$$
$$+\mathbb{E}_{S_y|Y=y}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_g\frac{1}{n_y}\sum_{j=1}^{n_y}(-\sigma_j)\psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j),y)\right]+\mathbb{E}_{S_y|Y=y}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_g\frac{1}{n_y}\sum_{j=1}^{n_y}\sigma_j\psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j),k)\right] \tag{46}$$

$$\leq 2(\mathfrak{R}(\mathcal{H}^{(y)};n_y)+\mathfrak{R}(\mathcal{H}^{(k)};n_y)), \tag{47}$$

where (44) and (46) follow since the subadditivity of the supremum, (45) follows from the fact that the distribution $S_y|Y=y$ and $S'_y|Y=y$ are the same, and in the last inequality we the fact that $-\sigma_j$ and σ_j follow the same distribution. Repeating the same argument and together with (42), de Morgan's Laws and the union bound, wit probability at least $1-\frac{\delta}{K+1}$, we have

$$\sup_g\left|\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g)\right|\leq 2(\mathfrak{R}(\mathcal{H}^{(y)};n_y)+\mathfrak{R}(\mathcal{H}^{(k)};n_y))+2\sqrt{2}C_\psi\sqrt{\ln\frac{2(K+1)}{\delta}}\frac{1}{\sqrt{n_y}}. \tag{48}$$

Therefore, Lemma 17 holds as a consequence of the inequality

$$\sup_{g\in\mathcal{G}}\left|\mathcal{S}_{\mathrm{L}}(g)-\widehat{\mathcal{S}}_{\mathrm{L}}(g)\right|=\sup_g\left|\sum_{y\in\mathcal{Y}\backslash k}\pi_y(\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g))\right|\leq\sum_{y\in\mathcal{Y}\backslash k}\pi_y\sup_g\left|\mathcal{S}_{\mathrm{L},y}(g)-\widehat{\mathcal{S}}_{\mathrm{L},y}(g)\right|,$$

de Morgan's Laws, and the union bound. □

Lemma 18. *Assume that there exists a constant $C_\psi > 0$ such that $\psi(y, \boldsymbol{\alpha}) \leq C_\psi$ for any $y \in \mathcal{Y}$, $\boldsymbol{\alpha} \in \mathbb{R}^{K-1}$. Then, for any $\delta > 0$, with probability at least $1 - \frac{\delta}{K+1}$,*

$$\sup_{g \in \mathcal{G}} \left| \mathcal{S}_{\mathrm{U}}(g) - \widehat{\mathcal{S}}_{\mathrm{U}}(g) \right| \leq 2\mathfrak{R}(\mathcal{H}^{(k)}; n_{\mathrm{U}}) + \sqrt{2} C_\psi \sqrt{\ln \frac{2(K+1)}{\delta}} \frac{1}{\sqrt{n_{\mathrm{U}}}}.$$

Proof. This lemma can be proven similarly to Lemma 17. □

Lemma 19. *Assume that there exists a constant $C_\psi > 0$ such that $\psi(y, \boldsymbol{\alpha}) \leq C_\psi$ for any $y \in \mathcal{Y}$, $\boldsymbol{\alpha} \in \mathbb{R}^{K-1}$. Then, for any $\delta > 0$, with probability at least $1 - \frac{\delta}{K+1}$,*

$$\sup_{g \in \mathcal{G}} \left| \mathcal{S}(g) - \widehat{\mathcal{S}}(g) \right| \leq 2\mathfrak{R}(\mathcal{H}; n_{\mathrm{L}}) + \sqrt{2} C_\psi \sqrt{\ln \frac{2(K+1)}{\delta}} \frac{1}{\sqrt{n_{\mathrm{L}}}}.$$

Proof. This lemma can be proven similarly to Lemma 17. □

Proof of Theorem 9. Let us bound the estimation error $\mathcal{S}(\widehat{g}^{\backslash k}) - \mathcal{S}(g^*)$. Note that the equality $\mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(g) = \mathcal{S}(g)$ holds for any $g \in \mathcal{G}$. Then, we have

$$\begin{aligned}
\mathcal{S}(\widehat{g}^{\backslash k}) - \mathcal{S}(g^*) &= \mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(\widehat{g}^{\backslash k}) - \mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(g^*) \\
&= \left(\mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(\widehat{g}^{\backslash k}) - \widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(\widehat{g}^{\backslash k}) \right) + \left(\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(\widehat{g}^{\backslash k}) - \widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g^*)) \right) \\
&\quad + \left(\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g^*) - \mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(g^*) \right) \\
&\leq \left(\mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(\widehat{g}^{\backslash k}) - \widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(\widehat{g}^{\backslash k}) \right) + 0 + \left(\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g^*) - \mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(g^*) \right) \qquad (49) \\
&\leq 2 \sup_{g \in \mathcal{G}} \left| \mathcal{S}^{\backslash k}_{\text{SEMI-}\gamma}(g) - \widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g) \right| \\
&\leq 2\gamma \left(\sup_{g \in \mathcal{G}} \left| \mathcal{S}_{\mathrm{L}}(g) - \widehat{\mathcal{S}}_{\mathrm{L}}(g) \right| + \sup_{g \in \mathcal{G}} \left| \mathcal{S}_{\mathrm{U}}(g) - \widehat{\mathcal{S}}_{\mathrm{U}}(g) \right| \right) \\
&\quad + 2(1-\gamma) \sup_{g \in \mathcal{G}} \left| \mathcal{S}(g) - \widehat{\mathcal{S}}(g) \right|, \qquad (50)
\end{aligned}$$

where (49) follows since $\widehat{g}^{\backslash k}$ is the minimizer of $\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}$, and the last inequality follows from the sub-additivity of the supremum. Hence, combining Eq. (50) and Lemmas 17 to 19, with probability at least $1 - \delta$, we have

$$\begin{aligned}
&\mathcal{S}(\widehat{g}^{\backslash k}) - \mathcal{S}(g^*) \\
&\leq 4 \left(\gamma \left(\sum_{y \in \mathcal{Y}^{\backslash k}} \pi_y \left(\mathfrak{R}(\mathcal{H}^{(y)}; n_y) + \mathfrak{R}(\mathcal{H}^{(k)}; n_y) \right) + \mathfrak{R}(\mathcal{H}^{(k)}; n_{\mathrm{U}}) \right) + (1-\gamma) \mathfrak{R}(\mathcal{H}; n_{\mathrm{L}}) \right) \\
&\quad + 2\sqrt{2} C_\psi \sqrt{\ln \frac{2(K+1)}{\delta}} \left(\gamma \left(2 \sum_{y \in \mathcal{Y}^{\backslash k}} \frac{\pi_y}{\sqrt{n_y}} + \frac{1}{\sqrt{n_{\mathrm{U}}}} \right) + (1-\gamma) \frac{1}{\sqrt{n_{\mathrm{L}}}} \right). \qquad (51)
\end{aligned}$$

Combining the above inequality with Theorems 7 and 8, we completed the proof of Theorem 9. □

E Upper Bounds for Rademacher Complexities in Ordinal Regression (Proof of Lemma 10)

Before proving Lemma 10, we prepare the following lemma and two theorems. In this section, we write $[n] \coloneqq \{1, \ldots, n\}$ for $n \in \mathbb{N}$. Recall that

$$\begin{aligned}
\mathcal{H}_{\mathrm{AT}} &= \{(\boldsymbol{x}, y) \mapsto \psi_{\mathrm{AT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\}, \\
\mathcal{H}_{\mathrm{IT}} &= \{(\boldsymbol{x}, y) \mapsto \psi_{\mathrm{IT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\}, \\
\mathcal{H}_{\mathrm{LS}} &= \{(\boldsymbol{x}, y) \mapsto \psi_{\mathrm{LS}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}\}.
\end{aligned} \tag{52}$$

and

$$\begin{aligned}
\mathcal{H}^{(y)}_{\mathrm{AT}} &= \{\boldsymbol{x} \mapsto \psi_{\mathrm{AT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\}, \\
\mathcal{H}^{(y)}_{\mathrm{IT}} &= \{\boldsymbol{x} \mapsto \psi_{\mathrm{IT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta\}, \\
\mathcal{H}^{(y)}_{\mathrm{LS}} &= \{\boldsymbol{x} \mapsto \psi_{\mathrm{LS}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) : f \in \mathcal{F}\}.
\end{aligned} \tag{53}$$

Lemma 20. *Let $n \in \mathbb{N}$ and $\mathcal{Z} \subset \mathcal{Y}$, and assume that*

$$\mathfrak{R}(\mathcal{H}^{(y)}; n) = 0 \tag{54}$$

for $y \in \mathcal{Z}$. Then,

$$\mathfrak{R}(\mathcal{H}; n) \leq \sum_{y \in \mathcal{Z}} \mathfrak{R}(\mathcal{H}^{(y)}; n). \tag{55}$$

Proof. Using the identity $\mathbb{1}[y = y_i] = \frac{1}{2} + \frac{2\,\mathbb{1}[y=y_i]-1}{2}$, we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}; n) &= \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\boldsymbol{x}_i, y_i)\right] \\
&= \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sum_{y \in \mathcal{Y}} \sigma_i h(\boldsymbol{x}_i, y) \mathbb{1}[y = y_i]\right] \\
&\leq \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sum_{y \in \mathcal{Y}} \sigma_i h(\boldsymbol{x}_i, y) \mathbb{1}[y = y_i]\right] && (56) \\
&= \sum_{y \in \mathcal{Y}} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\boldsymbol{x}_i, y) \left(\frac{1}{2} + \frac{2\,\mathbb{1}[y = y_i] - 1}{2}\right)\right] \\
&\leq \frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\boldsymbol{x}_i, y)\right] + \frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sigma_i (2\,\mathbb{1}[y = y_i] - 1) h(\boldsymbol{x}_i, y)\right] && (57) \\
&= \frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\boldsymbol{x}_i, y)\right] + \frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\boldsymbol{x}_i, y)\right] && (58) \\
&= \sum_{y \in \mathcal{Y}} \mathfrak{R}(\mathcal{H}^{(y)}; n) \\
&= \sum_{y \in \mathcal{Z}} \mathfrak{R}(\mathcal{H}^{(y)}; n), && (59)
\end{aligned}$$

where (56) and (57) follow from the subadditivity of the supremum, (58) follows from the fact that $\sigma_i(2\,\mathbb{1}[y = y_i] - 1)$ and σ_i follow the same distribution, and the last inequality follows from the assumption. □

Next, we give the upper bounds for the Rademacher complexities with $\mathcal{H}^{(y)}_{\mathrm{AT}}$, $\mathcal{H}^{(y)}_{\mathrm{IT}}$, and $\mathcal{H}^{(y)}_{\mathrm{LS}}$ without any assumptions on the decision functions class $\mathcal{F}$. Let us recall the theorem for the convenience.

Theorem 7. *Fix $y \in \mathcal{Y}$. Let $n \in \mathbb{N}$ assume for the AT and IT losses that the binary surrogate loss ℓ is ρ-Lipschitz. Then, the expected Rademacher complexities of $\mathcal{H}^{(y)}_{\mathrm{AT}}$, $\mathcal{H}^{(y)}_{\mathrm{IT}}$, and $\mathcal{H}^{(y)}_{\mathrm{LS}}$ are bounded as*

$$\begin{aligned}
&\mathfrak{R}(\mathcal{H}^{(y)}_{\mathrm{AT}}; n) \le \rho \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n), \\
&\mathfrak{R}(\mathcal{H}^{(y)}_{\mathrm{IT}}; n) \le \rho \left(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)\right), \\
&\mathfrak{R}(\mathcal{H}^{(y)}_{\mathrm{LS}}; n) \le 2|y + \theta_1 - 3/2|\mathfrak{R}(\mathcal{F}; n) + \mathfrak{R}(\mathrm{sq} \circ \mathcal{F}; n),
\end{aligned} \tag{60}$$

where $\mathrm{sq} : \mathrm{Im}\, f \ni z \mapsto z^2 \in \mathbb{R}$.

Proof. We prove the theorem for each loss.
AT loss. We have

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}^{(y)}_{\mathrm{AT}}; n) &= \mathbb{E}\left[\sup_{h \in \mathcal{H}^{(y)}_{\mathrm{AT}}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\boldsymbol{x}_i, y)\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(\sum_{j=1}^{y-1} \ell(-\alpha_j(\boldsymbol{x}_i)) + \sum_{j=y}^{K-1} \ell(\alpha_j(\boldsymbol{x}_i))\right)\right] \\
&\le \mathbb{E}\left[\sup_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(\sum_{j=1}^{y-1} \ell(-\alpha_j(\boldsymbol{x}_i))\right)\right] + \mathbb{E}\left[\sup_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(\sum_{j=y}^{K-1} \ell(\alpha_j(\boldsymbol{x}_i))\right)\right], \quad (61) \\
&\le \sum_{j=1}^{y-1} \mathbb{E}\left[\sup_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(\ell(-\alpha_j(\boldsymbol{x}_i))\right)\right] + \sum_{j=y}^{K-1} \mathbb{E}\left[\sup_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(\ell(\alpha_j(\boldsymbol{x}_i))\right)\right], \quad (62) \\
&\le \sum_{j=1}^{K-1} \mathfrak{R}(\ell \circ \mathcal{A}_j; n) \quad (63) \\
&\le \rho \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n), \quad (64)
\end{aligned}$$

where (61) and (62) follow from the subadditivity of the supremum, (63) follow from the fact that $-\sigma_i$ and σ_i follows the same distribution, and the last inequality follows from the Talagrand's lemma (Lemma 16 in Appendix C).
IT loss. The proof for the case of the IT loss follows a very similar argument to that of the AT loss, and

we include the proof for completeness. We have

$$
\begin{aligned}
\mathfrak{R}(\mathcal{H}_{\mathrm{IT}}^{(y)}; n) &= \mathbb{E}\left[\sup_{h \in \mathcal{H}_{\mathrm{IT}}^{(y)}} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(\ell(-\alpha_{y-1}(\boldsymbol{x}_i)) + \ell(\alpha_y(\boldsymbol{x}_i))\right)\right] \\
&\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i \in [n]} \sigma_i \ell(-\alpha_{y-1}(\boldsymbol{x}_i))\right] + \mathbb{E}\left[\sup_{f \in \mathcal{F}, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i \in [n]} \sigma_i \ell(\alpha_y(\boldsymbol{x}_i))\right], \qquad (65) \\
&\leq \mathfrak{R}(\ell \circ (-\mathcal{A}_{y-1}); n) + \mathfrak{R}(\ell \circ \mathcal{A}_y; n) \qquad (66) \\
&\leq \rho(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)), \qquad (67)
\end{aligned}
$$

where (65) and (66) follows from the subadditivity of the supremum, the last inequality follows from the fact that $-\sigma_i$ and σ_i follows the same distribution and the Talagrand's lemma.

LS loss. We give the upper bound for the case of the LS loss. We have

$$
\begin{aligned}
\mathfrak{R}(\mathcal{H}_{\mathrm{LS}}^{(y)}; n) &= \mathbb{E}\left[\sup_{h \in \mathcal{H}_{\mathrm{LS}}^{(y)}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\alpha(\boldsymbol{x}_i), y)\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(y - f(\boldsymbol{x}_i) + \theta_1 - \frac{3}{2}\right)^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{n} \sum_{i \in [n]} \sigma_i \left(y + \theta_1 - \frac{3}{2}\right)\right] + \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} \sigma_i f(\boldsymbol{x}_i)^2\right] \\
&\quad + 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} (-\sigma_i) f(\boldsymbol{x}_i)(y + \theta_1 - 3/2)\right]. \qquad (68)
\end{aligned}
$$

The first term in (68) is 0. The third term in (68) can be rewritten as

$$
\begin{aligned}
&2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} (-\sigma_i) f(\boldsymbol{x}_i)(y + \theta_1 - 3/2)\right] \\
&= 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} (-\sigma_i) \operatorname{sign}(y + \theta_1 - 3/2) |y + \theta_1 - 3/2| f(\boldsymbol{x}_i)\right] \\
&= 2|y + \theta_1 - 3/2| \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} (-\sigma_i) \operatorname{sign}(y + \theta_1 - 3/2) f(\boldsymbol{x}_i)\right] \\
&= 2|y + \theta_1 - 3/2| \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} \sigma_i f(\boldsymbol{x}_i)\right] \qquad (69) \\
&= 2|y + \theta_1 - 3/2| \ \mathfrak{R}(\mathcal{F}; n), \qquad (70)
\end{aligned}
$$

where in (69) we used the fact that $-\sigma_i \operatorname{sign}(y + \theta_1 - 3/2)$ and σ_i follow the same distribution. Summing up the above argument, the proof is completed. □

Next, we give the upper bounds for the Rademacher complexities with $\mathcal{H}_{\mathrm{AT}}$, $\mathcal{H}_{\mathrm{IT}}$, and $\mathcal{H}_{\mathrm{LS}}$ without any assumptions on a class of decision functions $\mathcal{F}$ using Lemma 20 and Theorem 7 Let us recall the theorem for the convenience.

Theorem 8 (Upper bounds of Rademacher complexities with general decision functions class)**.** *Let* $n \in \mathbb{N}$ *and assume for the AT and IT losses that the binary surrogate loss* ℓ *is* ρ*-Lipschitz. Then, the expected Rademacher complexities of* $\mathcal{H}_{\mathrm{AT}}, \mathcal{H}_{\mathrm{IT}}$*, and* $\mathcal{H}_{\mathrm{LS}}$ *are bounded as*

$$\begin{aligned}\mathfrak{R}(\mathcal{H}_{\mathrm{AT}}; n) &\le \rho K \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n),\\ \mathfrak{R}(\mathcal{H}_{\mathrm{IT}}; n) &\le \rho \sum_{y \in \mathcal{Y}} \left(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)\right),\\ \mathfrak{R}(\mathcal{H}_{\mathrm{LS}}; n) &\le 2(K + |\theta_1 - 3/2|)\mathfrak{R}(\mathcal{F}; n) + \mathfrak{R}(\mathrm{sq} \circ \mathcal{F}; n),\end{aligned} \tag{71}$$

where $\mathrm{sq} : \mathrm{Im}\, f \ni z \mapsto z^2 \in \mathbb{R}$.

Proof of Theorem 8. For the AT and IT losses, the result of Theorem 8 directly follows from Lemma 20 and Theorem 7.
AT loss. For the case of the AT loss, we have

$$\mathfrak{R}(\mathcal{H}_{\mathrm{AT}}; n) \le \sum_{y \in \mathcal{Y}} \mathfrak{R}(\mathcal{H}_{\mathrm{AT}}^{(y)}; n) \le \rho \sum_{y \in \mathcal{Y}} \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n) = \rho K \sum_{j=1}^{K-1} \mathfrak{R}(\mathcal{A}_j; n), \tag{72}$$

where in the first inequality we used Lemma 20, and in the last inequality we used Theorem 7.
IT loss. In a similar manner, we can bound for the case of IT loss as

$$\mathfrak{R}(\mathcal{H}_{\mathrm{IT}}; n) \le \sum_{y \in \mathcal{Y}} \mathfrak{R}(\mathcal{H}_{\mathrm{IT}}^{(y)}; n) \le \rho \sum_{y \in \mathcal{Y}} \left(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)\right). \tag{73}$$

LS loss. Next, we give the upper bound for the case of the LS loss, where we will not rely on Lemma 20 and Theorem 7. We have

$$\begin{aligned}\mathfrak{R}(\mathcal{H}_{\mathrm{LS}}; n) &= \mathbb{E}\left[\sup_{h \in \mathcal{H}_{\mathrm{LS}}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\alpha(\boldsymbol{x}_i), y_i)\right]\\ &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} \sigma_i \left(y_i - f(\boldsymbol{x}_i) + \theta_1 - \frac{3}{2}\right)^2\right]\\ &\le \mathbb{E}\left[\frac{1}{n} \sum_{i \in [n]} \sigma_i \left(y_i^2 + 2y_i(\theta_1 - 3/2) + (\theta_1 - 3/2)^2\right)\right] + 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} (-\sigma_i) y_i f(\boldsymbol{x}_i)\right]\\ &\quad + \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} \sigma_i f(\boldsymbol{x}_i)^2\right] + 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} (-\sigma_i) f(\boldsymbol{x}_i)(\theta_1 - 3/2)\right].\end{aligned} \tag{74}$$

The first term in (74) is 0. The second term in (74) is bounded as

$$2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} (-\sigma_i) y_i f(\boldsymbol{x}_i)\right] \le 2K\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} \sigma_i f(\boldsymbol{x}_i)\right] = 2K\mathfrak{R}(\mathcal{F}; n), \tag{75}$$

since σ_i and $-\sigma_i$ follow the same distribution. The forth term in (74) can be rewritten as

$$\begin{aligned}
2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i\in[n]}(-\sigma_i)f(\boldsymbol{x}_i)(\theta_1-3/2)\right] &= 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i\in[n]}(-\sigma_i)\operatorname{sign}(\theta_1-3/2)|\theta_1-3/2|f(\boldsymbol{x}_i)\right] \\
&= 2|\theta_1-3/2|\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i\in[n]}(-\sigma_i)\operatorname{sign}(\theta_1-3/2)f(\boldsymbol{x}_i)\right] \\
&= 2|\theta_1-3/2|\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i\in[n]}\sigma_i f(\boldsymbol{x}_i)\right] \qquad (76)\\
&= 2|\theta_1-3/2|\,\mathfrak{R}(\mathcal{F};n), \qquad (77)
\end{aligned}$$

where in (76) we used the fact that $-\sigma_i\operatorname{sign}(\theta_1-3/2)$ and σ_i follow the same distribution. Summing up the above argument, the proof is completed. □

Remark. The upper bound for the case of the LS loss can be similarly obtained as the cases of the AT and IT losses. However, it loosen the bound by the factor of K compared to the analysis given above.

The next lemma is the well-known bound for the Rademacher complexity of the linear-in-parameter models.

Lemma 21 (Theorem 5.10 in Mohri et al. (2018))**.** *Let $\mathcal{F}$ be the class of linear-in-parameter models, that is,*

$$\mathcal{F}=\{f(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x})\colon\|\boldsymbol{w}\|_2\le C_{\boldsymbol{w}},\ \|\boldsymbol{\phi}(\boldsymbol{x})\|_2\le C_{\boldsymbol{\phi}}\} \tag{78}$$

for positive constants $C_{\boldsymbol{w}}$ and $C_{\boldsymbol{\phi}}$. Then,

$$\mathfrak{R}(\mathcal{F};n)\le\frac{C_{\boldsymbol{w}}C_{\boldsymbol{\phi}}}{\sqrt{n}}. \tag{79}$$

Finally, we prove the Lemma 10 in the following.

Proof of Lemma 10. The result of Lemma 10 directly follows from Theorems 7, 8 and Lemma 21 as follows.

AT and IT losses. Fix $j\in[K]$. Being aware that we have

$$\alpha_j(\boldsymbol{x})=\theta_j-w^\top\phi(\boldsymbol{x})=(\theta_j,\boldsymbol{w}^\top)\begin{pmatrix}1\\-\boldsymbol{x}\end{pmatrix}, \tag{80}$$

and

$$\|(\theta_j,\boldsymbol{w}^\top)\|_2\le\|\boldsymbol{\theta}\|_2+\|\boldsymbol{w}\|_2\le C_{\boldsymbol{\theta}}+C_{\boldsymbol{w}}, \tag{81}$$

we can use Lemma 21 and have

$$\mathfrak{R}(\mathcal{A}_j;n)\le\frac{(C_{\boldsymbol{\theta}}+C_{\boldsymbol{w}})\,C_{\boldsymbol{\phi}}}{\sqrt{n}}. \tag{82}$$

In a similar manner, using Theorem 8, we have

$$\mathfrak{R}(\mathcal{H}_{\mathrm{AT}}^{(y)};n)\le\rho\,K\mathfrak{R}(\mathcal{A}_j;n)\le\frac{(C_{\boldsymbol{\theta}}+C_{\boldsymbol{w}})\,C_{\boldsymbol{\phi}}\rho K}{\sqrt{n}} \tag{83}$$

and

$$\mathfrak{R}(\mathcal{H}^{(y)}_{\mathrm{IT}}; n) \leq \rho(\mathfrak{R}(\mathcal{A}_{y-1}; n) + \mathfrak{R}(\mathcal{A}_y; n)) \leq \frac{2(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}} \rho}{\sqrt{n}}, \tag{84}$$

which completes the proof for the cases of the AT and IT losses.
LS loss. Since $|f(\boldsymbol{x})| \leq \|w\|_2 \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \leq C_{\boldsymbol{w}} C_{\boldsymbol{\phi}}$ by the Cauchy–Schwarz inequality, the map $\mathrm{sq} : \operatorname{Im} f \to \mathbb{R}$ is $2C_{\boldsymbol{w}} C_{\boldsymbol{\phi}}$-Lipschitz. Therefore, using Theorem 8 and the Talagrand's lemma, we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}^{(y)}_{\mathrm{LS}}; n) &\leq 2|y + \theta_1 - 3/2| \mathfrak{R}(\mathcal{F}; n) + \mathfrak{R}(\mathrm{sq} \circ \mathcal{F}; n) \\
&\leq 2|y + \theta_1 - 3/2| \mathfrak{R}(\mathcal{F}; n) + 2C_{\boldsymbol{w}} C_{\boldsymbol{\phi}} \mathfrak{R}(\mathcal{F}; n) \\
&\leq \frac{2(|y + \theta_1 - 3/2| + C_{\boldsymbol{w}} C_{\boldsymbol{\phi}})(C_{\boldsymbol{\theta}} + C_{\boldsymbol{w}})\, C_{\boldsymbol{\phi}} K}{\sqrt{n}},
\end{aligned} \tag{85}$$

where in the last inequality we used Lemma 21.

Using similar manners as the above arguments, we can obtain the upper bounds for the cases with $\mathcal{H}_{\mathrm{AT}}$, $\mathcal{H}_{\mathrm{IT}}$, and $\mathcal{H}_{\mathrm{LS}}$. □

F Proof of Theorem 12

Proof. Take any $g \in \mathcal{G}$. Recalling that $\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g) = \gamma \widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}} + (1-\gamma)\widehat{\mathcal{S}}$, we can rewrite the variance of the empirical semi-supervised risk $\operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g)]$ as

$$\operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g)] = \gamma^2 \operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g)] + (1-\gamma)^2 \operatorname{Var}[\widehat{\mathcal{S}}(g)] + 2\gamma(1-\gamma) \operatorname{Cov}(\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g), \widehat{\mathcal{S}}(g)). \tag{86}$$

Hence, the condition $\operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g)] < \operatorname{Var}[\widehat{\mathcal{S}}(g)]$ is equivalent to

$$\begin{aligned}
&(\operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g)] + \operatorname{Var}[\widehat{\mathcal{S}}(g)] - 2\operatorname{Cov}(\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g), \widehat{\mathcal{S}}(g)))\gamma^2 \\
&\qquad - 2(\operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g)] - \operatorname{Cov}(\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g), \widehat{\mathcal{S}}(g)))\gamma < 0.
\end{aligned} \tag{87}$$

This condition is indeed satisfied by taking γ satisfying

$$0 < \gamma < \frac{2\left(\operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g)] - \operatorname{Cov}(\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g), \widehat{\mathcal{S}}(g))\right)}{\operatorname{Var}[\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g)] + \operatorname{Var}[\widehat{\mathcal{S}}(g)] - 2\operatorname{Cov}(\widehat{\mathcal{S}}^{\backslash k}_{\mathrm{LU}}(g), \widehat{\mathcal{S}}(g))}, \tag{88}$$

and the proof is completed. □

G Proof of Theorem 14

Proof. If $\widehat{\mathcal{S}}_{\mathrm{L}}(g)$ is shown to be convex, the objective function $\widehat{J}_{\ell}(\boldsymbol{w}, \boldsymbol{\theta}) = \widehat{\mathcal{S}}^{\backslash k}_{\text{SEMI-}\gamma}(g) + \Omega(\boldsymbol{\theta})$ is convex with respect to $\boldsymbol{w}$ and $\boldsymbol{\theta}$ since the other terms are convex.
AT loss. Recalling that the AT loss is defined as

$$\psi_{\mathrm{AT}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) = \sum_{i=1}^{y-1} \ell(-\alpha_i(\boldsymbol{x})) + \sum_{i=y}^{K-1} \ell(\alpha_i(\boldsymbol{x})), \tag{89}$$

we have

$$
\begin{aligned}
\widehat{\mathbb{S}}_{\mathrm{L}}(g) &= \sum_{y\in\mathcal{Y}\setminus k} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \Big(\psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), y) - \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), k)\Big) \\
&= \sum_{y\in\mathcal{Y}\setminus k} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \left(\sum_{i=1}^{y-1} \ell(-\alpha_i(\boldsymbol{x}_j^y)) + \sum_{i=y}^{K-1} \ell(\alpha_i(\boldsymbol{x}_j^y)) - \sum_{i=1}^{k-1} \ell(-\alpha_i(\boldsymbol{x}_j^y)) - \sum_{i=k}^{K-1} \ell(\alpha_i(\boldsymbol{x}_j^y)) \right).
\end{aligned}
\tag{90}
$$

Then, the term inside the sum can be rewritten as

$$
\begin{aligned}
&\sum_{i=1}^{y-1} \ell(-\alpha_i(\boldsymbol{x}_j^y)) + \sum_{i=y}^{K-1} \ell(\alpha_i(\boldsymbol{x}_j^y)) - \sum_{i=1}^{k-1} \ell(-\alpha_i(\boldsymbol{x}_j^y)) - \sum_{i=k}^{K-1} \ell(\alpha_i(\boldsymbol{x}_j^y)) \\
&= \begin{cases} \sum_{i=y}^{k-1} \Big(-\ell(-\alpha_i(\boldsymbol{x}_j^y)) + \ell(\alpha_i(\boldsymbol{x}_j^y))\Big) & (y < k) \\ -\sum_{i=k}^{y-1} \Big(-\ell(-\alpha_i(\boldsymbol{x}_j^y)) + \ell(\alpha_i(\boldsymbol{x}_j^y))\Big) & (y > k) \\ 0 & (y = k) \end{cases} \\
&= \begin{cases} -C_\ell \sum_{i=y}^{k-1} \alpha_i(\boldsymbol{x}_j^y) & (y < k) \\ C_\ell \sum_{i=k}^{y-1} \alpha_i(\boldsymbol{x}_j^y) & (y > k) \\ 0 & (y = k). \end{cases}
\end{aligned}
\tag{91}
$$

Therefore, $\widehat{\mathbb{S}}_{\mathrm{L}}(g)$ is convex and the objective function $\widehat{J}_\ell(\boldsymbol{w}, \boldsymbol{\theta})$ is convex.
LS loss. Recalling that the LS loss is defined as

$$
\psi_{\mathrm{LS}}(\boldsymbol{\alpha}(\boldsymbol{x}), y) = \left(y + \alpha_1(\boldsymbol{x}) - \frac{3}{2}\right)^2,
$$

we have

$$
\begin{aligned}
\widehat{\mathbb{S}}_{\mathrm{L}}(g) &= \sum_{y\in\mathcal{Y}\setminus k} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \Big(\psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), y) - \psi(\boldsymbol{\alpha}(\boldsymbol{x}_j^y), k)\Big) \\
&= \sum_{y\in\mathcal{Y}\setminus k} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \left\{ \left(y + \alpha_1(\boldsymbol{x}_j^y) - \frac{3}{2}\right)^2 - \left(k + \alpha_1 - \frac{3}{2}\right)^2 \right\} \\
&= \sum_{y\in\mathcal{Y}\setminus k} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \left\{ y^2 + k^2 + 2(y + k) \left(\alpha_1(\boldsymbol{x}_j^y) - \frac{3}{2}\right) \right\}.
\end{aligned}
\tag{92}
$$

This is convex and subsequently the objective function $\widehat{J}_\ell(\boldsymbol{w}, \boldsymbol{\theta})$ is convex, and the proof is completed. □

H Dataset Statistics and Additional Results of Experiments

H.1 Benchmark Dataset Statistics

The detailed dataset statistics used in the small and large scale experiments are given in Tables 7 and 8. The other statistics is given in Section 6.

Table 7: Dataset statistics of small scale experiments.

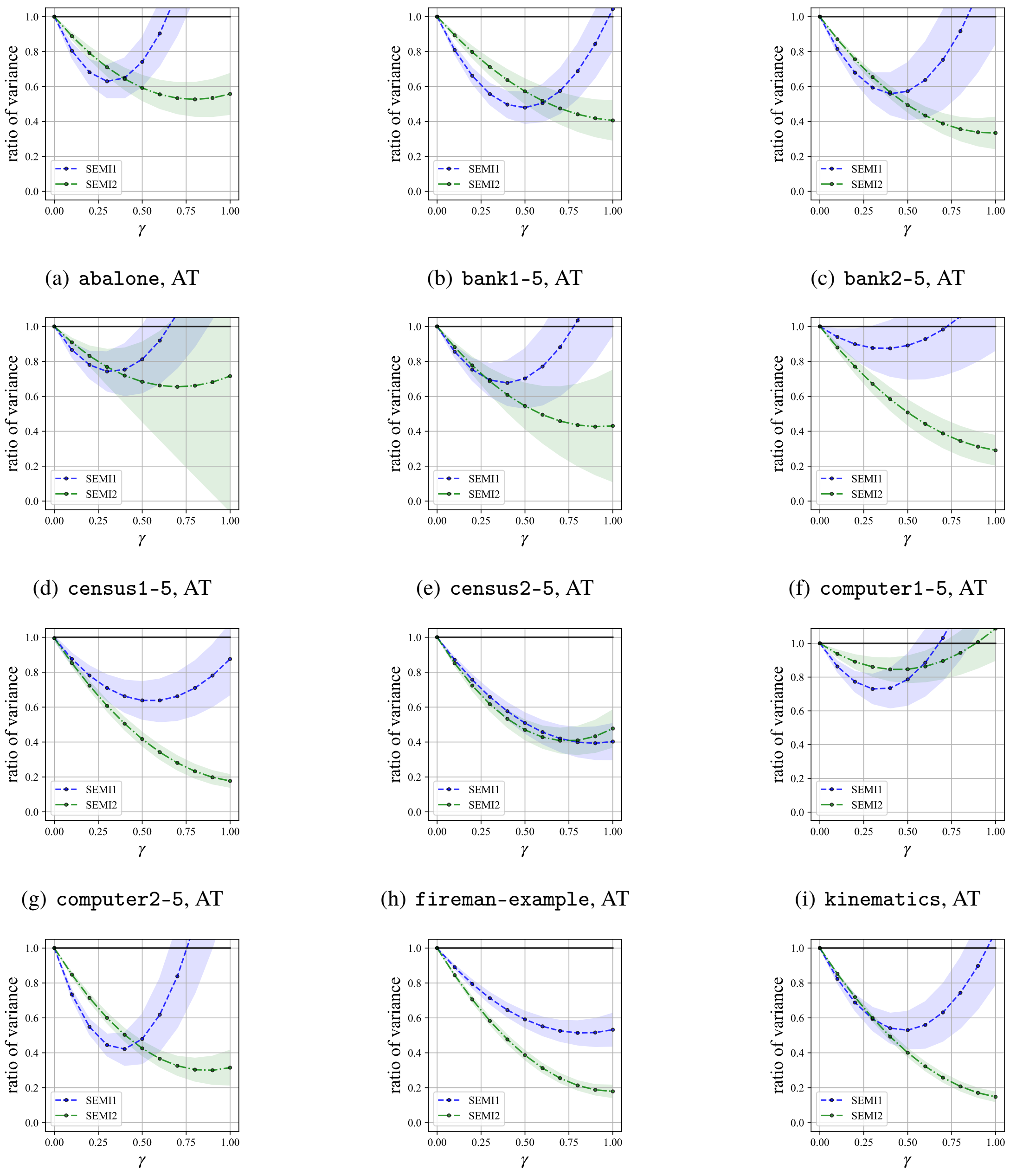
	dim.	# of unlabeled data	# of test data	class prior
abalone	10	871	1253	[0.2, 0.6, 0.2]
bank1-5	8	1851	2000	[0.2, 0.6, 0.2]
bank2-5	32	1851	2000	[0.2, 0.6, 0.2]
census1-5	8	2000	2000	[0.2, 0.6, 0.2]
census2-5	16	2000	2000	[0.2, 0.6, 0.2]
computer1-5	12	1851	2000	[0.2, 0.6, 0.2]
computer2-5	21	1851	2000	[0.2, 0.6, 0.2]
fireman-example	9	2000	2000	[0.5, 0.25, 0.25]
kinematics	7	1851	2000	[0.38, 0.38, 0.25]
lev	4	204	300	[0.09, 0.68, 0.22]
swd	10	204	300	[0.38, 0.4, 0.22]
toy	2	57	90	[0.12, 0.78, 0.1]

Table 8: Dataset statistics of large scale experiments.

	# of class	class prior
bank1-5	5	[0.2, 0.2, 0.2, 0.2, 0.2]
bank2-5	5	[0.2, 0.2, 0.2, 0.2, 0.2]
census1-5	5	[0.2, 0.2, 0.2, 0.2, 0.2]
census2-5	5	[0.2, 0.2, 0.2, 0.2, 0.2]
computer1-5	5	[0.2, 0.2, 0.2, 0.2, 0.2]
computer2-5	5	[0.2, 0.2, 0.2, 0.2, 0.2]
fireman-example	16	around 0.0625 for all classes
kinematics	8	around 0.125 for all classes

H.2 Additional Results of Experiments

Here we show the extended experimental results for the variance reduction experiments discussed in Section 6. Figures 4 to 6 show the results of the variance comparison. We can see that the variances of the semi-supervised risks are smaller than that of the supervised risk. Figures 7 to 9 show that the effect of the number of unlabeled on the performance improvement. Note that the experiments on toy dataset is not conducted because its number of unlabeled data is too small. We can see that the performance of the proposed method is indeed improved, but the improvement by adding a large amount of data is limited except when using LS as the task surrogate loss.



(a) `abalone`, AT

(b) `bank1-5`, AT

(c) `bank2-5`, AT

(d) `census1-5`, AT

(e) `census2-5`, AT

(f) `computer1-5`, AT

(g) `computer2-5`, AT

(h) `fireman-example`, AT

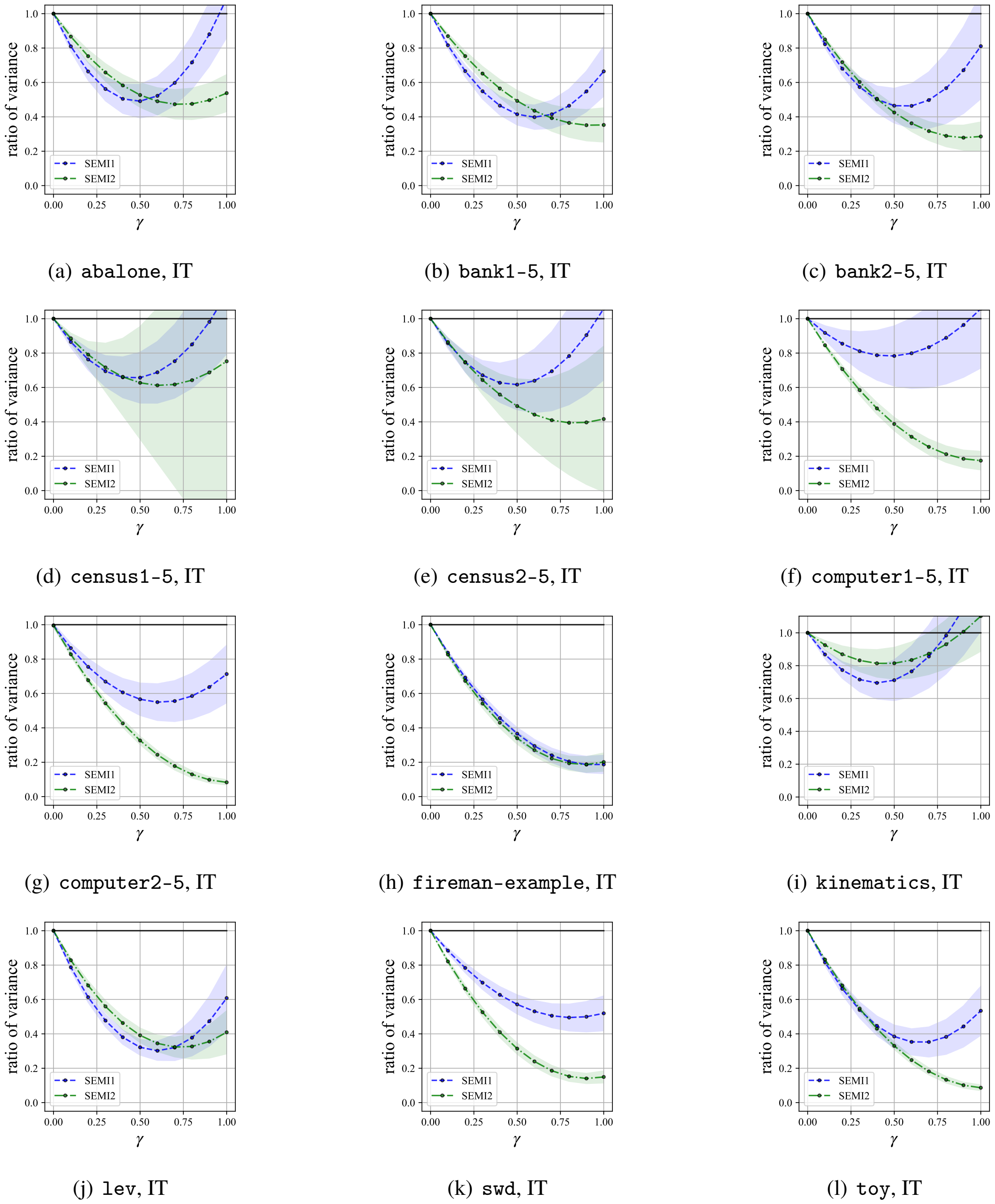
(i) `kinematics`, AT

(j) `lev`, AT

(k) `swd`, AT

(l) `toy`, AT

Figure 4: The ratio between the variance of the empirical semi-supervised risk to the supervised risk when we adopted AT as the task surrogate loss and used a linear-in-input model.



(a) `abalone`, IT

(b) `bank1-5`, IT

(c) `bank2-5`, IT

(d) `census1-5`, IT

(e) `census2-5`, IT

(f) `computer1-5`, IT

(g) `computer2-5`, IT

(h) `fireman-example`, IT

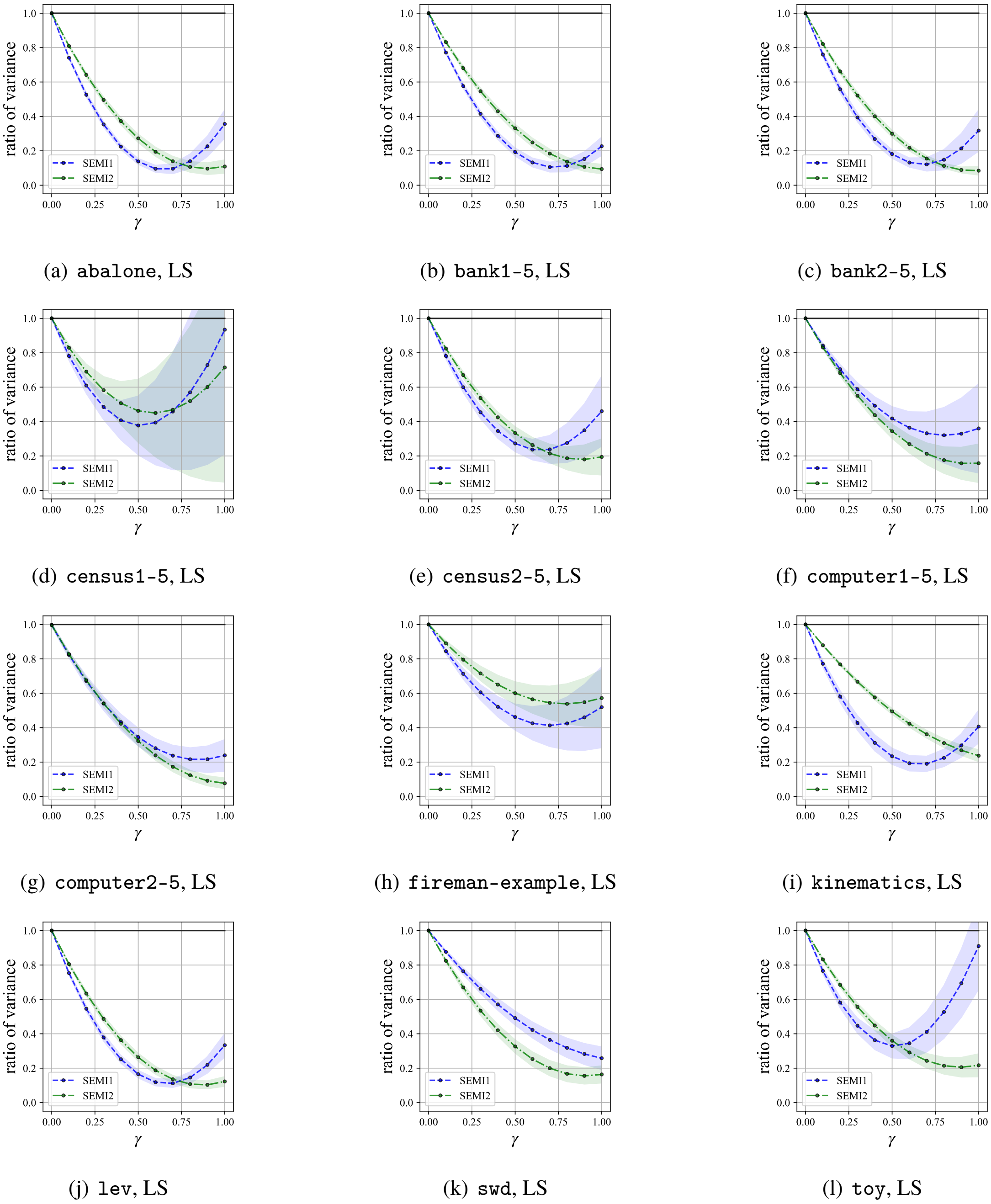
(i) `kinematics`, IT

(j) `lev`, IT

(k) `swd`, IT

(l) `toy`, IT

Figure 5: The ratio between the variance of the empirical semi-supervised risk to the supervised risk when we adopted IT as the task surrogate loss and used a linear-in-input model.



(a) `abalone`, LS

(b) `bank1-5`, LS

(c) `bank2-5`, LS

(d) `census1-5`, LS

(e) `census2-5`, LS

(f) `computer1-5`, LS

(g) `computer2-5`, LS

(h) `fireman-example`, LS

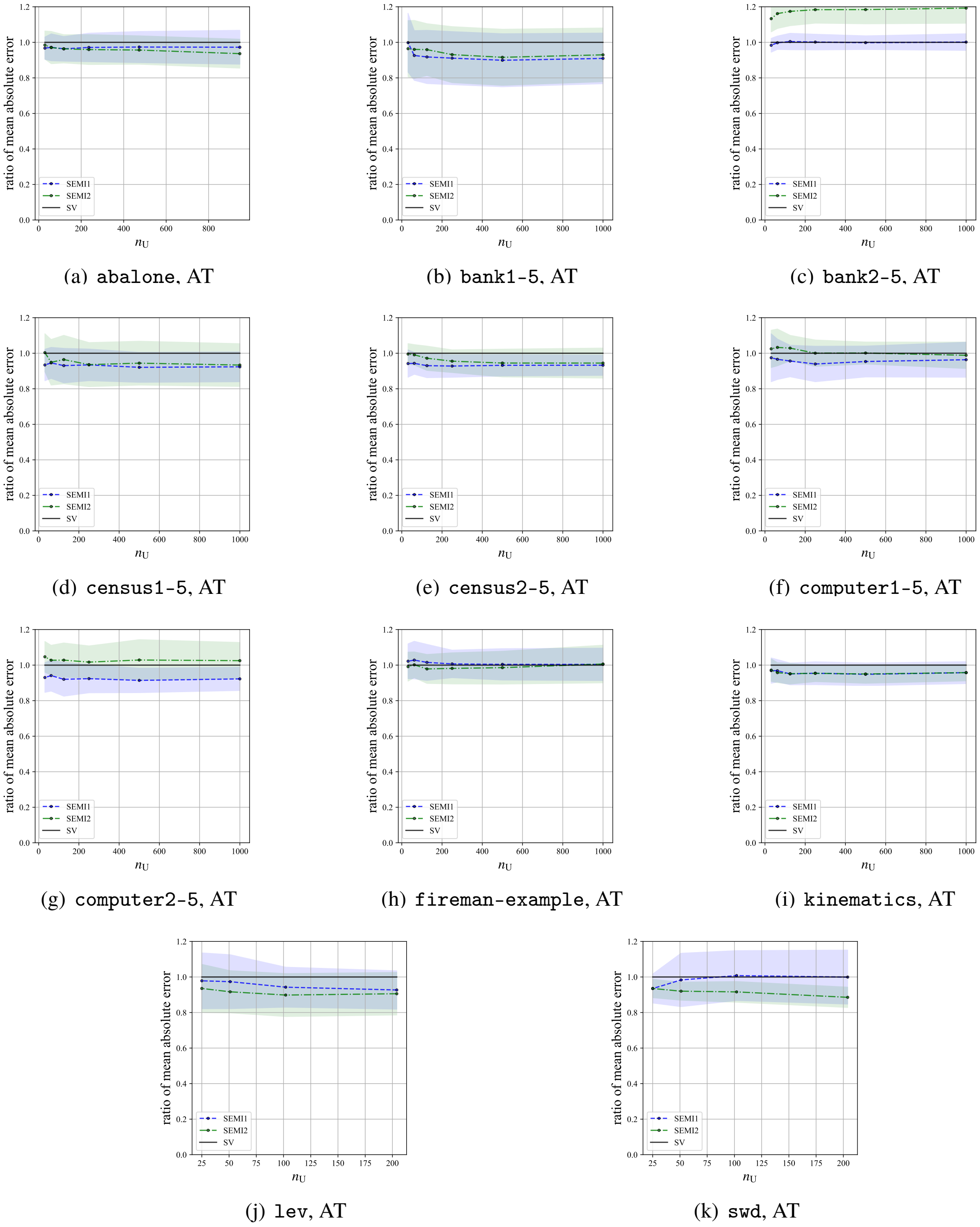
(i) `kinematics`, LS

(j) `lev`, LS

(k) `swd`, LS

(l) `toy`, LS

Figure 6: The ratio between the variance of the empirical semi-supervised risk to the supervised risk when we adopted LS as the task surrogate loss and used a linear-in-input model.



(a) `abalone`, AT

(b) `bank1-5`, AT

(c) `bank2-5`, AT

(d) `census1-5`, AT

(e) `census2-5`, AT

(f) `computer1-5`, AT

(g) `computer2-5`, AT

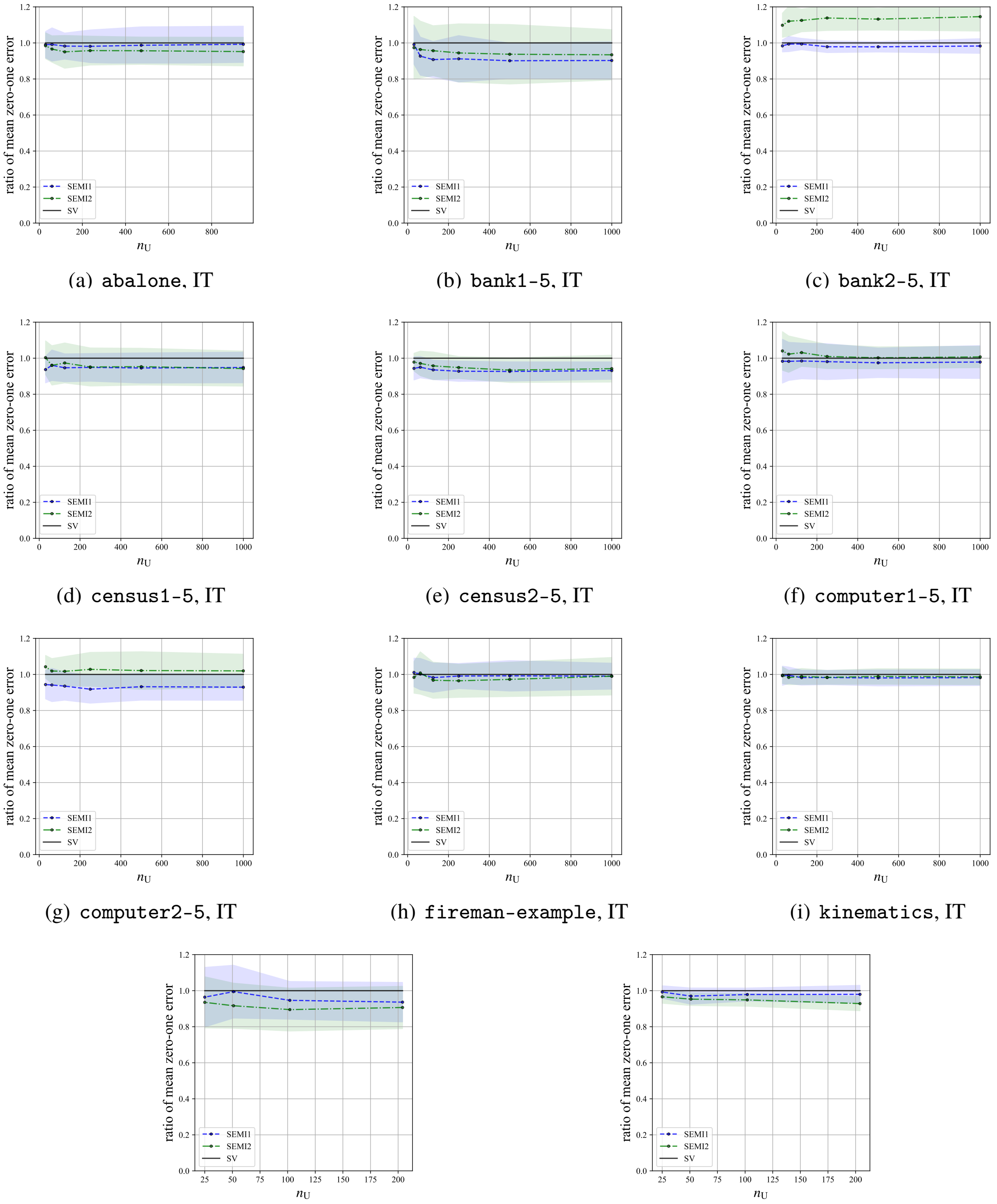
(h) `fireman-example`, AT

(i) `kinematics`, AT

(j) `lev`, AT

(k) `swd`, AT

Figure 7: The ratio between the error of the empirical semi-supervised risk to the supervised risk when we adopted AT as the task surrogate loss and used linear-in-input as a model. The ratio below 1.0 implies that the error of the semi-supervised method is better than that of the supervised risk.



(a) abalone, IT

(b) bank1-5, IT

(c) bank2-5, IT

(d) census1-5, IT

(e) census2-5, IT

(f) computer1-5, IT

(g) computer2-5, IT

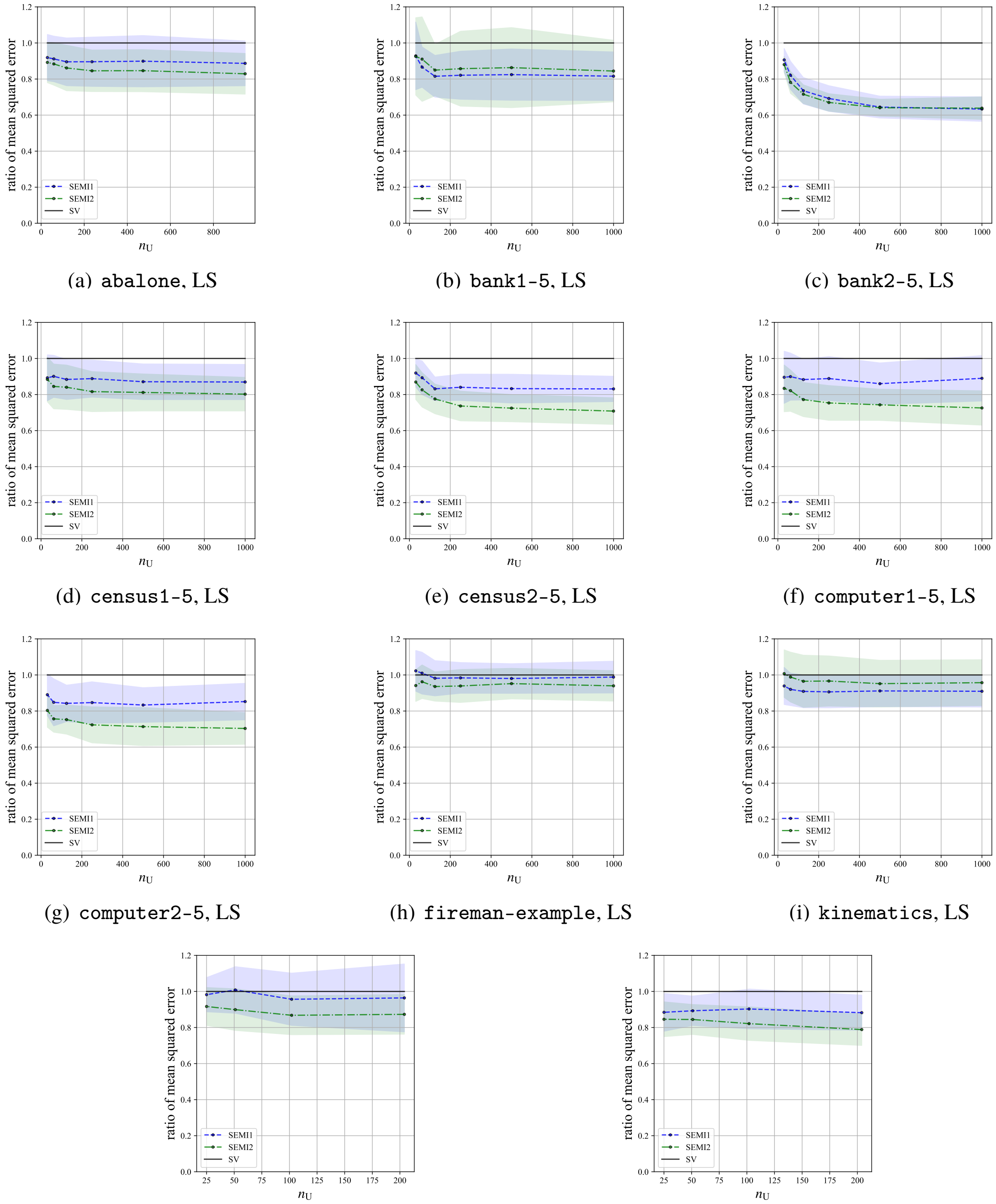
(h) fireman-example, IT

(i) kinematics, IT

(j) lev, IT

(k) swd, IT

Figure 8: The ratio between the error of the empirical semi-supervised risk to the supervised risk when we adopted IT as the task surrogate loss and used a linear-in-input model. The ratio below 1.0 implies that the error of the semi-supervised method is better than that of the supervised risk.



(a) `abalone`, LS

(b) `bank1-5`, LS

(c) `bank2-5`, LS

(d) `census1-5`, LS

(e) `census2-5`, LS

(f) `computer1-5`, LS

(g) `computer2-5`, LS

(h) `fireman-example`, LS

(i) `kinematics`, LS

(j) `lev`, LS

(k) `swd`, LS

Figure 9: The ratio between the error of the empirical semi-supervised risk to the supervised risk when we adopted LS as the task surrogate loss and used a linear-in-input model. The ratio below 1.0 implies that the error of the semi-supervised method is better than that of the supervised risk.