Journal of Machine Learning Research 14 (2013) 1461-1504 Submitted 4/12; Revised 1/13; Published 6/13

Conjugate Relation between Loss Functions and Uncertainty Sets in Classification Problems

Takafumi Kanamori KANAMORI@IS.NAGOYA-U.AC.JP
Department of Computer Science and Mathematical Informatics
Nagoya University
Nagoya 464-8603 Japan

Akiko Takeda TAKEDA@MIST.I.U-TOKYO.AC.JP
Taiji Suzuki S-TAIJI@STAT.T.U-TOKYO.AC.JP
Department of Mathematical Informatics
The University of Tokyo
Tokyo 113-8656, Japan

Editor: John Shawe-Taylor

Abstract

There are two main approaches to binary classification problems: the loss function approach and the uncertainty set approach. The loss function approach is widely used in real-world data analysis. Statistical decision theory has been used to elucidate its properties such as statistical consistency. Conditional probabilities can also be estimated by using the minimum solution of the loss function. In the uncertainty set approach, an uncertainty set is defined for each binary label from training samples. The best separating hyperplane between the two uncertainty sets is used as the decision function. Although the uncertainty set approach provides an intuitive understanding of learning algorithms, its statistical properties have not been sufficiently studied. In this paper, we show that the uncertainty set is deeply connected with the convex conjugate of a loss function. On the basis of the conjugate relation, we propose a way of revising the uncertainty set approach so that it will have good statistical properties such as statistical consistency. We also introduce statistical models corresponding to uncertainty sets in order to estimate conditional probabilities. Finally, we present numerical experiments, verifying that the learning with revised uncertainty sets improves the prediction accuracy.

Keywords: loss function, uncertainty set, convex conjugate, consistency

1. Introduction

In classification problems, the goal is to predict output labels for given input vectors. For this purpose, a decision function defined on the input space is estimated from training samples. The output value of the decision function is used for predicting the labels. In binary classification problems, the sign of the decision function is expected to provide an accurate prediction of the labels. Many learning algorithms use loss functions as a penalty of misclassifications. A decision function minimizing the empirical mean of the loss function over the training samples is employed as an estimator (Cortes and Vapnik, 1995; Schölkopf et al., 2000; Freund and Schapire, 1997; Hastie et al., 2001). For example, the hinge loss, exponential loss and logistic loss are used for support vector machine (SVM), Adaboost and logistic regression, respectively. In regards to binary classification tasks, recent studies have elucidated the statistical properties of learning algorithms using loss functions

(see Bartlett et al. 2006; Steinwart 2005, 2003; Schapire et al. 1998; Zhang 2004; Vapnik 1998 for details).

The loss function approach provides not only an estimator of the decision function, but also an estimator of the conditional probability of binary labels for a given input. The sign of the estimated decision function is used for the label prediction, and the magnitude of the decision function is connected to the conditional probability via the loss function. This connection has been studied by many researchers (Friedman et al., 1998; Bartlett and Tewari, 2007). For example, the logistic loss and exponential loss produce logistic models, whereas the hinge loss cannot be used to estimate the conditional probability except the probability 0.5 (Bartlett and Tewari, 2007).

Another approach to binary classification problems, the maximum-margin criterion, is taken in statistical learning. Under the maximum-margin criterion, the best separating hyperplane between the two output labels is used as the decision function. Hard-margin SVM (Vapnik, 1998) defines a convex-hull of input vectors for each binary label, and takes into account the maximum-margin between the two convex-hulls. For the non-separable case, ν-SVM gives us a similar picture (Schölkopf et al., 2000; Bennett and Bredensteiner, 2000). Ellipsoidal sets as well as polyhedral sets such as the convex-hull of finite input points can be used to solve classification problems (Lanckriet et al., 2003; Nath and Bhattacharyya, 2007). In this paper, the set used in the maximum-margin criterion is referred to as an *uncertainty set*. This term comes from the field of robust optimization in mathematical programming (Ben-Tal et al., 2009).

There have been studies on the statistical properties of learning with uncertainty sets. For example, Lanckriet et al. (2003) proposed minimax probability machine (MPM) using ellipsoidal uncertainty sets and studied its statistical properties in the worst-case setting. In statistical learning using uncertainty sets, the main concern is to develop optimization algorithms under the maximum margin criterion (Mavroforakis and Theodoridis, 2006). So far, however, the statistical properties of learning with uncertainty sets have not been studied as much as those of learning with loss functions.

The main purpose of this paper is to study the relation between the loss function approach and uncertainty set approach, and to use the relation to transform learning with uncertainty sets into loss-based learning in order to clarify the statistical properties of learning algorithms. As mentioned above, loss functions naturally involve statistical models of conditional probabilities. As a result, we can establish a correspondence between uncertainty sets and statistical models of conditional probabilities. Note that some of the existing learning methods using uncertainty sets do not necessarily have good statistical properties, such as the statistical consistency. We propose a way of revising uncertainty sets to establish statistical consistency.

Figure 1 shows how uncertainty sets, loss functions and statistical models are related. Starting from a learning algorithm with uncertainty sets, we obtain the corresponding loss function and statistical model via the convex conjugate. Usually, uncertainty sets are designed on the basis of an intuitive understandings of real-world data. By revising uncertainty sets, we can obtain the corresponding loss functions and statistical models. We also derive sufficient conditions under which the corresponding loss function produces a statistically consistent estimator. We think that our method of revising uncertainty sets can bridge the gap between intuitive statistical modeling and the nice statistical properties of learning algorithms.

The paper is organized as follows. Section 2 reviews the existing learning methods using loss functions and uncertainty sets. We describe the relation between the loss function and uncertainty set in ν-SVM. Section 3 is an investigation of the general relation between loss functions and uncertainty sets. In addition, we describe statistical models derived from loss functions. Section 4 shows

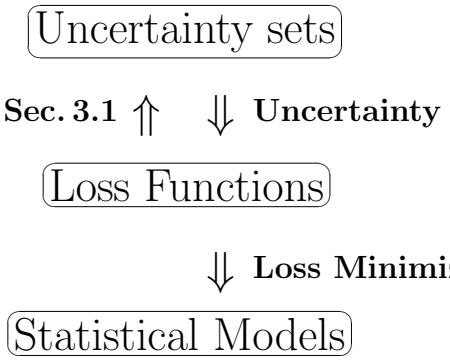


Figure 1: Relations among uncertainty sets, loss functions and statistical models. In Section 3.1, we derive uncertainty sets from loss functions by using the convex conjugate of loss functions. In Section 3.2, we derive statistical models from loss functions. Section 4.2 shows how to revise uncertainty sets in order to obtain loss functions from them. By applying the relations in the diagram, we can transform learning with uncertainty sets into loss-based learning so that we can benefit from good statistical properties such as statistical consistency.

how to revising the uncertainty set so that it will have good statistical properties. Section 5 describes a kernel-based learning algorithm derived from uncertainty sets. Section 6 proves that the kernel-based algorithm has statistical consistency. The results of numerical experiments are described in Section 7. We conclude in section 8. The details of the proofs are shown in the Appendix.

Let us summarize the notations to be used throughout the paper. The indicator function is denoted as $[\![A]\!]$; that is, $[\![A]\!]$ equals 1 if A is true, and 0 otherwise. The column vector $\boldsymbol{x}$ in Euclidean space is written in boldface. The transposition of $\boldsymbol{x}$ is denoted as $\boldsymbol{x}^T$. The Euclidean norm of the vector $\boldsymbol{x}$ is expressed as $\|\boldsymbol{x}\|$. For a set S in a linear space, the convex hull of S is denoted as conv$S$ or conv(S). The number of elements in S is denoted as $|S|$. The expectation of the random variable Z w.r.t. the probability distribution P is described as $\mathbb{E}_P[Z]$. We will drop the subscript P when it is clear from the context. The set of all measurable functions on the set $\mathcal{X}$ relative to the measure P is denoted by L_0. The supremum norm of $f \in L_0$ is denoted as $\|f\|_\infty$. Elements in $\mathcal{X}$ are written in Roman alphabets such as $x \in \mathcal{X}$ if $\mathcal{X}$ is not necessarily a subset of the Euclidean space. For the reproducing kernel Hilbert space $\mathcal{H}$, $\|f\|_{\mathcal{H}}$ is the norm of $f \in \mathcal{H}$ defined from the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ on $\mathcal{H}$.

2. Preliminaries and Previous Studies

We define $\mathcal{X}$ as the input space and $\{+1,-1\}$ as the set of binary labels. Suppose that the training samples $(x_1,y_1),\ldots,(x_m,y_m) \in \mathcal{X} \times \{+1,-1\}$ are drawn i.i.d. according to a probability distribution P on $\mathcal{X} \times \{+1,-1\}$. The goal is to estimate a decision function $f : \mathcal{X} \to \mathbb{R}$ such that the sign of $f(x)$ provides an accurate prediction of the unknown binary label associated with the input x under the probability distribution P. In other words, the probability of $\mathrm{sign}(f(x)) \neq y$ for the estimated

decision function f is expected to be as small as possible.[1] In this article, the composite function of the sign function and the decision function, $\mathrm{sign}(f(x))$, is referred to as classifier.

2.1 Learning with Loss Functions

In binary classification problems, the prediction accuracy of the decision function f is measured by the 0-1 loss $[\![\mathrm{sign}(f(x)) \neq y]\!]$, which equals 1 when the sign of $f(x)$ is different from y and 0 otherwise.

The average prediction performance of the decision function f is evaluated by the expected 0-1 loss, that is,

$$\mathcal{E}(f) = \mathbb{E}[\,[\![\mathrm{sign}(f(x)) \neq y]\!]\,].$$

The Bayes risk $\mathcal{E}^*$ is defined as the minimum value of the expected 0-1 loss over all the measurable functions on $\mathcal{X}$,

$$\mathcal{E}^* = \inf\{\mathcal{E}(f) : f \in L_0\}. \tag{1}$$

The Bayes risk is the lowest achievable error rate given the probability P. Given a set of training samples, $T = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, the empirical 0-1 loss is expressed as

$$\widehat{\mathcal{E}}_T(f) = \frac{1}{m}\sum_{i=1}^{m} [\![\mathrm{sign}(f(x_i)) \neq y_i]\!].$$

In what follows, the subscript T in $\widehat{\mathcal{E}}_T(f)$ will be dropped if it is clear from the context.

In general, minimization of $\widehat{\mathcal{E}}_T(f)$ is a hard problem (Arora et al., 1997). The main difficulty comes from the non-convexity of the 0-1 loss $[\![\mathrm{sign}(f(x)) \neq y]\!]$ as a function of f. Hence, many learning algorithms use a surrogate loss in order to make the computation tractable. For example, SVM uses the hinge loss, $\max\{1 - yf(x), 0\}$, and Adaboost uses the exponential loss, $\exp\{-yf(x)\}$. Both the hinge loss and the exponential loss are convex in f, and they provide an upper bound of the 0-1 loss. Thus, the minimizer under the surrogate loss is also expected to minimize the 0-1 loss. The quantitative relation between the 0-1 loss and the surrogate loss was studied by Bartlett et al. (2006) and Zhang (2004).

Regularization is used to avoid overfitting of the estimated decision function to the training samples. The complexity of the estimated classifier is limited by adding a regularization term such as the squared norm of the decision function to an empirical surrogate loss. The balance between the regularization term and the surrogate loss is adjusted by using a regularization parameter (Evgeniou et al., 1999; Steinwart, 2005). Accordingly, regularization controls the deviation of the empirical loss from the expected loss. The optimization is computationally tractable when both the regularization term and the surrogate loss are convex.

Besides computational tractability, surrogate loss functions have another benefit. As discussed by Friedman et al. (1998) and Bartlett and Tewari (2007), surrogate loss functions provide statistical models for the conditional probability of a label y for a given x, that is, $P(y|x)$. A brief introduction to this idea is given below.

1. As Bartlett et al. (2006) pointed out, the particular choice of the value of $\mathrm{sign}(0)$ is not important, but we need to choose some value in $\{+1, -1\}$.

Let us consider a minimization problem of the expected loss, $\min_{f \in L_0} \mathbb{E}[\ell(-yf(x))]$, where $\ell(-yf(x))$ is a surrogate loss of a decision function $f(x)$. The function $\ell : \mathbb{R} \to \mathbb{R}$ is assumed to be differentiable. In a similar way to Lemma 1 of Friedman et al. (1998), it is sufficient to minimize the loss function conditional on x:

$$\mathbb{E}[\ell(-yf(x))|x] = P(y = +1|x)\ell(-f(x)) + P(y = -1|x)\ell(f(x)).$$

At the optimal solution, the derivative is equal to zero, that is,

$$\frac{\partial}{\partial f(x)} \mathbb{E}[\ell(-yf(x))|x] = -P(y = +1|x)\ell'(-f(x)) + P(y = -1|x)\ell'(f(x)),$$

where ℓ' is the derivative of ℓ. Therefore, we have

$$P(y = +1|x) = \frac{\ell'(f(x))}{\ell'(f(x)) + \ell'(-f(x))}$$

for the optimal solution f. An estimator of the conditional probability can be obtained by substituting an estimated decision function into the above expression. For example, the exponential loss $\exp\{-yf(x)\}$ yields the logistic model

$$P(y = +1|x) = \frac{e^{f(x)}}{e^{f(x)} + e^{-f(x)}}.$$

The relation between surrogate losses and statistical models was extensively studied by Bartlett and Tewari (2007).

2.2 Learning with Uncertainty Sets

Besides statistical learning using loss functions, there is another approach to binary classification problems, that is, statistical learning based on the *uncertainty set*. What follows is a brief introduction to the basic idea of the uncertainty set. We assume that $\mathcal{X}$ is a subset of Euclidean space.

Uncertainty sets describe uncertainties or ambiguities present in robust optimization problems (Ben-Tal et al., 2009). The parameter in the optimization problem may not be precisely determined. For example, in portfolio optimization, the objective function may depend on a future stock price. Instead of precise information, we have an uncertainty set which probably includes the true parameter of the optimization problem. Typically, the worst case is the setting in which the robust optimization problem with uncertainty sets is solved.

Statistical learning with uncertainty sets is an application of robust optimization to classification problems. An uncertainty set is prepared for each binary label. Each uncertainty set is assumed to include the mean vector of the distribution of input point $\boldsymbol{x}$ conditioned on each label (Takeda et al., 2013). For example, $\mathcal{U}_p$ and $\mathcal{U}_n$ are confidence regions such that the conditional probabilities, $P(\boldsymbol{x} \in \mathcal{U}_p|y = +1)$ and $P(\boldsymbol{x} \in \mathcal{U}_n|y = -1)$, are both equal to 0.95. Another example is one in which the uncertainty set $\mathcal{U}_p$ (resp. $\mathcal{U}_n$) consists of the convex hull of input vectors in training samples having the positive (resp. negative) label. The convex hull of data points is used in hard margin SVM (Bennett and Bredensteiner, 2000). An ellipsoidal uncertainty set is also used for the robust classification in the worst-case setting (Lanckriet et al., 2003; Nath and Bhattacharyya, 2007).

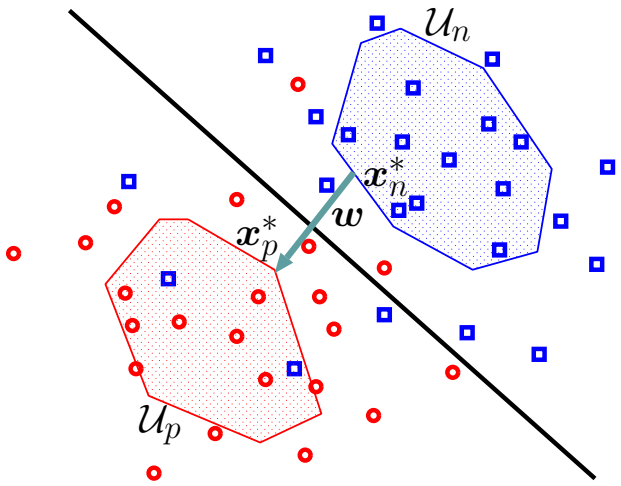


Figure 2: Decision boundary estimated by solving the minimum distance problem with the uncertainty sets $\mathcal{U}_p$ and $\mathcal{U}_n$.

We use the uncertainty set to estimate the linear decision function $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$. Here, let us consider the *minimum distance problem*

$$\min_{\boldsymbol{x}_p, \boldsymbol{x}_n} \|\boldsymbol{x}_p - \boldsymbol{x}_n\| \quad \text{subject to } \boldsymbol{x}_p \in \mathcal{U}_p, \boldsymbol{x}_n \in \mathcal{U}_n. \tag{2}$$

Let $\boldsymbol{x}_p^*$ and $\boldsymbol{x}_n^*$ be optimal solutions of (2). Then, the normal vector of the decision function, $\boldsymbol{w}$, can be estimated with $c(\boldsymbol{x}_p^* - \boldsymbol{x}_n^*)$, where c is a positive real number. Figure 2 illustrates the estimated decision boundary. When both $\mathcal{U}_p$ and $\mathcal{U}_n$ are compact subsets satisfying $\mathcal{U}_p \cap \mathcal{U}_n = \emptyset$, the estimated normal vector cannot be the null vector. The minimum distance problem appears in the hard margin SVM (Vapnik, 1998; Bennett and Bredensteiner, 2000), ν-SVM (Schölkopf et al., 2000; Crisp and Burges, 2000) and the learning algorithms proposed by Nath and Bhattacharyya (2007) and Mavroforakis and Theodoridis (2006). Section 2.3 briefly describes the relation between ν-SVM and the minimum distance problem. Another criterion is used to estimate the linear decision function in minimax probability machine (MPM) proposed by Lanckriet et al. (2003), but the ellipsoidal uncertainty set also plays an important role in MPM.

The minimum distance problem is equivalent to the maximum margin principle (Vapnik, 1998; Bennett and Bredensteiner, 2000). When the bias term b in the linear decision function is estimated such that the decision boundary bisects the line segment connecting $\boldsymbol{x}_p^*$ and $\boldsymbol{x}_n^*$, the estimated decision boundary will have the maximum margin between the uncertainty sets, $\mathcal{U}_p$ and $\mathcal{U}_n$. Takeda et al. (2013) studied the relation between the minimum distance problem and the maximum margin principle.

2.3 Loss Functions and Uncertainty Sets in ν-SVM

Here, we will describe how the loss function approach and uncertainty set approach are related to each other in ν-SVM (Schölkopf et al., 2000). We will follow the presentation laid out in Crisp and

Burges (2000) and Bennett and Bredensteiner (2000). We will extend this relation to more general learning algorithms in Section 3.

Suppose that the input space $\mathcal{X}$ is a subset of Euclidean space $\mathbb{R}^d$, and we have the linear decision function, $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$, where the normal vector $\boldsymbol{w} \in \mathbb{R}^d$ and the bias term $b \in \mathbb{R}$ are parameters to be estimated based on training samples. By applying the kernel trick (Berlinet and Thomas-Agnan, 2004; Schölkopf and Smola, 2002), we can obtain rich statistical models for the decision function, while maintaining computational tractability.

The decision function used in ν-SVM is estimated as the optimal solution of

$$\min_{w,b,\rho} \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} \max\{\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), 0\}, \quad \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R},\ \rho \in \mathbb{R}, \tag{3}$$

where $\nu \in (0,1)$ is a prespecified constant that acts as the regularization parameter. ν-SVM uses a variant of the hinge loss, $\max\{\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), 0\}$, as a surrogate loss. As Schölkopf et al. (2000) pointed out, the parameter ν controls the margin errors and number of support vectors. Roughly speaking, the derivative of the objective function with respect to ρ yields

$$\frac{1}{m}\sum_{i=1}^{m} [\![y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < \rho]\!] = \nu, \tag{4}$$

where the subdifferential at $\rho = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$ has been ignored for simplicity. The left side of (4) is called the margin error. The quantity $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$ is referred to as the margin, and the equality above implies that an optimal ρ is the ν-quantile of the empirical distribution of margins $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$, $i = 1, \ldots, m$. The empirical loss in ν-SVM is minimized over training samples such that $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < \rho$, and training samples having a large margin, that is, $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq \rho$, do not contribute to the loss function $\max\{\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), 0\}$. As a result, the sum of the second and third terms in ν-SVM (3) imply the mean of the negative margin $-y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$ such that $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < \rho$, that is,

$$-\nu\rho + \frac{1}{m}\sum_{i=1}^{m} \max\{\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), 0\} = \nu \cdot \frac{1}{m\nu} \sum_{i:y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < \rho} (-y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

at the optimal solution. The above loss function is known as the conditional value-at-risk in the field of mathematical finance (Rockafellar and Uryasev, 2002). The relation between ν-SVM and the conditional value-at-risk was studied by Takeda and Sugiyama (2008).

The original formulation of ν-SVM uses a non-negativity constraint, $\rho \geq 0$. As shown by Crisp and Burges (2000), the non-negativity constraint is redundant. Indeed, for an optimal solution $\widehat{\boldsymbol{w}}, \widehat{b}, \widehat{\rho}$, we have

$$-\nu\widehat{\rho} \leq \frac{1}{2}\|\widehat{\boldsymbol{w}}\|^2 - \nu\widehat{\rho} + \frac{1}{m}\sum_{i=1}^{m} \max\{\widehat{\rho} - y_i(\widehat{\boldsymbol{w}}^T\boldsymbol{x}_i + \widehat{b}), 0\} \leq 0,$$

where the last inequality comes from the fact that the parameter, $\boldsymbol{w} = \boldsymbol{0}$, $b = 0$, $\rho = 0$, is a feasible solution of (3). As a result, we have $\widehat{\rho} \geq 0$ for $\nu > 0$.

Now let us briefly show that the dual problem of (3) yields a minimum distance problem in which the reduced convex-hulls of training samples are used as uncertainty sets (See Bennett and

Bredensteiner 2000 for details). Problem (3) is equivalent to

$$\min_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i, \quad \text{subject to } \xi_i \geq 0,\ \xi_i \geq \rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b),\ i = 1,\ldots,m.$$

The Lagrangian function is defined as

$$L(\boldsymbol{w},b,\rho,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - \xi_i) - \sum_{i=1}^{m}\beta_i\xi_i,$$

where α_i, β_i, $i = 1,\ldots,m$ are non-negative Lagrange multipliers. For the training samples, we define M_p and M_n as the set of sample indices for each label, that is,

$$M_p = \{i \mid y_i = +1\}, \quad M_n = \{i \mid y_i = -1\}. \tag{5}$$

The min-max theorem (Bertsekas et al., 2003, Proposition 6.4.3) provides

$$\begin{aligned}
&\inf_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}} \sup_{\boldsymbol{\alpha}\geq\mathbf{0},\boldsymbol{\beta}\geq\mathbf{0}} L(\boldsymbol{w},b,\rho,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) \\
&= \sup_{\boldsymbol{\alpha}\geq\mathbf{0},\boldsymbol{\beta}\geq\mathbf{0}} \inf_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}} L(\boldsymbol{w},b,\rho,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) \\
&= \sup_{\boldsymbol{\alpha}\geq\mathbf{0},\boldsymbol{\beta}\geq\mathbf{0}} \inf_{b,\rho,\boldsymbol{\xi}} -\frac{1}{2}\Big\|\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i\Big\|^2 + \sum_{i=1}^{m}\xi_i\left(\frac{1}{m} - \alpha_i - \beta_i\right) + \rho\left(\sum_{i=1}^{m}\alpha_i - \nu\right) - b\sum_{i=1}^{m}\alpha_i y_i \qquad (6) \\
&= \sup_{\boldsymbol{\alpha}}\left\{ -\frac{1}{2}\Big\|\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i\Big\|^2 \ :\ \sum_{i=1}^{m}\alpha_i = \nu,\ \sum_{i=1}^{m}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq \frac{1}{m}\right\} \\
&= -\frac{\nu^2}{8}\inf_{\boldsymbol{\alpha}}\left\{ \Big\|\sum_{i\in M_p}\gamma_i\boldsymbol{x}_i - \sum_{j\in M_n}\gamma_j\boldsymbol{x}_j\Big\|^2 \ :\ \sum_{i\in M_p}\gamma_i = \sum_{i\in M_n}\gamma_i = 1,\ 0 \leq \gamma_i \leq \frac{2}{m\nu}\right\}. \qquad (7)
\end{aligned}$$

The following equalities should hold in (6) above

$$\frac{1}{m} - \alpha_i - \beta_i = 0,\ (i = 1,\ldots,m), \quad \sum_{i=1}^{m}\alpha_i - \nu = 0, \quad \sum_{i=1}^{m}\alpha_i y_i = 0.$$

Otherwise the objective value tends to $-\infty$. The last equality (7) is obtained by changing the variable from α_i to $\gamma_i = 2\alpha_i/\nu$.

For the positive (resp. negative) label, we introduce the uncertainty set $\mathcal{U}_p$ (reps. $\mathcal{U}_n$) defined by the reduced convex-hull, that is,

$$o \in \{p,n\}, \quad \mathcal{U}_o = \left\{ \sum_{i\in M_o}\gamma_i\boldsymbol{x}_i \ :\ \sum_{i\in M_o}\gamma_i = 1,\ 0 \leq \gamma_i \leq \frac{2}{m\nu}, i \in M_o\right\}.$$

When the upper limit of γ_i is less than one, the reduced convex-hull is a subset of the convex-hull of training samples. Hence, solving problem (7) is identical to solving the minimum distance problem with the uncertainty set of reduced convex hulls,

$$\inf_{\boldsymbol{x}_p,\boldsymbol{x}_n} \|\boldsymbol{x}_p - \boldsymbol{x}_n\| \quad \text{subject to } \boldsymbol{x}_p \in \mathcal{U}_p,\ \boldsymbol{x}_n \in \mathcal{U}_n.$$

If the loss function in ν-SVM is scaled, such as,

$$\frac{1}{\nu}\|\boldsymbol{w}\|^2 - 2\rho + \frac{1}{m}\sum_{i=1}^{m}\frac{2}{\nu}\max\{\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), 0\}, \tag{8}$$

the variable change from the Lagrange multipliers $\alpha_1,\ldots,\alpha_m$ to $\gamma_1,\ldots,\gamma_m$, as shown in (7), is not required to obtain uncertainty sets, $\mathcal{U}_p$ and $\mathcal{U}_n$.

3. Relation between Loss Functions and Uncertainty Sets

Here, we present an extension of ν-SVM with which we can investigate the relation between loss functions and uncertainty sets.

3.1 Uncertainty Sets Associated with Loss Functions

The decision function is defined as $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ on $\mathbb{R}^d$, and let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex non-decreasing function. For training samples, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$, we propose the following learning method, which is an extension of ν-SVM with the expression (8),

$$\inf_{\boldsymbol{w},b,\rho} -2\rho + \frac{1}{m}\sum_{i=1}^{m} \ell(\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) \ \text{ subject to } \ \|\boldsymbol{w}\|^2 \le \lambda^2,\ b \in \mathbb{R}, \rho \in \mathbb{R}. \tag{9}$$

The regularization effect is introduced by the constraint $\|\boldsymbol{w}\|^2 \le \lambda^2$, where λ is a regularization parameter which may depend on the sample size. The above formulation makes the proof of statistical consistency in Section 6 rather simple. Note that ν-SVM is recovered by setting $\ell(z) = \max\{2z/\nu, 0\}$ with an appropriate λ.

Let us consider the role of the parameter ρ in (9). As described in Section 2.3, ρ in ν-SVM is chosen adaptively, and as a result, training samples with a small margin such as $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < \rho$ suffer a penalty. The number of training samples suffering a penalty is determined by the parameter ν, and the optimal ρ is the ν-quantile of the empirical margin distribution. As shown below, the ν parameter of ν-SVM is related to the slope of the loss function $\ell(z)$ in (9). In the extended formulation, the ρ parameter in (9) is also adaptively estimated, and it is regarded as a soft-threshold; that is, training samples with margins less than ρ suffer large penalties. The number of such training samples is determined by the extremal condition of (9) with respect to ρ:

$$\frac{1}{m}\sum_{i=1}^{m} \ell'(\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) = 2,$$

where ℓ is assumed to have a derivative ℓ'. In the generalized learning algorithm, the magnitude of the derivative ℓ' roughly controls the optimal ρ and the samples size such that margins are smaller than ρ. Note that by placing a mild assumption on ℓ, the first term -2ρ in (9) prevents ρ from going to $-\infty$. The factor 2 in -2ρ can be replaced with an arbitrary positive constant, since multiplying a positive constant by the objective function does not change the optimal solution. However, the factor 2 makes the calculation and interpretation of the dual expression somewhat simpler, as described in the previous section.

We can derive the uncertainty set associated with the loss function ℓ in (9) in a similar way to what was done with ν-SVM. We introduce slack variables $\xi_i, i = 1, \ldots, m$ satisfying inequalities $\xi_i \ge \rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), i = 1, \ldots, m$. Accordingly, the Lagrangian (9) becomes

$$L(\boldsymbol{w}, b, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mu) = -2\rho + \frac{1}{m}\sum_{i=1}^{m} \ell(\xi_i) + \sum_{i=1}^{m} \alpha_i(\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - \xi_i) + \mu(\|\boldsymbol{w}\|^2 - \lambda^2),$$

where $\alpha_1, \ldots, \alpha_m$ and μ are non-negative Lagrange multipliers. We define the convex conjugate of $\ell(z)$ as

$$\ell^*(\alpha) = \sup_{z \in \mathbb{R}} \{z\alpha - \ell(z)\}.$$

The properties of the convex conjugate are summarized in Appendix A. The convex conjugate is mainly used to improve the computational efficiency of learning algorithms (Sun and Shawe-Taylor, 2010). Here, we use the convex conjugate of the loss function to connect seemingly different styles of learning algorithms.

The min-max theorem leads us to the dual problem as follows,

$$\begin{aligned}
&\inf_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}} \sup_{\boldsymbol{\alpha}\geq\mathbf{0},\mu\geq 0} L(\boldsymbol{w},b,\rho,\boldsymbol{\xi},\boldsymbol{\alpha},\mu)\\
&= \sup_{\boldsymbol{\alpha}\geq\mathbf{0},\mu\geq 0} \inf_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}} L(\boldsymbol{w},b,\rho,\boldsymbol{\xi},\boldsymbol{\alpha},\mu)\\
&= \sup_{\boldsymbol{\alpha}\geq\mathbf{0},\mu\geq 0} \inf_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}} \left\{ \rho\left(\sum_{i=1}^m \alpha_i - 2\right) - b\sum_{i=1}^m \alpha_i y_i \right.\\
&\qquad\qquad \left. - \frac{1}{m}\sum_{i=1}^m (m\alpha_i\xi_i - \ell(\xi_i)) - \sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i^T\boldsymbol{w} + \mu(\|\boldsymbol{w}\|^2 - \lambda^2) \right\}\\
&= -\inf_{\boldsymbol{\alpha}\geq\mathbf{0},\mu\geq 0} \left\{ \frac{1}{m}\sum_{i=1}^m \ell^*(m\alpha_i) + \frac{1}{4\mu}\Big\|\sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i\Big\|^2 + \mu\lambda^2 \,:\, \sum_{i=1}^m \alpha_i - 2 = 0, \sum_{i=1}^m \alpha_i y_i = 0 \right\}\\
&= -\inf_{\boldsymbol{\alpha}} \left\{ \frac{1}{m}\sum_{i=1}^m \ell^*(m\alpha_i) + \lambda\Big\|\sum_{i\in M_p}\alpha_i\boldsymbol{x}_i - \sum_{i\in M_n}\alpha_i\boldsymbol{x}_i\Big\| \,:\, \sum_{i\in M_p}\alpha_i = \sum_{i\in M_n}\alpha_i = 1, \alpha_i\geq 0 \right\}. \qquad (10)
\end{aligned}$$

Section 6 presents a rigorous proof that by placing certain assumptions on $\ell(\xi)$, the min-max theorem works in the above Lagrangian function; that is, there is no duality gap. For each binary label, we define parametrized uncertainty sets, $\mathcal{U}_p[c]$ and $\mathcal{U}_n[c]$, by

$$o\in\{p,n\},\quad \mathcal{U}_o[c] = \left\{ \sum_{i\in M_o}\alpha_i\boldsymbol{x}_i \,:\, \alpha_i\geq 0,\ \sum_{i\in M_o}\alpha_i = 1,\ \frac{1}{m}\sum_{i\in M_o}\ell^*(m\alpha_i)\leq c \right\}. \qquad (11)$$

Accordingly, the optimization problem in (10) can be represented as

$$\begin{aligned}
&\inf_{c_p,c_n,\boldsymbol{z}_p,\boldsymbol{z}_n} c_p + c_n + \lambda\big\|\boldsymbol{z}_p - \boldsymbol{z}_n\big\|\\
&\quad \text{subject to } \ \boldsymbol{z}_p\in\mathcal{U}_p[c_p],\ \boldsymbol{z}_n\in\mathcal{U}_n[c_n],\ c_p, c_n\in\mathbb{R}.
\end{aligned} \qquad (12)$$

Let $\widehat{\boldsymbol{z}}_p$ and $\widehat{\boldsymbol{z}}_n$ be the optimal solution of $\boldsymbol{z}_p$ and $\boldsymbol{z}_n$ in (12). Let $\widehat{\boldsymbol{w}}$ be an optimal solution of $\boldsymbol{w}$ in (9). The saddle point of the above min-max problem (10) leads to the relation between $\widehat{\boldsymbol{z}}_p$, $\widehat{\boldsymbol{z}}_n$ and $\widehat{\boldsymbol{w}}$. Some calculation yields that $\widehat{\boldsymbol{w}} = \lambda(\widehat{\boldsymbol{z}}_p - \widehat{\boldsymbol{z}}_n)/\|\widehat{\boldsymbol{z}}_p - \widehat{\boldsymbol{z}}_n\|$ holds for $\widehat{\boldsymbol{z}}_p\neq\widehat{\boldsymbol{z}}_n$, and for $\widehat{\boldsymbol{z}}_p = \widehat{\boldsymbol{z}}_n$ any vector such that $\|\widehat{\boldsymbol{w}}\|^2\leq\lambda^2$ satisfies the KKT condition of (9).

The shape of uncertainty sets and the max-margin criterion respectively correspond to the loss function and the regularization principle. Moreover, the size of the uncertainty set is determined by the regularization parameter. Now let us show some examples of uncertainty sets (11) associated with popular loss functions. The index sets in the following examples, M_p and M_n, are defined by (5) for the training samples $(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_m,y_m)$, and m_p and m_n be $m_p = |M_p|$ and $m_n = |M_n|$.

Example 1 (ν-SVM) *Problem* (9) *with* $\ell(z) = \max\{2z/\nu, 0\}$ *reduces to* ν*-SVM. The conjugate function of* ℓ *is*

$$\ell^*(\alpha) = \begin{cases} 0, & \alpha\in[0,2/\nu],\\ \infty, & \alpha\notin[0,2/\nu], \end{cases}$$

and the associated uncertainty set is defined by

$$o \in \{p,n\}, \quad \mathcal{U}_o[c] = \begin{cases} \left\{ \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i \,:\, \sum_{i \in M_o} \alpha_i = 1, 0 \le \alpha_i \le \dfrac{2}{m\nu}, i \in M_o \right\}, & c \ge 0, \\ \emptyset, & c < 0. \end{cases}$$

For $c \ge 0$, the uncertainty set consists of the reduced convex hull of the training samples, and it does not depend on the parameter c. In addition, a negative c is infeasible. Hence, the optimal solutions of c_p and c_n in the problem (12) *are $c_p = c_n = 0$, and the problem reduces to a simple minimum distance problem.*

Example 2 (Truncated quadratic loss) *Let us now consider $\ell(z) = (\max\{1+z, 0\})^2$. The conjugate function is*

$$\ell^*(\alpha) = \begin{cases} -\alpha + \dfrac{\alpha^2}{4}, & \alpha \ge 0, \\ \infty, & \alpha < 0. \end{cases}$$

For $o \in \{p,n\}$, we define $\bar{\boldsymbol{x}}_o$ and $\widehat{\Sigma}_o$ as the empirical mean and the empirical covariance matrix of the samples $\{\boldsymbol{x}_i : i \in M_o\}$, that is,

$$\bar{\boldsymbol{x}}_o = \frac{1}{m_o} \sum_{i \in M_o} \boldsymbol{x}_i, \quad \widehat{\Sigma}_o = \frac{1}{m_o} \sum_{i \in M_o} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_o)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_o)^T.$$

Suppose that $\widehat{\Sigma}_o$ is invertible. Then, the uncertainty set corresponding to the truncated quadratic loss is

$$\begin{aligned} o \in \{p,n\}, \quad \mathcal{U}_o[c] &= \left\{ \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i \,:\, \sum_{i \in M_o} \alpha_i = 1, \alpha_i \ge 0, i \in M_o, \sum_{i \in M_o} \alpha_i^2 \le \frac{4(c+1)}{m} \right\} \\ &= \left\{ \boldsymbol{z} \in \mathrm{conv}\{\boldsymbol{x}_i : i \in M_o\} \,:\, (\boldsymbol{z} - \bar{\boldsymbol{x}}_o)^T \widehat{\Sigma}_o^{-1} (\boldsymbol{z} - \bar{\boldsymbol{x}}_o) \le \frac{4(c+1)m_o}{m} \right\}. \end{aligned}$$

To prove the second equality, let us define a matrix $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m_o}) \in \mathbb{R}^{d \times m_o}$. For $\boldsymbol{\alpha}_o = (\alpha_i)_{i \in M_o}$ satisfying the constraints, we get

$$\boldsymbol{z} = \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i = (X - \bar{\boldsymbol{x}}_o \mathbf{1}^T) \boldsymbol{\alpha}_o + \bar{\boldsymbol{x}}_o,$$

where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^{m_o}$. The singular value decomposition of the matrix $X - \bar{\boldsymbol{x}}_o \mathbf{1}^T$ and the constraint $\|\boldsymbol{\alpha}_o\|^2 \le 4(c+1)/m$ yield the second equality. A similar uncertainty set is used in minimax probability machine (MPM) (Lanckriet et al., 2003) and maximum margin MPM (Nath and Bhattacharyya, 2007), though the constraint, $\boldsymbol{z} \in \mathrm{conv}\{\boldsymbol{x}_i : i \in M_o\}$, is not imposed.

Example 3 (exponential loss) *The loss function $\ell(z) = e^z$ is used in Adaboost (Freund and Schapire, 1997; Friedman et al., 1998). The conjugate function is equal to*

$$\ell^*(\alpha) = \begin{cases} -\alpha + \alpha \log \alpha, & \alpha \ge 0, \\ \infty, & \alpha < 0. \end{cases}$$

Hence, the corresponding uncertainty set is

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i : \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0, i \in M_o, \sum_{i \in M_o} \alpha_i \log \frac{\alpha_i}{1/m_o} \leq c + 1 + \log \frac{m_o}{m} \right\}$$

for $o \in \{p,n\}$. The Kullback-Leibler divergence from the weights $\alpha_i, i \in M_o$ to the uniform weight is bounded from above in the uncertainty set.

3.2 Statistical Models Associated with Uncertainty Sets

The extended minimum distance problem (12) with the parametrized uncertainty set (11) corresponds to the loss function in (9). We will show the relation between decision functions and conditional probabilities in a similar way to what is shown in Section 2.1. However, instead of the linear decision function $\boldsymbol{w}^T \boldsymbol{x} + b$, we will consider any measurable function $f \in L_0$.

In the learning algorithm (9), the loss function $-2\rho + \ell(\rho - yf(x))$ is used for estimating the decision function. When the sample size tends to infinity, the objective function converges in probability to $\mathbb{E}[-2\rho + \ell(\rho - yf(x))]$. We will show a minimum solution of the expected loss for $f \in L_0$. As described in Section 2.1, it is sufficient to minimize the loss function conditional on x. Suppose that ρ^* is the optimal solution of $\mathbb{E}[-2\rho + \ell(\rho - yf(x))]$, and let us minimize $\mathbb{E}[-2\rho^* + \ell(\rho^* - yf(x))|x]$ with respect to $f(x)$, which leads to solving

$$\begin{aligned} &\frac{\partial}{\partial f(x)} \mathbb{E}[-2\rho^* + \ell(\rho^* - yf(x))|x] \\ &= -P(y = +1|x)\ell'(\rho^* - f(x)) + P(y = -1|x)\ell'(\rho^* + f(x)) = 0. \end{aligned}$$

The extremal condition yields

$$P(y = +1|x) = \frac{\ell'(\rho^* + f(x))}{\ell'(\rho^* + f(x)) + \ell'(\rho^* - f(x))} \tag{13}$$

for the optimal solution $\rho^* \in \mathbb{R}$ and $f \in L_0$. An estimator of the conditional probability can be obtained by substituting estimated parameters into the above expression. Given the uncertainty set (11), the corresponding statistical model is defined as (13) via the loss function $\ell(z)$.

4. Revision of Uncertainty Sets

Section 3.1 derived parametrized uncertainty sets associated with convex loss functions. Conversely, if an uncertainty set is represented as the form of (11), a corresponding loss function exists. There are many mathematical tools to analyze loss-based estimators. However, if the uncertainty set does not have the form of (11), the corresponding loss function does not exist. One way to deal with the drawback is to revise the uncertainty set so that it possesses a corresponding loss function. This section is devoted to this idea.

Let us consider two different representations of a parametrized uncertainty set: the vertex representation, and the level-set representation. For index sets M_p and M_n defined in (5), let $m_p = |M_p|$ and $m_n = |M_n|$. For $o \in \{p,n\}$, let L_o be a closed, convex, proper function on $\mathbb{R}^{m_o}$, and L_o^* be the conjugate function of L_o. The argument of L_o^* is represented by $\boldsymbol{\alpha}_o = (\alpha_i)_{i \in M_o}$. The *vertex*

representation of the uncertainty set is defined as

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i : L_o^*(\boldsymbol{\alpha}_o) \leq c \right\}, \quad o \in \{p, n\}. \tag{14}$$

Example 2 uses the function $L_o^*(\boldsymbol{\alpha}_o) = \frac{m}{4} \sum_{i \in M_o} \alpha_i^2 - 1$. On the other hand, let $h_o : \mathbb{R}^d \to \mathbb{R}$ be a closed, convex, proper function and h_o^* be the conjugate of h_o. The *level-set representation* of the uncertainty set is defined by

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i : h_o^*\Big(\sum_{i \in M_o} \alpha_i \boldsymbol{x}_i \Big) \leq c \right\}, \quad o \in \{p, n\}. \tag{15}$$

The function h_o^* may depend on the population distribution. Now suppose that h_o^* does not depend on sample points, $\boldsymbol{x}_i, i \in M_o$. In Example 2, the second expression of the uncertainty set involves the convex function $h_o^*(\boldsymbol{z}) = (\boldsymbol{z} - \bar{\boldsymbol{x}}_o)^T \widehat{\Sigma}_o^{-1} (\boldsymbol{z} - \bar{\boldsymbol{x}}_o)$. This function does not satisfy the assumption, since h_o^* depends on the training samples via $\bar{\boldsymbol{x}}_o$ and $\widehat{\Sigma}_o$. Instead, the function $h_o^*(\boldsymbol{z}) = (\boldsymbol{z} - \boldsymbol{\mu}_o)^T \Sigma_o^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_o)$ with the population mean $\boldsymbol{\mu}_o$ and the population covariance matrix Σ_o satisfies the condition. When $\boldsymbol{\mu}_o$ and Σ_o are replaced with the estimated parameters based on prior knowledge or samples that are different from the ones used for training, h_o^* with the estimated parameters still satisfies the condition imposed above.

4.1 From Uncertainty Sets to Loss Functions

In popular learning algorithms using uncertainty sets such as hard-margin SVM, ν-SVM, and maximum margin MPM, the decision function is estimated by solving the minimum distance problem (2) with $\mathcal{U}_p = \mathcal{U}_p[\bar{c}_p]$ and $\mathcal{U}_n = \mathcal{U}_n[\bar{c}_n]$, where $\bar{c}_p$ and $\bar{c}_n$ are fixed constants. To investigate the statistical properties of learning algorithms using uncertainty sets, we will consider the primal expression of a variant of the minimum distance problem (2).

In Section 3, we expressed problem (12) as the dual form of (9). Here, let us consider the following optimization problem to obtain a loss function corresponding to a given uncertainty set:

$$\begin{aligned} &\min_{c_p, c_n, \boldsymbol{z}_p, \boldsymbol{z}_n} \; c_p + c_n + \lambda \|\boldsymbol{z}_p - \boldsymbol{z}_n\| \\ &\text{subject to} \;\; c_p, c_n \in \mathbb{R}, \\ &\qquad\qquad \boldsymbol{z}_p \in \mathcal{U}_p[c_p] \cap \mathrm{conv}\{\boldsymbol{x}_i : i \in M_p\}, \\ &\qquad\qquad \boldsymbol{z}_n \in \mathcal{U}_n[c_n] \cap \mathrm{conv}\{\boldsymbol{x}_i : i \in M_n\}. \end{aligned} \tag{16}$$

The constraints, $\boldsymbol{z}_o \in \mathrm{conv}\{\boldsymbol{x}_i : i \in M_o\}, o \in \{p, n\}$, are added because the corresponding uncertainty set (11) has them. Suppose that $\mathcal{U}_p[c_p]$ and $\mathcal{U}_n[c_n]$ have the vertex representation (14). Then, (16) is equivalent to

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \; L_p^*(\boldsymbol{\alpha}_p) + L_n^*(\boldsymbol{\alpha}_n) + \lambda \Big\| \sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i \Big\| \\ &\text{subject to} \;\; \sum_{i \in M_p} \alpha_i = 1, \; \sum_{j \in M_n} \alpha_j = 1, \; \alpha_i \geq 0 \; (i = 1, \ldots, m). \end{aligned}$$

If there is no duality gap, the corresponding primal formulation is

$$\begin{aligned} &\inf_{\boldsymbol{w}, b, \rho, \boldsymbol{\xi}_p, \boldsymbol{\xi}_n} \; -2\rho + L_p(\boldsymbol{\xi}_p) + L_n(\boldsymbol{\xi}_n), \\ &\text{subject to} \;\; \rho - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \leq \xi_i, \; i = 1, \ldots, m, \quad \|\boldsymbol{w}\|^2 \leq \lambda^2, \end{aligned} \tag{17}$$

where $\boldsymbol{\xi}_o$ is defined as $\boldsymbol{\xi}_o = (\xi_i)_{i \in M_o}$ for $o \in \{p, n\}$.

In the primal expression (17), L_p and L_n are regarded as loss functions for the decision function $\boldsymbol{w}^T \boldsymbol{x} + b$ on the training samples. In general, however, the loss function is not represented as the empirical mean over training samples.

4.2 Revised Uncertainty Sets and Corresponding Loss Functions

The uncertainty sets can be revised such that the primal form (17) is represented as minimization of the empirical mean of a loss function. Theorem 1 below is the justification for this revision.

Revision of uncertainty set defined by vertex representation: Suppose that the uncertainty set is described by (14). For $o \in \{p, n\}$, we define m_o-dimensional vectors $\mathbf{1}_o = (1, \ldots, 1)$ and $\mathbf{0}_o = (0, \ldots, 0)$. For the convex function $L_o^* : \mathbb{R}^{m_o} \to \mathbb{R}$, we define $\bar{\ell}^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ by

$$\bar{\ell}^*(\alpha) = \begin{cases} L_p^*(\frac{\alpha}{m}\mathbf{1}_p) + L_n^*(\frac{\alpha}{m}\mathbf{1}_n) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases} \tag{18}$$

The revised uncertainty set $\bar{\mathcal{U}}_o[c]$, $o \in \{p, n\}$ is defined as

$$\bar{\mathcal{U}}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i \,:\, \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0, i \in M_o, \frac{1}{m} \sum_{i \in M_o} \bar{\ell}^*(\alpha_i m) \leq c \right\}. \tag{19}$$

Revision of uncertainty set defined by level-set representation: Suppose that the uncertainty set is described by (15) and that the mean of the input vector $\boldsymbol{x}$ conditioned on the positive (resp. negative) label is given as $\boldsymbol{\mu}_p$ (resp. $\boldsymbol{\mu}_n$). The null vector is denoted as $\mathbf{0}$. We define the function $\bar{\ell}^* : \mathbb{R} \to \mathbb{R}$ by

$$\bar{\ell}^*(\alpha) = \begin{cases} h_p^*(\alpha\frac{m_p}{m}\boldsymbol{\mu}_p) + h_n^*(\alpha\frac{m_n}{m}\boldsymbol{\mu}_n) - h_p^*(\mathbf{0}) - h_n^*(\mathbf{0}) & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases} \tag{20}$$

For $\bar{\ell}^*(\alpha)$ in (20), the revised uncertainty set $\bar{\mathcal{U}}_o[c], o \in \{p, n\}$ is defined in the same way as (19). We apply a parallel shift to the training samples so as to be $\boldsymbol{\mu}_p \neq \mathbf{0}$ or $\boldsymbol{\mu}_n \neq \mathbf{0}$.

Now let us explain why the revised uncertainty set is defined as it is. When the function $L_p^* + L_n^*$ is described in additive form such as $\sum_{i=1}^m g(\alpha_i)$ for a function g, the uncertainty set defined by the revision (18) does not change. Indeed, Theorem 1 below implies that the transformation of $L_p^* + L_n^*$ into $\frac{1}{m}\sum_{i=1}^m \bar{\ell}^*(\alpha_i m)$ is a projection onto the set of functions with an additive form. In other words, performing the revision twice is the same as performing it once. In addition, the second statement of Theorem 1 means that the projection is uniquely determined when we impose the condition in which the function values on the diagonal $\{(\alpha, \ldots, \alpha) \in \mathbb{R}^m \,:\, \alpha \geq 0\}$ remain unchanged.

Theorem 1 *Let $L_o^* : \mathbb{R}^{m_o} \to \mathbb{R}$, $o \in \{p, n\}$ be convex functions and $\bar{\ell}^*$ be the function defined by* (18) *for given L_p^* and L_n^*. Suppose that $\ell : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is a closed, convex, proper function such that $\ell^*(0) = 0$ and $\ell^*(\alpha) = \infty$ for $\alpha < 0$ hold.*

1. *Suppose that*

$$L_p^*(\boldsymbol{\alpha}_p) + L_n^*(\boldsymbol{\alpha}_n) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) = \frac{1}{m}\sum_{i=1}^{m} \ell^*(\alpha_i m) \tag{21}$$

holds for all non-negative α_i, $i = 1, \ldots, m$. *Then, the equality* $\bar{\ell}^* = \ell^*$ *holds.*

2. *Suppose further that*

$$L_p^*(\alpha \mathbf{1}_p) + L_n^*(\alpha \mathbf{1}_n) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) = \frac{1}{m}\sum_{i=1}^{m} \ell^*(\alpha m) = \ell^*(\alpha m)$$

holds for all $\alpha \geq 0$. *Then, the equality* $\bar{\ell}^* = \ell^*$ *holds.*

Proof Let us prove the first statement. From the definition of $\bar{\ell}^*$ and the assumption placed on ℓ^*, the equality $\ell^*(\alpha) = \bar{\ell}^*(\alpha)$ holds for $\alpha < 0$. Next, suppose $\alpha \geq 0$. The assumption (21) leads to $L_p^*(\frac{\alpha}{m}\mathbf{1}_p) + L_n^*(\frac{\alpha}{m}\mathbf{1}_n) - L_p^*(\mathbf{0}_p) - L_n^*(\mathbf{0}_n) = \ell^*(\alpha)$. Hence, we have $\ell^* = \bar{\ell}^*$. The second statement of the theorem is straightforward. ■

Next, we show that the formula (20) is valid. We want to find a function $\bar{\ell}^*(\alpha)$ such that $h_p^*(\sum_{i\in M_p} \alpha_i \boldsymbol{x}_i) + h_n^*(\sum_{i\in M_n} \alpha_i \boldsymbol{x}_i) - h_p^*(\mathbf{0}) - h_n^*(\mathbf{0})$ is close to $\frac{1}{m}\sum_{i=1}^{m} \bar{\ell}^*(m\alpha_i)$ in some sense. To do so, we substitute $\alpha_i = \alpha/m$ into $h_o^*(\sum_{i\in M_o} \alpha_i \boldsymbol{x}_i)$, $o \in \{p, n\}$. In the large sample limit, $h_o^*(\sum_{i\in M_o} \frac{\alpha}{m} \boldsymbol{x}_i)$ is approximated by $h_o^*(\alpha \frac{m_o}{m} \boldsymbol{\mu}_o)$. Suppose that

$$h_p^*(\alpha \frac{m_p}{m} \boldsymbol{\mu}_p) + h_n^*(\alpha \frac{m_n}{m} \boldsymbol{\mu}_n) - h_p^*(\mathbf{0}) - h_n^*(\mathbf{0})$$

is represented as $\frac{1}{m}\sum_{i=1}^{m} \bar{\ell}^*(\frac{\alpha}{m} m) = \bar{\ell}^*(\alpha)$. As a result, we get (20).

The expanded minimum distance problem using the revised uncertainty sets $\bar{\mathcal{U}}_p[c]$ and $\bar{\mathcal{U}}_n[c]$ is

$$\min_{c_p, c_n, \boldsymbol{z}_p, \boldsymbol{z}_n} c_p + c_n + \lambda \|\boldsymbol{z}_p - \boldsymbol{z}_n\| \quad \text{subject to } \boldsymbol{z}_p \in \bar{\mathcal{U}}_p[c_p],\ \boldsymbol{z}_n \in \bar{\mathcal{U}}_n[c_n]. \tag{22}$$

The corresponding primal problem is

$$\inf_{w, b, \rho, \xi_p, \xi_n} -2\rho + \frac{1}{m}\sum_{i=1}^{m} \bar{\ell}(\xi_i) \qquad \text{subject to } \rho - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \leq \xi_i,\ i = 1, \ldots, m,\ \ \|\boldsymbol{w}\|^2 \leq \lambda^2.$$

The revision of uncertainty sets leads to the empirical mean of the revised loss function $\bar{\ell}$. Asymptotic analysis can be used to study the statistical properties of the estimator given by the optimal solution of (22), since the objective in the primal expression is described by the empirical mean of the revised loss function.

Now let us show some examples to illustrate how revision of uncertainty sets works.

Example 4 *Let* $L_o^*, o \in \{p, n\}$ *be the convex function* $L_o^*(\boldsymbol{\alpha}_o) = \boldsymbol{\alpha}_o^T C_o \boldsymbol{\alpha}_o$, *where* C_o *is a positive definite matrix. When both* C_p *and* C_n *are the identity matrix, the following equality holds:*

$$L_p^*(\boldsymbol{\alpha}_p) + L_n^*(\boldsymbol{\alpha}_n) = \frac{1}{m}\sum_{i=1}^{m} \bar{\ell}^*(\alpha_i m) = \sum_{i=1}^{m} \alpha_i^2.$$

The revised function defined by (18) *is*

$$\bar{\ell}^*(\alpha) = \alpha^2 \frac{\mathbf{1}_p^T C_p \mathbf{1}_p + \mathbf{1}_n^T C_n \mathbf{1}_n}{m^2}$$

for $\alpha \geq 0$. *Accordingly, we get*

$$\frac{1}{m}\sum_{i=1}^{m} \bar{\ell}^*(\alpha_i m) = \frac{\mathbf{1}_p^T C_p \mathbf{1}_p + \mathbf{1}_n^T C_n \mathbf{1}_n}{m} \sum_{i=1}^{m} \alpha_i^2.$$

Let k *be* $k = \mathbf{1}_p^T C_p \mathbf{1}_p + \mathbf{1}_n^T C_n \mathbf{1}_n$. *The revised uncertainty set is*

$$o \in \{p,n\}, \quad \bar{\mathcal{U}}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i \boldsymbol{x}_i \,:\, \sum_{i \in M_o} \alpha_i = 1, \alpha_i \geq 0\,(i \in M_o), \sum_{i \in M_o} \alpha_i^2 \leq \frac{cm}{k} \right\}.$$

For $o \in \{p,n\}$, *let* $\bar{\boldsymbol{x}}_o$ *and* $\widehat{\Sigma}_o$ *be the empirical mean and the empirical covariance matrix,*

$$\bar{\boldsymbol{x}}_o = \frac{1}{m_o} \sum_{i \in M_o} \boldsymbol{x}_i, \quad \widehat{\Sigma}_o = \frac{1}{m_o} \sum_{i \in M_o} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_o)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_o)^T.$$

If $\widehat{\Sigma}_o$ *is invertible, we have*

$$\bar{\mathcal{U}}_o[c] = \left\{ \boldsymbol{z} \in \mathrm{conv}\{\boldsymbol{x}_i : i \in M_o\} \,:\, (\boldsymbol{z} - \bar{\boldsymbol{x}}_o)^T \widehat{\Sigma}_o^{-1} (\boldsymbol{z} - \bar{\boldsymbol{x}}_o) \leq \frac{cmm_o}{k} \right\}.$$

In the learning algorithm based on the revised uncertainty set, the estimator is obtained by solving

$$\begin{aligned} &\min_{c_p, c_n, \boldsymbol{z}_p, \boldsymbol{z}_n} c_p + c_n + \lambda \|\boldsymbol{z}_p - \boldsymbol{z}_n\| \;\; \textit{subject to} \;\; \boldsymbol{z}_p \in \bar{\mathcal{U}}_p[c_p],\, \boldsymbol{z}_n \in \bar{\mathcal{U}}_n[c_n] \\ \Longleftrightarrow &\min_{c_p, c_n, \boldsymbol{z}_p, \boldsymbol{z}_n} c_p + c_n + \frac{m^2\lambda}{4k} \|\boldsymbol{z}_p - \boldsymbol{z}_n\| \;\; \textit{subject to} \;\; \boldsymbol{z}_p \in \bar{\mathcal{U}}_p\left[\frac{4c_p k}{m^2}\right],\, \boldsymbol{z}_n \in \bar{\mathcal{U}}_n\left[\frac{4c_n k}{m^2}\right]. \end{aligned}$$

The corresponding primal expression is

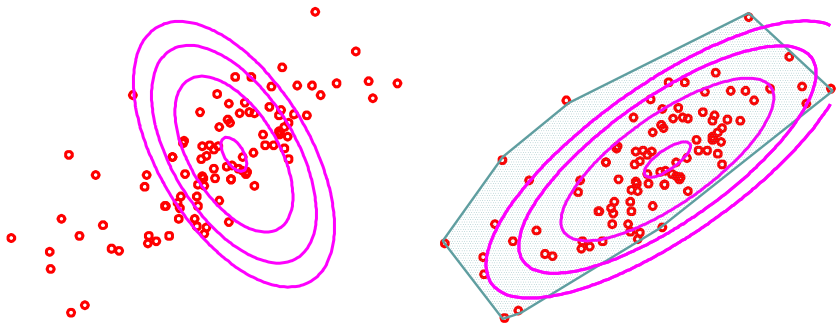
$$\min_{\boldsymbol{w}, b, \rho, \boldsymbol{\xi}} -2\rho + \frac{1}{m} \sum_{i \in M_p} \xi_i^2 \;\; \textit{subject to} \;\; \rho - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \leq \xi_i,\, 0 \leq \xi_i, \forall i,\; \|\boldsymbol{w}\|^2 \leq \left(\frac{m^2\lambda}{4k}\right)^2.$$

Example 5 *We define* $h_o^* : \mathcal{X} \to \mathbb{R}$ *for* $o \in \{p,n\}$ *by*

$$h_o^*(\boldsymbol{z}) = (\boldsymbol{z} - \boldsymbol{\mu}_o)^T C_o (\boldsymbol{z} - \boldsymbol{\mu}_o)$$

where $\boldsymbol{\mu}_o$ *is the mean vector of the input vector* $\boldsymbol{x}$ *conditioned on each label and* C_o *is a positive definite matrix. In practice, the mean vector is estimated by using prior knowledge which is independent of training samples* $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, m\}$. *Suppose that* $\boldsymbol{\mu}_o \neq \mathbf{0}$. *Accordingly, for* $\alpha \geq 0$, *the revision of* (20) *leads to*

$$\begin{aligned} \bar{\ell}^*(\alpha) &= \left((\alpha\frac{m_p}{m} - 1)^2 - 1\right) \boldsymbol{\mu}_p^T C_p \boldsymbol{\mu}_p + \left((\alpha\frac{m_n}{m} - 1)^2 - 1\right) \boldsymbol{\mu}_n^T C_n \boldsymbol{\mu}_n \\ &= b_1 \alpha + b_2 \alpha^2, \end{aligned}$$



original uncertainty set $\mathcal{U}_p[c]$ revised uncertainty set $\bar{\mathcal{U}}_p[c]$

Figure 3: Training samples and uncertainty sets. Left panel: original uncertainty set for the positive label. Right panel: revised uncertainty set consisting of the intersection of ellipsoid and convex-hull of input vectors with the positive label.

where b_1 and $b_2(>0)$ are constant numbers. Thus, we have

$$\begin{aligned}\bar{\mathcal{U}}_o[c] &= \left\{ \sum_{i\in M_o} \alpha_i \boldsymbol{x}_i \,:\, \sum_{i\in M_o} \alpha_i = 1, \alpha_i \geq 0\,(i\in M_o), \sum_{i\in M_o} \alpha_i^2 \leq \frac{c-b_1}{mb_2} \right\} \\ &= \left\{ \boldsymbol{z} \in \mathrm{conv}\{\boldsymbol{x}_i : i\in M_o\} \,:\, (\boldsymbol{z}-\bar{\boldsymbol{x}}_o)^T \widehat{\Sigma}_o^{-1} (\boldsymbol{z}-\bar{\boldsymbol{x}}_o) \leq m_o \cdot \frac{c-b_1}{mb_2} \right\},\end{aligned}$$

where $\bar{\boldsymbol{x}}_o$ and $\widehat{\Sigma}_o$ are the estimators of the mean vector and the covariance matrix for $\{\boldsymbol{x}_i : i\in M_o\}$. The corresponding loss function is obtained in the same way as Example 4. Figure 3 illustrates an example of the revision of the uncertainty set. In the left panel, the uncertainty set does not match the distribution of the training samples. On the other hand, the revised uncertainty set in the right panel well approximates the dispersal of the training samples.

Example 6 *Suppose that for $o\in\{p,n\}$, $\boldsymbol{\mu}_o$ is the mean vector and Σ_o is the covariance matrix of the input vector conditioned on each label. We define the uncertainty set by*

$$o\in\{p,n\}, \quad \mathcal{U}_o[c] = \left\{ z \in \mathrm{conv}\{\boldsymbol{x}_i : i\in M_o\} \,:\, (\boldsymbol{z}-\boldsymbol{\mu})^T \Sigma_o^{-1} (\boldsymbol{z}-\boldsymbol{\mu}) \leq c, \forall \boldsymbol{\mu} \in \mathcal{A} \right\},$$

where $\mathcal{A}$ denotes the estimation error of the mean vector $\boldsymbol{\mu}$. For a fixed radius $r>0$, $\mathcal{A}$ is defined as

$$\mathcal{A} = \left\{ \boldsymbol{\mu} \in \mathcal{X} \,:\, (\boldsymbol{\mu}-\boldsymbol{\mu}_o)^T \Sigma_o^{-1} (\boldsymbol{\mu}-\boldsymbol{\mu}_o) \leq r^2 \right\}.$$

The uncertainty set with the estimation error is used by Lanckriet et al. (2003) in MPM. The above uncertainty set is useful when the probability in the training phase is slightly different from that in the test phase. A brief calculation yields a representation of $\mathcal{U}_o[c]$ in terms of the level set of the convex function,

$$h_o^*(\boldsymbol{z}) = \max_{\boldsymbol{\mu}\in\mathcal{A}} (\boldsymbol{z}-\boldsymbol{\mu})^T \Sigma_o^{-1} (\boldsymbol{z}-\boldsymbol{\mu}) = \left(\sqrt{(\boldsymbol{z}-\boldsymbol{\mu}_o)^T \Sigma_o^{-1} (\boldsymbol{z}-\boldsymbol{\mu}_o)} + r \right)^2.$$

The revised uncertainty set $\bar{\mathcal{U}}_o[c]$ is defined by the function $\bar{\ell}^*$:

$$\bar{\ell}^*(\alpha) = \left(\left|\alpha\frac{m_p}{m} - 1\right|\sqrt{\boldsymbol{\mu}_p^T\Sigma_p^{-1}\boldsymbol{\mu}_p} + r\right)^2 - \left(\sqrt{\boldsymbol{\mu}_p^T\Sigma_p^{-1}\boldsymbol{\mu}_p} + r\right)^2 + \left(\left|\alpha\frac{m_n}{m} - 1\right|\sqrt{\boldsymbol{\mu}_n^T\Sigma_n^{-1}\boldsymbol{\mu}_n} + r\right)^2 - \left(\sqrt{\boldsymbol{\mu}_n^T\Sigma_n^{-1}\boldsymbol{\mu}_n} + r\right)^2. \tag{23}$$

Suppose that $\boldsymbol{\mu}_p \neq \mathbf{0}$ and $\boldsymbol{\mu}_n = \mathbf{0}$ hold. Let $d = \sqrt{\boldsymbol{\mu}_p^T\Sigma_p^{-1}\boldsymbol{\mu}_p}$ and $h = r/d(>0)$. The corresponding loss function is

$$\bar{\ell}(z) = \frac{md^2}{m_p}u\big(\frac{z}{d^2}\big),$$

where $u(z)$ as defined as

$$u(z) = \begin{cases} 0, & z \leq -2h-2, \\ \big(\frac{z}{2}+1+h\big)^2, & -2h-2 \leq z \leq -2h, \\ z+2h+1, & -2h \leq z \leq 2h, \\ \frac{z^2}{4} + z(1-h) + (1+h)^2, & 2h \leq z. \end{cases} \tag{24}$$

Figure 4 depicts the function $u(z)$ with $h = 1$. When $r = 0$ holds, $\bar{\ell}(z)$ reduces to the truncated quadratic function shown in Example 4 and 5. For positive r, $\bar{\ell}(z)$ is linear around $z = 0$. This implies that by introducing the confidence set of the mean vector $\mathcal{A}$, the penalty for the misclassification reduces from quadratic to linear around the decision boundary, though the original uncertainty set $\mathcal{U}_o[c]$ does not correspond to minimization of an empirical loss function.

5. Kernel-Based Learning Algorithm Derived from Uncertainty Set

Suppose that we have training samples $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{+1, -1\}$, where $\mathcal{X}$ is not necessarily a linear space. Let us define a kernel function $k : \mathcal{X}^2 \to \mathbb{R}$, and let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) endowed with the kernel function k; see Schölkopf and Smola (2002) for details about the kernel estimators in machine learning.

Let us start with the parametrized uncertainty sets $\mathcal{U}_p[c]$ and $\mathcal{U}_n[c]$ in $\mathcal{H}$. Given uncertainty sets, a kernel variant of (16) is expressed as

$$\begin{aligned} \inf_{c_p, c_n, f_p, f_n} \; & c_p + c_n + \lambda\|f_p - f_n\|_{\mathcal{H}} \\ \text{subject to } \; & c_p, c_n \in \mathbb{R}, \\ & f_p \in \mathcal{U}_p[c_p] \cap \operatorname{conv}\{k(\cdot, x_i) : i \in M_p\}, \\ & f_n \in \mathcal{U}_n[c_n] \cap \operatorname{conv}\{k(\cdot, x_j) : j \in M_n\}. \end{aligned} \tag{25}$$

Next, we find the corresponding loss function $\ell(z)$. Note that the revision of uncertainty sets presented in Section 4 can be used, if necessary. Suppose the uncertainty sets are represented as

$$\mathcal{U}_o[c] = \left\{ \sum_{i \in M_o} \alpha_i k(\cdot, x_i) \in \mathcal{H} : \frac{1}{m}\sum_{i \in M_o} \ell^*(m\alpha_i) \leq c \right\} \tag{26}$$

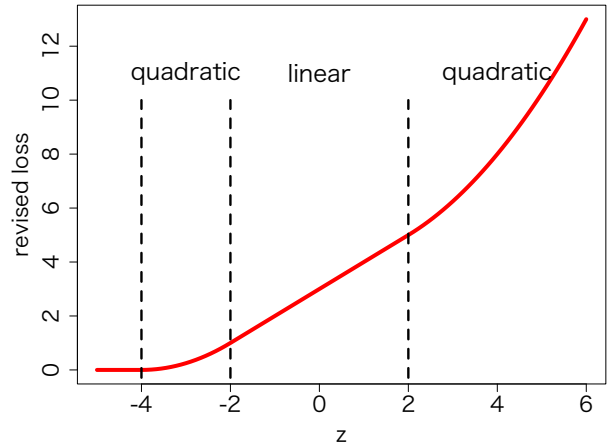


Figure 4: Loss function $u(z)$ in Example 6 that corresponds to the revised uncertainty set with the estimation error.

for $o \in \{p,n\}$. In the same way as in Section 3.1, we find that problem (25) is the dual representation of

$$\begin{aligned} &\min_{f,b,\rho} -2\rho + \frac{1}{m}\sum_{i=1}^{m} \ell(\rho - y_i(f(x_i)+b)) \\ &\text{subject to } \ f \in \mathcal{H}, b \in \mathbb{R}, \rho \in \mathbb{R}, \|f\|_{\mathcal{H}}^2 \le \lambda^2. \end{aligned} \tag{27}$$

We can obtain the estimated decision function $\widehat{f}+\widehat{b} \in \mathcal{H}+\mathbb{R}$ by solving the problem (27). A rigorous proof of the strong duality between (25) and (27) is presented in Section 6 and Appendix B.

Example 7 (ellipsoidal uncertainty sets in RKHS) *Let us consider an uncertainty set $\mathcal{U}[c]$ in RKHS $\mathcal{H}$ defined by*

$$\mathcal{U}[c] = \left\{ \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \,:\, \sum_{i=1}^{m} \alpha_i^2 \le c \right\} \subset \mathcal{H},$$

where $x_1,\ldots,x_m$ are points in $\mathcal{X}$. The corresponding loss is the truncated quadratic loss. Let us define $\bar{k} \in \mathcal{H}$ as $\frac{1}{m}\sum_{i=1}^{m} k(\cdot,x_i)$. Furthermore, let us define the empirical variance operator $\widehat{\Sigma} : \mathcal{H} \to \mathcal{H}$ as

$$\widehat{\Sigma} h = \frac{1}{m}\sum_{i=1}^{m} (k(\cdot,x_i) - \bar{k}) \langle k(\cdot,x_i) - \bar{k}, h \rangle_{\mathcal{H}}$$

for $h \in \mathcal{H}$. Some calculation yields

$$\begin{aligned} &\mathcal{U}[c] \cap \mathrm{conv}\{k(\cdot,x_i) \,:\, i = 1,\ldots,m\} \\ &= \left\{ \bar{k} + \widehat{\Sigma} h \,:\, \langle \widehat{\Sigma} h, h \rangle_{\mathcal{H}} \le mc - 1 \right\} \cap \mathrm{conv}\{k(\cdot,x_i) \,:\, i = 1,\ldots,m\}. \end{aligned}$$

This is the kernel variant of the ellipsoidal uncertainty set in Example 2.

By transforming the uncertainty-set-based learning into loss-based learning, we can obtain a statistical model for the conditional probability, as shown in Section 3.2. In addition, we can verify the statistical consistency of the learning algorithm with the (revised) uncertainty sets by taking the corresponding loss function into account. Other authors have proposed kernel-based learning algorithms with uncertainty sets (Lanckriet et al., 2003; Huang et al., 2004), but they did not deal with the issue of statistical consistency. In the following, we study the statistical properties of the learning algorithm based on (27).

6. Statistical Properties of Kernel-Based Learning Algorithms

Here, we prove that the expected 0-1 loss $\mathcal{E}(\widehat{f}+\widehat{b})$ converges to the Bayes risk $\mathcal{E}^*$ defined by (1). We also determine whether certain popular uncertainty sets produce consistent learning methods. All proofs are presented in Appendix B and Appendix C.

6.1 Assumptions for Statistical Consistency

Let us show four assumptions.

Assumption 1 (universal kernel) *The input space $\mathcal{X}$ is a compact metric space. The kernel function $k : \mathcal{X}^2 \to \mathbb{R}$ is continuous, and satisfies*

$$\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \leq K < \infty,$$

where K is a positive constant. In addition, k is universal, that is, the RKHS associated with k is dense in the set of all continuous functions on $\mathcal{X}$ with respect to the supremum norm (Steinwart and Christmann, 2008, Definition 4.52).

Assumption 2 (non-deterministic assumption) *For the probability distribution of training samples, there exists a positive constant $\varepsilon > 0$ such that*

$$P(\{x \in \mathcal{X} : \varepsilon \leq P(+1|x) \leq 1-\varepsilon\}) > 0,$$

where $P(y|x)$ is the conditional probability of the label y for the input x.

Assumption 3 (basic assumptions on loss functions) *The loss function $\ell : \mathbb{R} \to \mathbb{R}$ satisfies the following conditions.*

1. *ℓ is a non-decreasing, convex function that is non-negative, that is, $\ell(z) \geq 0$ for all $z \in \mathbb{R}$.*
2. *Let $\partial\ell(z)$ be the subdifferential of the loss function ℓ at $z \in \mathbb{R}$ (Rockafellar, 1970, Chapter 23). Then, the equality $\lim_{z\to\infty} \partial\ell(z) = \infty$ holds; that is, for any $M > 0$, there exists z_0 such that $g \geq M$ for all $z \geq z_0$ and all $g \in \partial\ell(z)$.*

Note that the second condition in Assumption 3 assures that ℓ is not a constant function and that $\lim_{z\to\infty} \ell(z) = \infty$.

Assumption 4 (modified classification-calibrated loss)

1. *$\ell(z)$ is first order differentiable for $z \geq -\ell(0)/2$, and $\ell'(z) > 0$ for $z \geq -\ell(0)/2$, where ℓ' is the derivative of ℓ.*

2. *Let $\psi(\theta,\rho)$ be the function*

$$\psi(\theta,\rho) = \ell(\rho) - \inf_{z\in\mathbb{R}} \left\{ \frac{1+\theta}{2}\ell(\rho - z) + \frac{1-\theta}{2}\ell(\rho + z) \right\}, \quad 0 \leq \theta \leq 1,\ \rho \in \mathbb{R}.$$

There exists a function $\widetilde{\psi}(\theta)$ and a positive real $\varepsilon > 0$ such that the following three conditions are satisfied:

   (a) $\widetilde{\psi}(0) = 0$ and $\widetilde{\psi}(\theta) > 0$ for $0 < \theta \leq \varepsilon$.

   (b) $\widetilde{\psi}(\theta)$ is a continuous and strictly increasing function on the interval $[0,\varepsilon]$.

   (c) The inequality $\widetilde{\psi}(\theta) \leq \inf_{\rho \geq -\ell(0)/2} \psi(\theta,\rho)$ holds for $0 \leq \theta \leq \varepsilon$.

Appendix B presents a rigorous proof of the duality between (27) and (25) with the uncertainty set (26). Appendix C.3 presents sufficient conditions for the existence of the function $\widetilde{\psi}$ in Assumption 4.

Under Assumptions 1–4 and another mild assumption, we prove that the expected 0-1 loss $\mathcal{E}(\widehat{f} + \widehat{b})$ converges to the Bayes risk $\mathcal{E}^*$. In the mild assumption, the covering number of the RKHS $\mathcal{H}$ is taken into account. The details of the conferring number are shown in Appendix C.1.

6.1.1 THEOREM (STATISTICAL CONSISTENCY)

For the RKHS $\mathcal{H}$ and the loss function ℓ, we assume Assumptions 1, 2, 3 and 4. Also, we assume that $\mathcal{H}$ satisfies the covering number condition, that is, (41) *in Appendix C.1 converges to zero for any positive ε, when the sample size m tends to infinity. Then, $\mathcal{E}(\widehat{f} + \widehat{b})$ converges to $\mathcal{E}^*$ in probability.*

Appendix C presents the necessary definitions, lemmas, and theorems, and Theorem 8 of Appendix C.1 and Theorem 9 of Appendix C.2 summarize the main results. The examples presented in Appendix C.3 show that some popular uncertainty sets and their revisions yield loss functions satisfying the above sufficient conditions.

6.2 Supplementary Explanations

Let us discuss Assumptions 1–4.

Universal kernel: The universality of RKHSs in Assumption 1 is usually assumed, when discussing the statistical consistency of kernel methods. If the RKHS under consideration is not universal, a decision function might exist that is not approximated well by any element in the RKHS. The Gaussian kernel is universal, while the polynomial kernel is not universal.

Non-deterministic assumption: In Assumption 2, the label y is assigned in a non-deterministic way. The label assignment is deterministic when the conditional probability $P(y = +1|x)$ is equal to 0 or 1 for all x. Steinwart (2005) introduced the y-degenerated condition defined as

$$P(\{x \in \mathcal{X} : P(y|x) = 1\}) = 1$$

for a label $y \in \{+1, -1\}$. If the y-degenerated condition holds, the proof of the consistency is straightforward for standard learning methods such as C-SVM. Involved mathematical arguments are needed to prove consistency if the y-degenerated condition does not apply. Here, the deterministic assumption means

$$P(\{x \in \mathcal{X} : P(y|x) = 1 \text{ or } 0\}) = 1.$$

In our setup, the parameter ρ is a variable, and it makes the situation somewhat difficult. Under the above deterministic assumption, the optimal value of (27) may go to $-\infty$, as the number of training samples tends to infinity; see Lemma 3 in Appendix C.1. In such a case, it would be impossible to make an empirical approximation of the objective value in (27). We introduced the non-deterministic assumption to avoid such a troublesome situation.

Basic assumptions on loss functions: Loss functions are based on Assumption 3 and Assumption 4. Conditions such that $\ell(z)$ is convex, non-decreasing, and bounded from below are standard ones, but the second condition in Assumption 3 is rather strong. The hinge loss and logistic loss do not satisfy this assumption, whereas the quadratic loss and exponential loss satisfy it. Assumption 3 is used to derive an upper bound of the optimal ρ in (27); see Lemma 5 in Appendix C.1.

Modified classification-calibrated loss: Assumption 4 is related to the classification-calibrated loss. Bartlett et al. (2006) introduced classification-calibrated losses to analyze the statistical consistency of binary classification problems. Roughly speaking, if a loss function is classification-calibrated, the minimizer of the loss function produces the minimizer of the 0-1 loss. See Bartlett et al. (2006) for details about classification-calibrated losses. Suppose that the function $\ell(\rho - z)$ with a fixed ρ is convex in $z \in \mathbb{R}$. Then, a sufficient condition for $\ell(\rho - z)$ to be a classification-calibrated loss is given as $\ell'(\rho) > 0$; that is, ℓ is differentiable at ρ and the derivative is positive. In our setup, ρ is variable, and hence the condition $\ell'(\rho) > 0$ is required for all possible values of ρ. As shown in the proof of Lemma 5 in Appendix C.1, the optimal ρ of the problem (27) is bounded from below by $-\ell(0)/2$. Thus, we assumed the differentiability of $\ell(z)$ for $z \geq -\ell(0)/2$. The second condition of Assumption 4 defines the functions, $\psi(\theta, \rho)$ and $\widetilde{\psi}(\theta)$. Bartlett et al. (2006) defined the function $\psi(\theta, 0)$ and derived the quantitative relation between the classification calibrated loss and the 0-1 loss via $\psi(\theta, 0)$. We extended $\psi(\theta, 0)$ to $\psi(\theta, \rho)$ having a variable ρ. The functions $\psi(\theta, \rho)$ and $\widetilde{\psi}(\theta)$ describe a qualitative relation between the convex loss ℓ and the 0-1 loss. Appendix C.2 uses the function $\widetilde{\psi}(\theta)$ to prove that the convergence of the expected loss guarantees the convergence of the expected 0-1 loss to the Bayes risk.

Appendix C.3 describes the sufficient conditions for the existence of the function $\widetilde{\psi}(\theta)$ in Assumption 4. It shows some simple conditions under which a given loss function $\ell(z)$ will possess $\widetilde{\psi}(\theta)$. As a result, it is shown that the existence of $\widetilde{\psi}$ is guaranteed for the truncated quadratic loss, exponential loss and the loss function derived from the uncertainty set with the estimation error in Example 6; see the examples provide in Appendix C.3.

7. Experiments

We conducted some numerical experiments to examine the prediction performance of our revision of uncertainty sets methods. The results indicate that the method improves the estimator. In addition, we evaluated the estimation accuracy of the conditional probability.

We compared the kernel-based learning algorithms using the Gaussian kernel. So far, many studies have compared linear models and kernel-based models. The conclusion is that linear models outperform kernel-based models when the linear models have good approximations of the decision boundary. Otherwise, linear models have an approximation bias, and kernel-based estimators with a nice regularization outperform linear models. For this reason, we focused on kernel-based estimators.

The following methods were examined using the synthetic data and the standard benchmark data sets: C-SVM, MPM, unbiased MPM, and the learning method with (27). C-SVM is the one implemented in the `kernlab` library (Karatzoglou et al., 2004). In the unbiased MPM, the bias term b of the model was estimated by minimizing the training error rate after estimating the function part, $\widehat{f} \in \mathcal{H}$. The unbiased estimator will outperform the original MPM when the probability of the class label is heavily unbalanced. The loss function $\ell(z)$ of the proposed method was the function $u(z)$ in (24). This loss function corresponds to the revised uncertainty set of the ellipsoidal uncertainty set with the estimation error. The parameter in the function $u(z)$ of (24) was set to $h = 0$ or $h = 1$. The kernel parameter and the regularization parameter were estimated by 5-fold cross validation.

We evaluated the learning results as follows. We used the test error over the test samples to evaluate the classification accuracy. We assessed the estimation accuracies of the conditional probabilities given by C-SVM and the proposed method. A C-SVM with such a probability estimation is included in the `kernlab` library; the probability model is shown in Karatzoglou et al. (2004). We used the squared loss to assess the estimation accuracy of the conditional probability:

$$\mathbb{E}\Big[\sum_{y=\pm 1}(\widehat{P}(y|x)-P(y|x))^2\Big] = \mathbb{E}\big[\widehat{P}(+1|x)^2+\widehat{P}(-1|x)^2\big] - 2\mathbb{E}\big[\widehat{P}(y|x)\big] + \mathbb{E}\big[P(+1|x)^2+P(-1|x)^2\big],$$

where $\widehat{P}(y|x)$ is an estimator of the true conditional probability $P(y|x)$. Since the last term of the above expression does not depend on the estimator, we used only the first two terms as the measure of estimation accuracy. As a result, given test samples $\{(\widetilde{x}_i,\widetilde{y}_i) : \ell = 1,\ldots,L\}$, the estimated conditional probability $\widehat{P}(y|x)$ can be approximately evaluated as follows:

$$\text{squared-loss} = \frac{1}{L}\sum_{\ell=1}^{L}\big[\widehat{P}(+1|\widetilde{x}_\ell)^2+\widehat{P}(-1|\widetilde{x}_\ell)^2\big] - \frac{2}{L}\sum_{\ell=1}^{L}\widehat{P}(\widetilde{y}_\ell|\widetilde{x}_\ell).$$

This measure works even for benchmark data sets in which the true probability is unknown. Note that the squared-loss above can take negative values, since the last term in the expansion of $\mathbb{E}\big[\sum_{y=\pm 1}(\widehat{P}(y|x)-P(y|x))^2\big]$ is not taken into account. We did not use the Kullback-Leibler divergence or logarithmic loss, since the estimator $\widehat{P}(y|x)$ can take zero.

7.1 Synthetic Data

The input points conditioned on the positive label were generated from a two dimensional normal distribution with mean $\boldsymbol{\mu}_p = (0,0)^T$ and variance-covariance matrix $\Sigma_p = I$, where I is the identity

$P(y=+1)$	C-SVM	MPM	unbiased MPM	$h=0$	$h=1$
0.2	15.81 ± 1.17	25.63 ± 2.27	16.51 ± 1.48	15.39 ± 0.99	15.37 ± 0.93
0.5	25.32 ± 1.49	25.33 ± 1.64	25.50 ± 1.47	24.81 ± 1.18	24.89 ± 1.27

Table 1: Test error (%) and standard deviation of each learning method. We compared C-SVM, MPM, unbiased MPM, and the proposed learning method using the loss function (24) with $h=0$ or $h=1$.

matrix. The conditional distribution of the input points with the negative label was a normal distribution with mean $\boldsymbol{\mu}_n = (1,1)^T$ and variance-covariance matrix $\Sigma_n = R^T \mathrm{diag}(0.5^2, 1.5^2) R$, where R is the $\pi/3$ radian counterclockwise rotation matrix. The label probability was $P(y=+1) = 0.2$ or 0.5. The size of the training samples was $m = 400$. We computed test errors by averaging over 100 iterations. For C-SVM and the proposed method, we computed the average squared-loss of the estimated conditional probability. We also evaluated average absolute difference between the true conditional probability $P(+1|\boldsymbol{x})$ and the estimator $\widehat{P}(+1|\boldsymbol{x})$ on the test set, that is, the average of $\frac{1}{L}\sum_{\ell=1}^{L} |P(+1|\widetilde{\boldsymbol{x}}_\ell) - \widehat{P}(+1|\widetilde{\boldsymbol{x}}_\ell)|$ over 100 iterations. This is possible, since the true probability of synthetic data is known.

Table 1 shows the test errors of C-SVM, MPM, unbiased MPM, and the proposed method using the loss function (24) with $h=0$ or $h=1$. The table shows that the MPM has an estimation bias for unbalanced samples, that is, the case of $P(y=+1) = 0.2$. MPM is slightly better than unbiased MPM on the setup of the balanced data. Overall, the proposed method is better than the other learning methods. Indeed, the difference between it and C-SVM is statistically significant. On the other hand, the parameter h in the loss function (24) does not significantly affect the experimental results.

Table 2 shows the accuracy of the estimated conditional probabilities measured by the squared loss and absolute difference. As shown in the lower table, the absolute error of the proposed method is about 5%, while the error of C-SVM is about 10%. The proposed method also outperforms C-SVM in terms of the squared-loss. C-SVM and the proposed method differ significantly in their estimation accuracy of the conditional probability, though the difference in classification error rate is less than 0.5%. Figure 5 presents the squared loss and absolute loss of the estimated conditional probability versus the size of the training samples for C-SVM and the proposed method with $h=0$ and $h=1$. The proposed method outperforms C-SVM. For each sample size, the parameter h does not significantly affect the estimation accuracy, though the loss function $u(z)$ with $h=1$ is consistently slightly better than $h=0$.

7.2 Benchmark Data

The experiments used thirteen artificial and real-world data sets from the UCI, DELVE, and STATLOG benchmark repositories: `banana`, `breast-cancer`, `diabetes`, `german`, `heart`, `image`, `ringnorm`, `flare-solar`, `splice`, `thyroid`, `titanic`, `twonorm`, and `waveform`. All data sets are in the IDA benchmark repository. See Rätsch et al. (2001) and Rätsch et al. (2000) for details about the data sets. The properties of each data set are shown in Table 3, where "dim", "$P(y=+1)$","#train", "#test" and "rep." respectively denote the input dimension, the ratio of the positive labels in training samples, the size of training set, the size of test set, and the number of

	squared-loss $\times 100$		
$P(y=+1)$	C-SVM	$h=0$	$h=1$
0.2	-75.43 ± 1.78	-77.36 ± 1.07	-77.42 ± 1.08
0.5	-65.63 ± 1.89	-67.90 ± 1.07	-67.83 ± 1.21

	absolute difference (%) between $P(+1\|\boldsymbol{x})$ and $\widehat{P}(+1\|\boldsymbol{x})$		
$P(y=+1)$	C-SVM	$h=0$	$h=1$
0.2	9.37 ± 1.96	4.57 ± 1.28	4.43 ± 1.14
0.5	10.10 ± 2.34	5.11 ± 1.10	5.19 ± 1.32

Table 2: Squared loss and absolute loss of the estimated conditional probability $\widehat{P}(y|x)$. We compared C-SVM and the proposed method using the loss function (24) with $h=0$ and $h=1$.

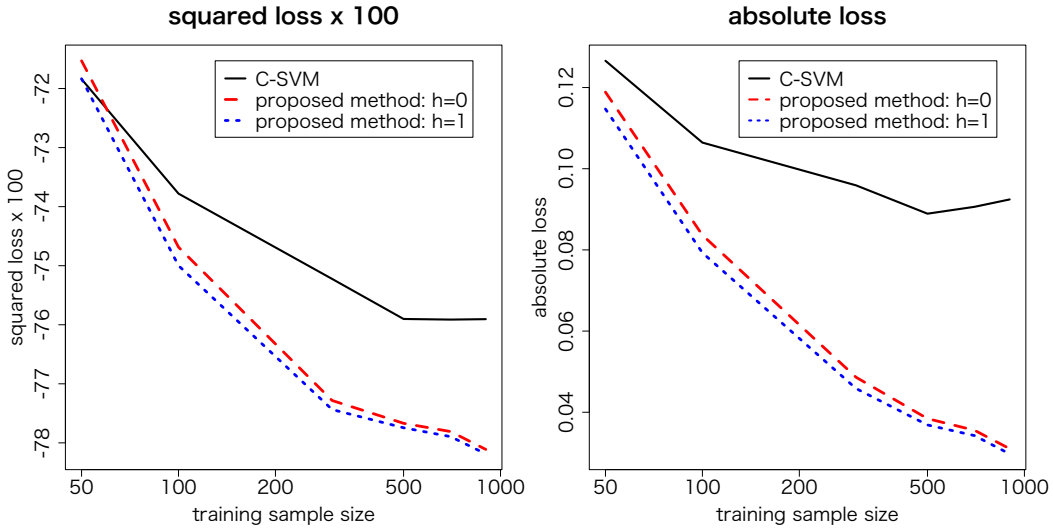


Figure 5: Squared loss and absolute loss of estimated conditional probability versus training sample size are presented for C-SVM and the proposed method with $h=0$ and $h=1$.

replications of learning to evaluate the average performance. Table 4 shows the test errors (%) and the standard deviation for the benchmark data sets.

First, we compared MPM, unbiased MPM and the proposed method with "$h=0$". The uncertainty set of MPM and unbiased MPM is an ellipsoid defined by the estimated covariance matrix. The corresponding loss function of the form of (9) does not exist, since the convex-hull of the input points is not taken into account. The uncertainty set of the proposed method with "$h=0$" is the intersection of an ellipsoid and the convex-hull of the input vectors. The revision of the ellipsoidal uncertainty set leads to the uncertainty set of our algorithm. The proposed method with "$h=0$" outperforms MPM and unbiased MPM for most data sets. Hence, the revision of uncertainty sets can improve the prediction accuracy of uncertainty-set-based learning.

data set	dim	$P(y=+1)$	#train	#test	rep.
banana	2	0.454	400	4900	100
breast-cancer	9	0.294	200	77	100
diabetis	8	0.350	468	300	100
flare-solar	9	0.552	666	400	100
german	20	0.301	700	300	100
heart	13	0.445	170	100	100
image	18	0.574	1300	1010	20
ringnorm	20	0.497	400	7000	100
splice	60	0.483	1000	2175	20
thyroid	5	0.305	140	75	80
titanic	3	0.322	150	2051	100
twonorm	20	0.505	400	7000	100
waveform	21	0.331	400	4600	100

Table 3: The properties of each data sets: "dim", "$P(y=+1)$","#train", "#test" and "rep." respectively denote the input dimension, the ratio of the positive label in the training samples, the size of the training set, the size of the test set, and the number of replications of learning.

The boldface letters in Table 4 indicate the smallest average test error for each data set. Overall, C-SVM and the learning method "$h=1$" outperform the others. C-SVM is significantly better than the proposed method with "$h=1$" on `flare-solar`, `ringnorm` and `twonorm`, but the proposed method with "$h=1$" is significantly better than C-SVM on `banana`, `diabetis`, `german` and `waveform`. These results show that the proposed method with "$h=1$" is comparable to C-SVM. Table 5 shows the squared-losses for estimated conditional probabilities. It shows that the proposed method with "$h=1$" outperforms the others in the conditional probability estimation.

In Section 6, we proved the statistical consistency of learning methods derived from the uncertainty set approach. The numerical experiments described in this section indicate that learning methods derived from revised uncertainty sets are an alternative for solving classification problems involving conditional probability estimations.

8. Conclusion

We studied the relation between the loss function approach and the uncertainty set approach in binary classification problems. We showed that these two approaches are connected via the convex conjugate of the loss function. Given a loss function, there exists a corresponding parametrized uncertainty set. In general, however, the uncertainty set does not correspond to the empirical loss function. We presented a way of revising the uncertainty set so that it will correspond to an empirical loss function. On the basis of this revision, we proposed a kernel-based learning algorithm and proved statistical consistency. The way to estimate the conditional probability was also proposed. Numerical experiments showed that learning methods derived from revised uncertainty sets are alternatives means for solving classification problems involving conditional probability estimation.

Some problems remains with our methodology. The proof of the statistical consistency does not include the hinge loss used in ν-SVM. Steinwart (2003) proved that ν-SVM is statistically consistent with a nice choice of the regularization parameter. However, such a regularization parameter heavily depends on the true probability distribution; that is, the parameter ν should be twice the Bayes error

	test error (%)				
data set	C-SVM	MPM	unbiased MPM	$h=0$	$h=1$
banana	10.74 ± 0.60	11.35 ± 0.87	11.49 ± 0.93	10.47 ± 0.48	$\mathbf{10.45 \pm 0.47}$
breast-cancer	26.96 ± 4.57	34.77 ± 4.53	33.26 ± 5.01	$\mathbf{26.60 \pm 4.64}$	26.77 ± 4.60
diabetis	23.94 ± 2.05	28.81 ± 2.61	28.42 ± 2.46	$\mathbf{23.30 \pm 1.85}$	23.36 ± 1.78
flare-solar	$\mathbf{33.71 \pm 2.11}$	34.92 ± 1.73	35.62 ± 1.83	34.09 ± 1.65	34.14 ± 1.86
german	23.84 ± 2.32	29.17 ± 2.43	28.53 ± 2.58	23.54 ± 2.20	$\mathbf{23.28 \pm 2.08}$
heart	$\mathbf{16.56 \pm 3.51}$	25.41 ± 4.34	25.82 ± 4.17	16.62 ± 3.48	16.70 ± 3.17
image	3.17 ± 0.66	$\mathbf{3.11 \pm 0.58}$	3.30 ± 0.73	3.20 ± 0.67	3.21 ± 0.62
ringnorm	$\mathbf{1.73 \pm 0.27}$	3.21 ± 0.49	2.81 ± 0.38	2.02 ± 0.25	2.01 ± 0.24
splice	$\mathbf{11.03 \pm 0.74}$	12.25 ± 1.71	11.74 ± 0.89	11.10 ± 0.72	11.07 ± 0.64
thyroid	5.25 ± 2.10	6.58 ± 2.96	6.83 ± 3.23	5.27 ± 2.16	$\mathbf{5.08 \pm 2.23}$
titanic	$\mathbf{22.47 \pm 0.81}$	24.27 ± 2.60	22.47 ± 1.23	22.59 ± 1.37	22.62 ± 1.36
twonorm	$\mathbf{2.67 \pm 0.41}$	4.50 ± 0.65	4.47 ± 0.66	2.98 ± 0.31	2.97 ± 0.30
waveform	10.22 ± 0.68	12.90 ± 0.79	12.73 ± 0.94	10.00 ± 0.50	$\mathbf{9.96 \pm 0.44}$

Table 4: Test errors (%) and the standard deviation for benchmark data sets. We compared C-SVM, MPM, unbiased MPM, and the proposed method with loss functions (24) having $h = 0$ and $h = 1$: boldface letters indicate that the average squared loss is the smallest.

	squared loss $\times 100$		
data set	C-SVM	$h=0$	$h=1$
banana	-83.99 ± 0.90	-84.98 ± 0.53	$\mathbf{-85.15 \pm 0.50}$
breast-cancer	-62.60 ± 4.15	-64.04 ± 3.88	$\mathbf{-64.11 \pm 4.09}$
diabetis	-67.51 ± 2.01	$\mathbf{-68.31 \pm 1.39}$	-68.13 ± 1.52
flare-solar	-56.52 ± 1.67	-59.69 ± 1.10	$\mathbf{-59.79 \pm 1.15}$
german	-67.29 ± 2.20	-67.69 ± 1.99	$\mathbf{-67.98 \pm 2.09}$
heart	$\mathbf{-74.75 \pm 4.02}$	-74.14 ± 3.23	-74.59 ± 3.37
image	-94.71 ± 0.81	-94.67 ± 0.68	$\mathbf{-94.72 \pm 0.72}$
ringnorm	$\mathbf{-97.30 \pm 0.75}$	-96.27 ± 0.27	-96.27 ± 0.29
splice	$\mathbf{-83.88 \pm 0.75}$	-83.30 ± 0.57	-83.32 ± 0.54
thyroid	-92.26 ± 3.12	-92.94 ± 2.48	$\mathbf{-93.11 \pm 2.53}$
titanic	-65.06 ± 1.10	$\mathbf{-66.26 \pm 1.28}$	-66.18 ± 1.40
twonorm	$\mathbf{-95.92 \pm 0.60}$	-95.07 ± 0.36	-95.11 ± 0.35
waveform	-85.50 ± 0.89	-85.39 ± 0.53	$\mathbf{-85.57 \pm 0.48}$

Table 5: Squared loss $\times 100$ of estimated conditional probability for C-SVM and the proposed method with loss functions (24) having $h = 0$ and $h = 1$: boldface letters indicate that the average test error is the smallest.

that cannot be obtained before the learning. We are currently investigating of the possibility of relaxing the assumptions so as to include the hinge loss and other popular loss functions such as the logistic loss. We focused on binary classification problems in this paper. An interesting direction of research is to extend the relation between loss-based learning and uncertainty-set-based learning to more general statistical problems such as ranking problems and multiclass classification problems. The statistical consistency of more general problem setups is an ongoing research topic, and we expect that the duality based on the convex conjugate can be used to devise a new approach to these problems.

The relation between the loss function approach and the uncertainty set approach is a useful tool for statistical modeling. In optimization and control theory, the modeling based on the uncertainty set is frequently applied to the real-world data; the reader may consult the modeling used in robust optimization and related work (Ben-Tal and Nemirovski, 2002). We believe that learning algorithms with revised uncertainty sets can bridge the gap between intuitive statistical modeling and nice statistical properties.

Acknowledgments

The authors are grateful to anonymous reviewers for their helpful comments. TK was partially supported by JSPS KAKENHI Grant Number 24500340. AT was partially supported by Grant-in-Aid for Young Scientists (23710174). TS was partially supported by MEXT Kakenhi 22700289 and the Aihara Project, the FIRST program from JSPS, initiated by CSTP. An earlier and shorter version of this paper appeared in the proceedings of COLT2012, and the present version has benefited from comments from the COLT referees.

Appendix A. Preliminaries on Convex Conjugates

A convex conjugate is a standard tool in convex analysis. The convex conjugate is also referred to as a Legendre transformation. We show a brief introduction of the convex conjugate. See Rockafellar (1970) for details.

Let $\ell : \mathbb{R}^k \to \mathbb{R}$ be a convex function. The *convex conjugate* $\ell^* : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ of the function ℓ is defined by

$$\ell^*(\boldsymbol{\alpha}) = \sup_{\boldsymbol{z} \in \mathbb{R}^k} \boldsymbol{z}^T \boldsymbol{\alpha} - \ell(\boldsymbol{z}).$$

Note that $\ell^*(\boldsymbol{\alpha}) = +\infty$ can occur. Under a mild assumption, the equality $(\ell^*)^* = \ell$ holds.

Suppose that the function $\ell(\boldsymbol{z})$ is decomposed into

$$\ell(\boldsymbol{z}) = \ell_1(\boldsymbol{z}_1) + \ell_2(\boldsymbol{z}_2), \quad \boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2) \in \mathbb{R}^k.$$

Then, the convex conjugate of $\ell(\boldsymbol{z})$ is the sum of convex conjugates of ℓ_1 and ℓ_2. Indeed,

$$\begin{aligned}
\ell^*(\boldsymbol{\alpha}) &= \sup_{\boldsymbol{z}} \boldsymbol{z}^T \boldsymbol{\alpha} - \ell_1(\boldsymbol{z}_1) - \ell_2(\boldsymbol{z}_2) \\
&= \sup_{\boldsymbol{z}_1, \boldsymbol{z}_2} \boldsymbol{z}_1^T \boldsymbol{\alpha}_1 + \boldsymbol{z}_2^T \boldsymbol{\alpha}_2 - \ell_1(\boldsymbol{z}_1) - \ell_2(\boldsymbol{z}_2) \\
&= \ell_1^*(\boldsymbol{\alpha}_1) + \ell_2^*(\boldsymbol{\alpha}_2), \quad \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \in \mathbb{R}^k.
\end{aligned}$$

The formula above is used in Section 4.2.

Appendix B. Proof of Strong Duality between (25) and (27)

We prove that there is no duality gap between (25) and (27).

Lemma 2 *Suppose that both $M_p = \{i : y_i = +1\}$ and $M_n = \{i : y_i = -1\}$ are non-empty, that is, $m_p = |M_p|$ and $m_n = |M_n|$ are positive numbers. Under Assumption 1 and 3 in Section 6, there exists an optimal solution for* (27)*. Moreover, the dual problem of* (27) *yields the problem* (25) *with the uncertainty set* (26)*.*

Proof First, we prove the existence of an optimal solution. According to the standard argument on the kernel estimator, we can restrict the function part f to be the form of

$$f(x) = \sum_{j=1}^{m} \alpha_j k(x, x_j).$$

Then, the problem is reduced to the finite-dimensional problem,

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, b, \rho} -2\rho + \frac{1}{m}\sum_{i=1}^{m} \ell(\rho - y_i(\sum_{j=1}^{m} \alpha_j k(x_i, x_j) + b)) \\ &\text{subject to } \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \le \lambda^2. \end{aligned} \tag{28}$$

Let $\zeta_0(\boldsymbol{\alpha}, b, \rho)$ be the objective function of (28). Let us define $\mathcal{S}$ be the linear subspace in $\mathbb{R}^m$ spanned by the column vectors of the gram matrix $(k(x_i, x_j))_{i,j=1}^m$. We can impose the constraint $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathcal{S}$, since the orthogonal complement of $\mathcal{S}$ does not affect the objective function and the constraint in (28). We see that Assumption 1 and the reproducing property yield the inequality $\|y_i \sum_{j=1}^m \alpha_j k(\cdot, x_j)\|_\infty \le K\lambda$. Due to this inequality and the assumptions on the function ℓ, the objective function $\zeta_0(\boldsymbol{\alpha}, b, \rho)$ is bounded below by

$$\zeta_1(b, \rho) = -2\rho + \frac{m_p}{m}\ell(\rho - b - K\lambda) + \frac{m_n}{m}\ell(\rho + b - K\lambda).$$

Hence, for any real number c, the inclusion relation

$$\begin{aligned} &\left\{ (\boldsymbol{\alpha}, b, \rho) \in \mathbb{R}^{m+2} \ : \ \zeta_0(\boldsymbol{\alpha}, b, \rho) \le c, \ \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \le \lambda^2, \boldsymbol{\alpha} \in \mathcal{S} \right\} \\ &\subset \left\{ (\boldsymbol{\alpha}, b, \rho) \in \mathbb{R}^{m+2} \ : \ \zeta_1(b, \rho) \le c, \ \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \le \lambda^2, \boldsymbol{\alpha} \in \mathcal{S} \right\} \end{aligned} \tag{29}$$

holds. Note that any vector $\boldsymbol{\alpha}$ satisfying $\sum_{i,j=1}^m \alpha_i \alpha_j k(x_i, x_j) \le \lambda^2$ and $\boldsymbol{\alpha} \in \mathcal{S}$ is included in a compact subset of $\mathbb{R}^m$. We shall prove that the subset (29) is compact, if they are not empty. We see that the two sets above are closed subsets, since both ζ_0 and ζ_1 are continuous. By the variable change from (b, ρ) to $(u_1, u_2) = (\rho - b, \rho + b)$, $\zeta_1(b, \rho)$ is transformed to a convex function $\zeta_2(u_1, u_2)$ defined by

$$\zeta_2(u_1, u_2) = -u_1 + \frac{m_p}{m}\ell(u_1 - K\lambda) - u_2 + \frac{m_n}{m}\ell(u_2 - K\lambda).$$

The subgradient of $\ell(z)$ diverges to infinity, when z tends to infinity. In addition, $\ell(z)$ is a non-decreasing and non-negative function. Hence we have

$$\lim_{|u_1| \to \infty} -u_1 + \frac{m_p}{m}\ell(u_1 - K\lambda) = \infty.$$

The same limit holds for $-u_2 + \frac{m_n}{m}\ell(u_2 - K\lambda)$. Hence, the level set of $\zeta_2(u_1,u_2)$ is closed and bounded, that is, compact. As a result, the level set of $\zeta_1(b,\rho)$ is also compact. Therefore, the subset (29) is also compact in $\mathbb{R}^{m+2}$. This implies that (28) has an optimal solution.

Next, we prove the duality between (25) and (27). Since (28) has an optimal solution, the problem with the slack variables $\xi_i, i = 1,\ldots,m$,

$$\begin{aligned}
&\min_{\alpha,b,\rho,\xi} -2\rho + \frac{1}{m}\sum_{i=1}^{m}\ell(\xi_i)\\
&\text{subject to } \sum_{i,j=1}^{m}\alpha_i\alpha_j k(x_i,x_j) \le \lambda^2,\\
&\qquad\qquad \rho - y_i\Big(\sum_{j=1}^{m}\alpha_i k(x_i,x_j) + b\Big) \le \xi_i,\ i = 1,\ldots,m,
\end{aligned}$$

also has an optimal solution and the finite optimal value. In addition, the above problem clearly satisfies the Slater condition (Bertsekas et al., 2003, Assumption 6.4.2). Indeed, at a feasible solution, $\alpha = \mathbf{0}, b = 0, \rho = 0$ and $\xi_i = 1, i = 1,\ldots,m$, the constraint inequalities are all inactive for positive λ. Hence, Proposition 6.4.3 in Bertsekas et al. (2003) ensures that the min-max theorem holds, that is, there is no duality gap. Then, in the same way as (10), we obtain (25) with the uncertainty set (26) as the dual problem of (27). ■

Appendix C. Proof of Consistency

We define some notations. For a measurable function $f : \mathcal{X} \to \mathbb{R}$ or $f \in \mathcal{H}$, and a real number $\rho \in \mathbb{R}$, we define the expected loss $\mathcal{R}(f,\rho)$ and the regularized expected loss $\mathcal{R}_\lambda(f,\rho)$ by

$$\begin{aligned}
\mathcal{R}(f,\rho) &= -2\rho + \mathbb{E}[\ell(\rho - yf(x))],\\
\mathcal{R}_\lambda(f,\rho) &= -2\rho + \mathbb{E}[\ell(\rho - yf(x))] + \theta(\|f\|_{\mathcal{H}}^2 \le \lambda^2),
\end{aligned}$$

where λ is a positive number and $\theta(A)$ equals 0 when A is true and ∞ otherwise. Let $\mathcal{R}^*$ be the infimum of $\mathcal{R}(f,\rho)$,

$$\mathcal{R}^* = \inf\{\mathcal{R}(f,\rho) : f \in L_0, \rho \in \mathbb{R}\}.$$

For the set of training samples, $T = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, the empirical loss $\widehat{\mathcal{R}}_T(f,\rho)$ and the regularized empirical loss $\widehat{\mathcal{R}}_{T,\lambda}(f,\rho)$ are defined by

$$\begin{aligned}
\widehat{\mathcal{R}}_T(f,\rho) &= -2\rho + \frac{1}{m}\sum_{i=1}^{m}\ell(\rho - y_i f(x_i)),\\
\widehat{\mathcal{R}}_{T,\lambda}(f,\rho) &= -2\rho + \frac{1}{m}\sum_{i=1}^{m}\ell(\rho - y_i f(x_i)) + \theta(\|f\|_{\mathcal{H}}^2 \le \lambda^2).
\end{aligned}$$

The subscript T is dropped if it is clear from the context. By a slight abuse of notation, for a function $f \in \mathcal{H}$ and a real number b, the regularized expected loss $\mathcal{R}_\lambda(f+b,\rho)$ denotes

$$\mathcal{R}_\lambda(f+b,\rho) = -2\rho + \mathbb{E}[\ell(\rho - y(f(x)+b))] + \theta(\|f\|_{\mathcal{H}}^2 \le \lambda^2).$$

The similar notation $\widehat{\mathcal{R}}_{T,\lambda}(f+b,\rho)$ is also used for the regularized empirical loss, that is,

$$\widehat{\mathcal{R}}_{T,\lambda}(f+b,\rho) = -2\rho + \frac{1}{m}\sum_{i=1}^{m}\ell(\rho - y_i(f(x_i)+b)) + \theta(\|f\|_{\mathcal{H}}^2 \le \lambda^2).$$

For the observed training samples T, clearly the problem (27) is identical to the minimization of $\widehat{\mathcal{R}}_{T,\lambda}(f+b,\rho)$. We define $\widehat{f},\widehat{b}$ and $\widehat{\rho}$ as an optimal solution of

$$\min_{f,b,\rho}\widehat{\mathcal{R}}_{T,\lambda_m}(f+b,\rho), \quad f\in\mathcal{H}, b\in\mathbb{R}, \rho\in\mathbb{R}, \tag{30}$$

where the regularization parameter λ_m may depend on the sample size, m.

In this section, we prove that the error rate $\mathcal{E}(\widehat{f}+\widehat{b})$ converges to the Bayes risk $\mathcal{E}^*$. The proof consists of two parts. In Section C.1, we prove that the expected loss for the estimated decision function, $\mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho})$, converges to the infimum of the expected loss $\mathcal{R}^*$, where $\widehat{f},\widehat{b}$ and $\widehat{\rho}$ are optimal solutions of (30). Here, we apply the mathematical tools developed by Steinwart (2005). In Section C.2, we prove the convergence of the error rate $\mathcal{E}(\widehat{f}+\widehat{b})$ to the Bayes risk $\mathcal{E}^*$. In the proof, the concept of the classification-calibrated loss (Bartlett et al., 2006) plays an important role.

In the following, Assumptions 1–4 are presented in Section 6.

C.1 Convergence to Optimal Expected Loss

In this section, we prove that $\mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho})$ converges to $\mathcal{R}^*$. Following lemmas show the relation between the expected loss and the regularized expected loss.

Lemma 3 *Under Assumption 2 and Assumption 3, we have* $\mathcal{R}^* > -\infty$.

Proof Let $S\subset\mathcal{X}$ be the subset $S=\{x\in\mathcal{X} : \varepsilon\le P(+1|x)\le 1-\varepsilon\}$, then Assumption 2 assures $P(S)>0$. Due to the non-negativity of the loss function ℓ, we have

$$\begin{aligned}\mathcal{R}(f,\rho) &\ge -2\rho + \int_S\Big\{P(+1|x)\ell(\rho-f(x)) + P(-1|x)\ell(\rho+f(x))\Big\}P(dx)\\ &= \int_S\Big\{-\frac{2}{P(S)}\rho + P(+1|x)\ell(\rho-f(x)) + P(-1|x)\ell(\rho+f(x))\Big\}P(dx).\end{aligned}$$

For given η satisfying $\varepsilon\le\eta\le 1-\varepsilon$, we define the function $\xi(f,\rho)$ by

$$\xi(f,\rho) = -\frac{2}{P(S)}\rho + \eta\ell(\rho-f) + (1-\eta)\ell(\rho+f), \quad f,\rho\in\mathbb{R}.$$

We derive a lower bound $\inf\{\xi(f,\rho) : f,\rho\in\mathbb{R}\}$. Since $\ell(z)$ is a finite-valued convex function on $\mathbb{R}$, the subdifferential $\partial\xi(f,\rho)\subset\mathbb{R}^2$ is given as

$$\partial\xi(f,\rho) = \left\{\begin{pmatrix}0\\-2/P(S)\end{pmatrix} + \eta u\begin{pmatrix}-1\\1\end{pmatrix} + (1-\eta)v\begin{pmatrix}1\\1\end{pmatrix} : u\in\partial\ell(\rho-f), v\in\partial\ell(\rho+f)\right\}.$$

Formulas of the subdifferential are presented in Theorem 23.8 and Theorem 23.9 of Rockafellar (1970). We prove that there exist f^* and ρ^* such that $(0,0)^T\in\partial\xi(f^*,\rho^*)$ holds. Since the second condition in Assumption 3 holds for the convex function ℓ, the union $\cup_{z\in\mathbb{R}}\partial\ell(z)$ includes all the

positive real numbers. Hence, there exist real numbers z_1 and z_2 satisfying $\frac{1}{\eta P(S)} \in \partial\ell(z_1)$ and $\frac{1}{(1-\eta)P(S)} \in \partial\ell(z_2)$. Then, for $f^* = (z_2 - z_1)/2$, $\rho^* = (z_1 + z_2)/2$, the null vector is an element of $\partial\xi(f^*, \rho^*)$. Since $\xi(f, \rho)$ is convex in (f, ρ), the minimum value of $\xi(f, \rho)$ is attained at (f^*, ρ^*). Define z_{up} as a real number satisfying

$$g > \frac{1}{\varepsilon P(S)}, \quad \forall g \in \partial\ell(z_{\mathrm{up}}).$$

Since $\varepsilon \leq \eta \leq 1 - \varepsilon$ is assumed, both z_1 and z_2 are less than z_{up} due to the monotonicity of the subdifferential. Then, the inequality

$$\xi(f, \rho) \geq \xi(f^*, \rho^*) = -\frac{z_1 + z_2}{P(S)} + \eta\ell(z_1) + (1-\eta)\ell(z_2) \geq -\frac{2z_{\mathrm{up}}}{P(S)}$$

holds for all $f, \rho \in \mathbb{R}$ and all η such that $\varepsilon \leq \eta \leq 1 - \varepsilon$. The right-side of the expression above depends only on $P(S)$ and ε. Hence, for any measurable function $f \in L_0$ and $\rho \in \mathbb{R}$, we have

$$\mathcal{R}(f, \rho) \geq \int_S \frac{-2z_{\mathrm{up}}}{P(S)} P(dx) \geq -2z_{\mathrm{up}}.$$

As a result, we have $\mathcal{R}^* \geq -2z_{\mathrm{up}} > -\infty$. ■

Lemma 4 *Under Assumption 1, 2 and 3, we have*

$$\lim_{\lambda \to \infty} \inf\{\mathcal{R}_\lambda(f + b, \rho) : f \in \mathcal{H}, b, \rho \in \mathbb{R}\} = \mathcal{R}^*. \tag{31}$$

Proof Corollary 5.29 of Steinwart and Christmann (2008) ensures that the equality

$$\inf\{\mathbb{E}[\ell(\rho - y(f(x) + b))] : f \in \mathcal{H}, b \in \mathbb{R}\} = \inf\{\mathbb{E}[\ell(\rho - yf(x))] : f \in L_0\}$$

holds for any $\rho \in \mathbb{R}$. Thus, we have

$$\inf\{\mathcal{R}(f + b, \rho) : f \in \mathcal{H}, b \in \mathbb{R}\} = \inf\{\mathcal{R}(f, \rho) : f \in L_0\}$$

for any $\rho \in \mathbb{R}$. Then, the equality

$$\inf\{\mathcal{R}(f + b, \rho) : f \in \mathcal{H}, b, \rho \in \mathbb{R}\} = \mathcal{R}^*$$

holds. Under Assumption 2 and Assumption 3, we have $\mathcal{R}^* > -\infty$ due to Lemma 3. Then, for any $\varepsilon > 0$, there exist $\lambda_\varepsilon > 0, f_\varepsilon \in \mathcal{H}$, $b_\varepsilon \in \mathbb{R}$ and $\rho_\varepsilon \in \mathbb{R}$ such that $\|f_\varepsilon\|_{\mathcal{H}} \leq \lambda_\varepsilon$ and $\mathcal{R}(f_\varepsilon + b_\varepsilon, \rho_\varepsilon) \leq \mathcal{R}^* + \varepsilon$ hold. For all $\lambda \geq \lambda_\varepsilon$ we have

$$\inf\{\mathcal{R}_\lambda(f + b, \rho) : f \in \mathcal{H}, b, \rho \in \mathbb{R}\} \leq \mathcal{R}_\lambda(f_\varepsilon + b_\varepsilon, \rho_\varepsilon) = \mathcal{R}(f_\varepsilon + b_\varepsilon, \rho_\varepsilon) \leq \mathcal{R}^* + \varepsilon.$$

On the other hand, it is clear that the inequality $\mathcal{R}^* \leq \inf\{\mathcal{R}_\lambda(f + b, \rho) : f \in \mathcal{H}, b, \rho \in \mathbb{R}\}$ holds. Thus we obtain Equation (31). ■

We derive an upper bound on the norm of the optimal solution in (30).

Lemma 5 *Suppose* $\lim_{m\to\infty}\lambda_m = \infty$. *Under Assumption 1, 2 and 3, there are positive constants c and C and a natural number M such that the optimal solution of* (30) *satisfies*

$$\|\widehat{f}\|_{\mathcal{H}} \leq \lambda_m, \quad |\widehat{b}| \leq C\lambda_m, \quad |\widehat{\rho}| \leq C\lambda_m \tag{32}$$

with the probability greater than $1-e^{-cm}$ *for* $m \geq M$.

More precisely, M depends on the label probability, the function value $\ell(0)$ and the divergence speed of the sequence $\{\lambda_m\}$, and c depends only on the label probability. We can choose $C = K+1$, where K is defined in Assumption 1.

Proof Under Assumption 2, the label probabilities, $P(y=+1)$ and $P(y=-1)$, are positive. We assume that the inequalities

$$\frac{1}{2}P(Y=+1) < \frac{m_p}{m}, \quad \frac{1}{2}P(Y=-1) < \frac{m_n}{m} \tag{33}$$

hold. Applying Chernoff bound, we see that there exists a positive constant $c>0$ depending only on the marginal probability of the label such that (33) holds with the probability higher than $1-e^{-cm}$.

Lemma 2 ensures that the problem (30) has optimal solution, $\widehat{f},\widehat{b},\widehat{\rho}$. The first inequality in (32), that is, $\|\widehat{f}\|_{\mathcal{H}} \leq \lambda_m$, is clearly satisfied. Then, we have $\|\widehat{f}\|_\infty \leq K\|\widehat{f}\|_{\mathcal{H}} \leq K\lambda_m$ from the reproducing property of the RKHSs. The definition of the estimator and the non-negativity of ℓ yield that

$$-2\widehat{\rho} \leq -2\widehat{\rho} + \frac{1}{m}\sum_{i=1}^{m}\ell(\widehat{\rho} - y_i(\widehat{f}(x_i)+\widehat{b})) \leq \widehat{\mathcal{R}}_{T,\lambda_m}(0,0) = \ell(0).$$

Then, we have

$$\widehat{\rho} \geq -\frac{\ell(0)}{2}. \tag{34}$$

Next, we consider the optimality condition of $\widehat{\mathcal{R}}_{T,\lambda_m}$. According to the calculus of subdifferential introduced in Section 23 of Rockafellar (1970), the derivative of the objective function with respect to ρ leads to an optimality condition,

$$0 \in -2 + \frac{1}{m}\sum_{i=1}^{m}\partial\ell(\widehat{\rho} - y_i(\widehat{f}(x_i)+\widehat{b})).$$

The monotonicity and non-negativity of the subdifferential and the bound of $\|f\|_\infty$ lead to

$$\begin{aligned}
2 &\geq \frac{1}{m}\sum_{i=1}^{m}\partial\ell(\widehat{\rho} - y_i\widehat{b} - K\lambda_m) \\
&= \frac{1}{m}\sum_{i=1}^{m_p}\partial\ell(\widehat{\rho} - \widehat{b} - K\lambda_m) + \frac{1}{m}\sum_{j=1}^{m_n}\partial\ell(\widehat{\rho} + \widehat{b} - K\lambda_m) \\
&\geq \frac{1}{m}\sum_{i=1}^{m_p}\partial\ell(\widehat{\rho} - \widehat{b} - K\lambda_m).
\end{aligned} \tag{35}$$

The above expression means that there exists a number in the subdifferential such that the inequality holds, where $\sum_{i=1}^{m_p}\partial\ell$ denotes the m_p-fold sum of the set $\partial\ell$, that is, $\{a_1+\cdots+a_{m_p} : a_i \in \partial\ell, i=$

$1,\ldots,m_p\}$. Let z_p be a real number satisfying $\frac{2m}{m_p} < \partial\ell(z_p)$, that is, all elements in $\partial\ell(z_p)$ are greater than $\frac{2m}{m_p}$. Then, the inequality $\widehat{\rho} - \widehat{b} - K\lambda_m < z_p$ should hold. Otherwise the inequality (35) does not hold. In the same way, for z_n satisfying $\frac{2m}{m_n} < \partial\ell(z_n)$, we have $\widehat{\rho} + \widehat{b} - K\lambda_m < z_n$. The existence of z_p and z_n is guaranteed by Assumption 3. Hence, the inequalities

$$-\frac{\ell(0)}{2} \leq \widehat{\rho} \leq K\lambda_m + \max\{z_p, z_n\}, \qquad |\widehat{b}| \leq \frac{\ell(0)}{2} + K\lambda_m + \max\{z_p, z_n\} \tag{36}$$

hold, in which $\widehat{\rho} \geq -\ell(0)/2$ is used in the second inequality. Define $\bar{z}$ as a positive real number such that

$$\forall g \in \partial\ell(\bar{z}), \quad \max\left\{\frac{4}{P(Y=+1)}, \frac{4}{P(Y=-1)}\right\} < g.$$

Inequalities in (33) lead to

$$\max\left\{\frac{2m}{m_p}, \frac{2m}{m_n}\right\} < \max\left\{\frac{4}{P(Y=+1)}, \frac{4}{P(Y=-1)}\right\}.$$

Hence, we can choose $\bar{z} > 0$ satisfying $\max\{z_p, z_n\} < \bar{z}$. Note that $\bar{z}$ depends only on the label probability. Suppose that $\ell(0)/2 + \bar{z} \leq \lambda_m$ holds for $m \geq M$. Then from (36) we have

$$|\widehat{\rho}| \leq (K+1)\lambda_m, \quad |\widehat{b}| \leq (K+1)\lambda_m.$$

for $m \geq M$. Then we obtain (32) with $C = K+1$, when (33) holds. ■

Let us define the covering number for a metric space.

Definition 6 (covering number) *For a metric space $\mathcal{G}$, the covering number of $\mathcal{G}$ is defined as*

$$\mathcal{N}(\mathcal{G}, \varepsilon) = \min\left\{n \in \mathbb{N} : g_1, \ldots, g_n \in \mathcal{G} \text{ such that } \mathcal{G} \subset \bigcup_{i=1}^{n} B(g_i, \varepsilon)\right\},$$

where $B(g, \varepsilon)$ denotes the closed ball with center g and radius ε.

According to Lemma 5, the optimal solution $(\widehat{f}, \widehat{b}, \widehat{\rho})$ is included in the set

$$\mathcal{G}_m = \{(f, b, \rho) \in \mathcal{H} \times \mathbb{R}^2 : \|f\|_{\mathcal{H}} \leq \lambda_m, |b| \leq C\lambda_m, |\rho| \leq C\lambda_m\}$$

with high probability. Suppose that the norm $\|f\|_\infty + |b| + |\rho|$ is introduced on $\mathcal{G}_m$. We define the function parametrized by (f, b, ρ),

$$L(x, y; f, b, \rho) = -2\rho + \ell(\rho - y(f(x) + b)),$$

and the function set

$$\mathcal{L}_m = \{L(x, y; f, b, \rho) : (f, b, \rho) \in \mathcal{G}_m\}.$$

The supremum norm is defined on $\mathcal{L}_m$. The expected loss and the empirical loss, $\mathcal{R}(f + b, \rho)$ and $\widehat{\mathcal{R}}_T(f + b, \rho)$, are represented as the expectation of $L(x, y; f, b, \rho)$ with respect to the population

distribution and the empirical distribution, respectively. Since $\ell : \mathbb{R} \to \mathbb{R}$ is a finite-valued convex function, ℓ is locally Lipschitz continuous. Then, for any sample size m, there exists a constant κ_m depending on m such that

$$|\ell(z) - \ell(z')| \leq \kappa_m |z - z'| \tag{37}$$

holds for all z and z' satisfying $|z|, |z'| \leq (K + 2C)\lambda_m$. Then, for any $(f, b, \rho), (f', b', \rho') \in \mathcal{G}_m$, we have

$$\begin{aligned} |L(x, y; f, b, \rho) - L(x, y; f', b', \rho')| &\leq 2|\rho - \rho'| + \kappa_m(|\rho - \rho'| + |b - b'| + \|f - f'\|_\infty) \\ &\leq (2 + \kappa_m)(|\rho - \rho'| + |b - b'| + \|f - f'\|_\infty) \end{aligned}$$

The covering number of $\mathcal{L}_m$ is evaluated by using that of $\mathcal{G}_m$ as follows:

$$\mathcal{N}(\mathcal{L}_m, \varepsilon) \leq \mathcal{N}\big(\mathcal{G}_m, \frac{\varepsilon}{2 + \kappa_m}\big). \tag{38}$$

Let the metric space $\mathcal{F}_m$ be

$$\mathcal{F}_m = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda_m\}$$

with the supremum norm, then we also have

$$\begin{aligned} \mathcal{N}\left(\mathcal{G}_m, \frac{\varepsilon}{2 + \kappa_m}\right) &\leq \mathcal{N}\left(\mathcal{F}_m, \frac{\varepsilon}{3(2 + \kappa_m)}\right) \left(\frac{2C\lambda_m}{\frac{\varepsilon}{3(2+\kappa_m)}}\right)^2 \\ &= \mathcal{N}\left(\mathcal{F}_m, \frac{\varepsilon}{3(2 + \kappa_m)}\right) \left(\frac{6C\lambda_m(2 + \kappa_m)}{\varepsilon}\right)^2. \end{aligned} \tag{39}$$

An upper bound of the covering number of $\mathcal{F}_m$ endowed with the supremum norm is given by Cucker and Smale (2002) and Zhou (2002).

We prove the uniform convergence of $\widehat{\mathcal{R}}(f + b, \rho)$.

Lemma 7 *Let b_m be $b_m = 4C\lambda_m + \ell((K + 2C)\lambda_m)$ in which C is the positive constant defined in Lemma 5. Under Assumption 1 and 3, the inequality*

$$\begin{aligned} &P\left(\sup_{(f,b,\rho)\in\mathcal{G}_m} |\widehat{\mathcal{R}}(f + b, \rho) - \mathcal{R}(f + b, \rho)| \geq \varepsilon\right) \\ &\leq 2\mathcal{N}(\mathcal{L}_m, \varepsilon/3) \exp\left\{-\frac{2m\varepsilon^2}{9b_m^2}\right\} \qquad (40) \\ &\leq 2\mathcal{N}\left(\mathcal{F}_m, \frac{\varepsilon}{9(2 + \kappa_m)}\right) \left(\frac{18C\lambda_m(2 + \kappa_m)}{\varepsilon}\right)^2 \exp\left\{-\frac{2m\varepsilon^2}{9b_m^2}\right\} \qquad (41) \end{aligned}$$

holds, where κ_m is the Lipschitz constant defined by (37).

Proof Since $\|f\|_\infty \leq K\lambda_m$ holds for $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} \leq \lambda_m$, we have the following inequality

$$\begin{aligned}
&\sup_{\substack{(x,y)\in\mathcal{X}\times\{+1,-1\}\\(f,b,\rho)\in\mathcal{G}_m}} L(x,y;f,b,\rho) - \inf_{\substack{(x,y)\in\mathcal{X}\times\{+1,-1\}\\(f,b,\rho)\in\mathcal{G}_m}} L(x,y;f,b,\rho) \\
&\leq 2C\lambda_m + \sup_{\substack{(x,y)\in\mathcal{X}\times\{+1,-1\}\\(f,b,\rho)\in\mathcal{G}_m}} \ell(\rho - y(f(x)+b)) - (-2C\lambda_m) \\
&\leq 4C\lambda_m + \ell(C\lambda_m + K\lambda_m + C\lambda_m) \\
&= b_m.
\end{aligned}$$

In the same way as the proof of Lemma 3.4 in Steinwart (2005), Hoeffding's inequality leads to the upper bound (40). Equation (41) is the direct conclusion of (38) and (39). ■

We present the main theorem of this section.

Theorem 8 *Suppose that* $\lim_{m\to\infty}\lambda_m = \infty$ *holds. Suppose that Assumption 1, 2 and 3 hold. Moreover we assume that* (41) *converges to zero for any* $\varepsilon > 0$*, when the sample size* m *tends to infinity. Then,* $\mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho})$ *converges to* $\mathcal{R}^*$ *in probability in the large sample limit of the data set* $T = \{(x_i,y_i) : i = 1,\ldots,m\}$.

Later on we show an example in which 7 converges to zero.

Proof Lemma 4 assures that, for any $\gamma > 0$, there exists sufficiently large M_1 such that

$$|\inf\{\mathcal{R}_{\lambda_m}(f+b,\rho) : f \in \mathcal{H}, b,\rho \in \mathbb{R}\} - \mathcal{R}^*| \leq \gamma$$

holds for all $m \geq M_1$. Thus, there exist f_γ, b_γ and ρ_γ such that

$$|\mathcal{R}_{\lambda_m}(f_\gamma + b_\gamma, \rho_\gamma) - \mathcal{R}^*| \leq 2\gamma$$

and $\|f_\gamma\|_{\mathcal{H}} \leq \lambda_m$ hold for $m \geq M_1$. Due to the law of large numbers, the inequality

$$|\widehat{\mathcal{R}}_T(f_\gamma + b_\gamma, \rho_\gamma) - \mathcal{R}(f_\gamma + b_\gamma, \rho_\gamma)| \leq \gamma$$

holds with high probability, say $1-\delta_m$, for $m \geq M_2$. The boundedness property in Lemma 5 leads to

$$P((\widehat{f},\widehat{b},\widehat{\rho}) \in \mathcal{G}_m) \geq 1 - e^{-cm}$$

for $m \geq M_3$. In addition, by the uniform bound shown in Lemma 7, the inequality

$$\sup_{(f,b,\rho)\in\mathcal{G}_m} |\widehat{\mathcal{R}}_T(f+b,\rho) - \mathcal{R}(f+b,\rho)| \leq \gamma$$

holds with probability $1-\delta'_m$. Hence, the probability such that the inequality

$$|\widehat{\mathcal{R}}_T(\widehat{f}+\widehat{b},\widehat{\rho}) - \mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho})| \leq \gamma$$

holds is greater than $1-e^{-cm}-\delta'_m$ for $m \geq M_3$. Let M_0 be $M_0 = \max\{M_1, M_2, M_3\}$. Then, for any $\gamma > 0$, the following inequalities hold with probability higher than $1-e^{-cm}-\delta'_m-\delta_m$ for $m \geq M_0$,

$$\begin{aligned}
\mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho}) &\leq \widehat{\mathcal{R}}_T(\widehat{f}+\widehat{b},\widehat{\rho})+\gamma \\
&\leq \widehat{\mathcal{R}}_T(f_\gamma+b_\gamma,\rho_\gamma)+\gamma \\
&\leq \mathcal{R}(f_\gamma+b_\gamma,\rho_\gamma)+2\gamma \\
&= \mathcal{R}_{\lambda_m}(f_\gamma+b_\gamma,\rho_\gamma)+2\gamma \\
&\leq \mathcal{R}^*+4\gamma.
\end{aligned} \tag{42}$$

The second inequality (42) above is given as

$$\widehat{\mathcal{R}}_T(\widehat{f}+\widehat{b},\widehat{\rho}) = \widehat{\mathcal{R}}_{T,\lambda_m}(\widehat{f}+\widehat{b},\widehat{\rho}) \leq \widehat{\mathcal{R}}_{T,\lambda_m}(f_\gamma+b_\gamma,\rho_\gamma) = \widehat{\mathcal{R}}_T(f_\gamma+b_\gamma,\rho_\gamma).$$

■

We show the order of λ_m admitting the assumption in Theorem 8.

Example 8 *Suppose that $\mathcal{X} = [0,1]^n \subset \mathbb{R}^n$ and the Gaussian kernel is used. According to Zhou (2002), we have*

$$\log \mathcal{N}\left(\mathcal{F}_m, \frac{\varepsilon}{9(2+\kappa_m)}\right) = O\left(\left(\log \frac{\lambda_m}{\frac{\varepsilon}{9(2+\kappa_m)}}\right)^{n+1}\right) = O\big(\big(\log(\lambda_m\kappa_m)\big)^{n+1}\big).$$

For any $\varepsilon > 0$, (41) *is bounded above by*

$$\exp\left\{O\left(-\frac{m}{b_m^2}+(\log(\lambda_m\kappa_m))^{n+1}\right)\right\}.$$

For the truncated quadratic loss, we have

$$\begin{aligned}
&\kappa_m \leq 2((K+2C)\lambda_m+1) = O(\lambda_m), \\
&b_m \leq 4C\lambda_m+((K+2C)\lambda_m+1)^2 = O(\lambda_m^2).
\end{aligned}$$

Let us define $\lambda_m = m^\alpha$ with $0 < \alpha < 1/4$. Then, for any $\varepsilon > 0$, (41) *converges to zero when m tends to infinity. In the same way, for the exponential loss we obtain*

$$\kappa_m = O(e^{(K+2C)\lambda_m}), \quad b_m = O(e^{(K+2C)\lambda_m}).$$

Hence, $\lambda_m = (\log m)^\alpha$ with $0 < \alpha < 1$ assures the convergence of (41).

C.2 Convergence to Bayes Risk

We prove that the expected 0-1 loss $\mathcal{E}(\widehat{f}+\widehat{b})$ converges to the Bayes risk $\mathcal{E}^*$, when the sample size m tends to infinity.

Theorem 9 *Suppose that $\mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho})$ converges to $\mathcal{R}^*$ in probability, when the sample size m tends to infinity. For the RKHS $\mathcal{H}$ and the loss function ℓ, we assume Assumption 1, 3 and 4. Then, $\mathcal{E}(\widehat{f}+\widehat{b})$ converges to $\mathcal{E}^*$ in probability.*

As a result, we find that the prediction error rate of $\widehat{f}+\widehat{b}$ converges to the Bayes risk under Assumption 1, 2, 3, 4, and the assumption on the covering number in Theorem 8.

Proof Suppose that ρ satisfies $\rho \geq -\ell(0)/2$. Since $\ell'(\rho) > 0$ holds, the loss function $\ell(\rho - z)$ is classification-calibrated (Bartlett et al., 2006). Hence, for $\rho \geq -\ell(0)/2$ Theorem 1 and Theorem 2 of Bartlett et al. (2006) guarantee that $\psi(\theta,\rho)$ in Assumption 4 satisfies $\psi(0,\rho) = 0$, $\psi(\theta,\rho) > 0$ for $0 < \theta \leq 1$ and that $\psi(\theta,\rho)$ is continuous and strictly increasing in $\theta \in [0,1]$. In addition, for all $f \in \mathcal{H}$ and $b \in \mathbb{R}$ the inequality

$$\psi(\mathcal{E}(f+b) - \mathcal{E}^*, \rho) \leq \mathbb{E}[\ell(\rho - y(f(x)+b))] - \inf_{f\in\mathcal{H}, b\in\mathbb{R}} \mathbb{E}[\ell(\rho - y(f(x)+b))]$$

holds. Here we used the equality

$$\inf\{\mathbb{E}[\ell(\rho - y(f(x)+b))] : f \in \mathcal{H}, b \in \mathbb{R}\} = \inf\{\mathbb{E}[\ell(\rho - y(f(x)+b))] : f \in L_0, b \in \mathbb{R}\},$$

which is shown in Corollary 5.29 of Steinwart and Christmann (2008). Hence, we have

$$\begin{aligned}\psi(\mathcal{E}(\widehat{f}+\widehat{b}) - \mathcal{E}^*, \widehat{\rho}) &\leq \mathbb{E}[\ell(\widehat{\rho} - y(\widehat{f}(x)+\widehat{b}))] - \inf_{f\in\mathcal{H}, b\in\mathbb{R}} \mathbb{E}[\ell(\widehat{\rho} - y(f(x)+b))] \\ &= \mathcal{R}(\widehat{f}+\widehat{b}, \widehat{\rho}) - \inf_{f\in\mathcal{H}, b\in\mathbb{R}} \mathcal{R}(f+b, \widehat{\rho}),\end{aligned}$$

since $\widehat{\rho} \geq -\ell(0)/2$ holds due to (34). Since $\mathcal{R}(\widehat{f}+\widehat{b}, \widehat{\rho})$ is assumed to converge to $\mathcal{R}^*$ in probability, for any $\varepsilon > 0$ the inequality

$$\mathcal{R}^* \leq \inf_{f\in\mathcal{H}, b\in\mathbb{R}} \mathcal{R}(f+b, \widehat{\rho}) \leq \mathcal{R}(\widehat{f}+\widehat{b}, \widehat{\rho}) \leq \mathcal{R}^* + \varepsilon$$

holds with high probability for sufficiently large m. Thus, $\psi(\mathcal{E}(\widehat{f}+\widehat{b}) - \mathcal{E}^*, \widehat{\rho})$ converges to zero in probability. The inequality

$$0 \leq \widetilde{\psi}(\mathcal{E}(\widehat{f}+\widehat{b}) - \mathcal{E}^*) \leq \psi(\mathcal{E}(\widehat{f}+\widehat{b}) - \mathcal{E}^*, \widehat{\rho})$$

and the assumption on the function $\widetilde{\psi}$ ensure that $\mathcal{E}(\widehat{f}+\widehat{b})$ converges to $\mathcal{E}^*$ in probability, when m tends to infinity. ■

C.3 Sufficient Conditions for Existence of the Function $\widetilde{\psi}$ in Assumption 4

We present some sufficient conditions for existence of the function $\widetilde{\psi}$ in Assumption 4.

Lemma 10 *Suppose that the first condition in Assumption 3 and the first condition in Assumption 4 hold. In addition, suppose that ℓ is first-order continuously differentiable on $\mathbb{R}$. Let d be $d = \sup\{z \in \mathbb{R} : \ell'(z) = 0\}$, where ℓ' is the derivative of ℓ. When $\ell'(z) > 0$ holds for all $z \in \mathbb{R}$, we define $d = -\infty$. We assume the following conditions:*

1. *$d < -\ell(0)/2$.*
2. *$\ell(z)$ is second-order continuously differentiable on the open interval (d, ∞).*

3. $\ell''(z) > 0$ *holds on* (d, ∞).

4. $1/\ell'(z)$ *is convex on* (d, ∞).

Then, for any $\theta \in [0,1]$*, the function* $\psi(\theta, \rho)$ *is non-decreasing as the function of* ρ *for* $\rho \geq -\ell(0)/2$.

When the condition in Lemma 10 is satisfied, we can choose $\psi(\theta, -\ell(0)/2)$ as $\widetilde{\psi}(\theta)$ for $0 \leq \theta \leq 1$, since $\psi(\theta, -\ell(0)/2)$ is classification-calibrated under the first condition in Assumption 4.

Proof For $\theta = 0$ and $\theta = 1$, we can directly confirm that the lemma holds. In the following, we assume $0 < \theta < 1$ and $\rho \geq -\ell(0)/2$. We consider the following optimization problem involved in $\psi(\theta, \rho)$,

$$\inf_{z \in \mathbb{R}} \frac{1+\theta}{2}\ell(\rho - z) + \frac{1-\theta}{2}\ell(\rho + z). \tag{43}$$

The function in the infimum is a finite-valued convex function for $z \in \mathbb{R}$, and diverges to infinity when z tends to $\pm\infty$. Thus the problem (43) has an optimal solution $z^* \in \mathbb{R}$. The optimality condition leads to the equality

$$(1+\theta)\ell'(\rho - z^*) - (1-\theta)\ell'(\rho + z^*) = 0.$$

We assumed that both $1+\theta$ and $1-\theta$ are positive and that $\rho \geq -\ell(0)/2 > d$ holds. Hence, both $\ell'(\rho - z^*)$ and $\ell'(\rho + z^*)$ should not be zero. Indeed, if one of them is equal to zero, the other is also zero, and we have $\rho - z^* \leq d$ and $\rho + z^* \leq d$. We find that these inequalities contradict $\rho > d$. As a result, we have $\rho - z^* > d$ and $\rho + z^* > d$, that is, $|z^*| < \rho - d$. In addition, we have

$$\frac{1+\theta}{2} = \frac{\ell'(\rho + z^*)}{\ell'(\rho + z^*) + \ell'(\rho - z^*)}.$$

Since $\ell''(z) > 0$ holds on (d, ∞), the second derivative of the objective in (43) satisfies the positivity condition,

$$(1+\theta)\ell''(\rho - z) + (1-\theta)\ell''(\rho + z) > 0$$

for all z such that $\rho - z > d$ and $\rho + z > d$. Therefore, z^* is uniquely determined. For a fixed $\theta \in (0,1)$, the optimal solution can be described as the function of ρ, that is, $z^* = z(\rho)$. By the implicit function theorem, $z(\rho)$ is continuously differentiable with respect to ρ. Then, the derivative of $\psi(\theta, \rho)$ is given as

$$\begin{aligned}
\frac{\partial}{\partial \rho}\psi(\theta, \rho) &= \frac{\partial}{\partial \rho}\left\{\ell(\rho) - \frac{1+\theta}{2}\ell(\rho - z(\rho)) - \frac{1-\theta}{2}\ell(\rho + z(\rho))\right\} \\
&= \ell'(\rho) - \frac{1+\theta}{2}\ell'(\rho - z(\rho))\left(1 - \frac{\partial z}{\partial \rho}\right) - \frac{1-\theta}{2}\ell'(\rho + z(\rho))\left(1 + \frac{\partial z}{\partial \rho}\right) \\
&= \ell'(\rho) - \frac{\ell'(\rho + z(\rho))}{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))}\ell'(\rho - z(\rho))\left(1 - \frac{\partial z}{\partial \rho}\right) \\
&\quad - \frac{\ell'(\rho - z(\rho))}{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))}\ell'(\rho + z(\rho))\left(1 + \frac{\partial z}{\partial \rho}\right) \\
&= \ell'(\rho) - \frac{2\ell'(\rho - z(\rho))\ell'(\rho + z(\rho))}{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))}.
\end{aligned}$$

The convexity of $1/\ell'(z)$ for $z > d$ leads to

$$0 < \frac{1}{\ell'(\rho)} \leq \frac{1}{2\ell'(\rho + z(\rho))} + \frac{1}{2\ell'(\rho - z(\rho))} = \frac{\ell'(\rho + z(\rho)) + \ell'(\rho - z(\rho))}{2\ell'(\rho - z(\rho))\ell'(\rho + z(\rho))}.$$

Hence, we have

$$\frac{\partial}{\partial \rho}\psi(\theta, \rho) \geq 0$$

for $\rho \geq -\ell(0)/2 > d$ and $0 < \theta < 1$. As a result, we see that $\psi(\theta, \rho)$ is non-decreasing as the function of ρ. ■

We give another sufficient condition for existence of the function $\widetilde{\psi}$ in Assumption 4.

Lemma 11 *Suppose that the first condition in Assumption 3 and the first condition in Assumption 4 hold. Let d be* $d = \sup\{z \in \mathbb{R} : \partial\ell(z) = \{0\}\}$. *When* $0 \notin \partial\ell(z)$ *holds for all* $z \in \mathbb{R}$, *we define* $d = -\infty$. *Suppose that the inequality* $-\ell(0)/2 > d$ *holds. For* $\rho \geq -\ell(0)/2$ *and* $z \geq 0$, *we define* $\xi(z, \rho)$ *by*

$$\xi(z, \rho) = \begin{cases} \dfrac{\ell(\rho + z) + \ell(\rho - z) - 2\ell(\rho)}{z\ell'(\rho)}, & z > 0, \\ 0, & z = 0. \end{cases}$$

Suppose that there exists a function $\bar{\xi}(z)$ *for* $z \geq 0$ *such that the following conditions hold:*

1. $\bar{\xi}(z)$ *is continuous and strictly increasing on* $z \geq 0$, *and satisfies* $\bar{\xi}(0) = 0$ *and* $\lim_{z\to\infty} \bar{\xi}(z) > 1$.
2. $\sup_{\rho \geq -\ell(0)/2} \xi(z, \rho) \leq \bar{\xi}(z)$ *holds.*

Then, there exists a function $\widetilde{\psi}$ *defined in the second condition of Assumption 4.*

Note that Lemma 11 does not require the second order differentiability of the loss function.

Proof We use the result of Bartlett et al. (2006). For a fixed ρ, the function $\xi(z, \rho)$ is continuous for $z \geq 0$, and the convexity of ℓ leads to the non-negativity of $\xi(z, \rho)$. Moreover, the convexity and the non-negativity of $\ell(z)$ lead to

$$\xi(z, \rho) \geq \frac{\ell(\rho + z) - \ell(\rho)}{z\ell'(\rho)} - \frac{\ell(\rho)}{z\ell'(\rho)} \geq 1 - \frac{\ell(\rho)}{z\ell'(\rho)}$$

for $z > 0$ and $\rho \geq -\ell(0)/2$, where $\ell(\rho)$ and $\ell'(\rho)$ are positive for $\rho > -\ell(0)/2$. The above inequality and the continuity of $\xi(\cdot, \rho)$ ensure that there exists z satisfying $\xi(z, \rho) = \theta$ for all θ such that $0 \leq \theta < 1$. We define the inverse function ξ_ρ^{-1} by

$$\xi_\rho^{-1}(\theta) = \inf\{z \geq 0 : \xi(z, \rho) = \theta\}$$

for $0 \leq \theta < 1$. For a fixed $\rho \geq -\ell(0)/2$, the loss function $\ell(\rho - z)$ is classification-calibrated (Bartlett et al., 2006). Hence, Lemma 3 in Bartlett et al. (2006) leads to the inequality

$$\psi(\theta, \rho) \geq \ell'(\rho)\frac{\theta}{2}\xi_\rho^{-1}\left(\frac{\theta}{2}\right),$$

for $0 \leq \theta < 1$. Define $\bar{\xi}^{-1}$ by

$$\bar{\xi}^{-1}(\theta) = \inf\{z \geq 0 : \bar{\xi}(z) = \theta\}.$$

From the definition of $\bar{\xi}(z)$, $\bar{\xi}^{-1}(\theta)$ is well-defined for all $\theta \in [0,1)$. Since $\xi(z,\rho) \leq \bar{\xi}(z)$ holds, we have $\xi_\rho^{-1}(\theta/2) \geq \bar{\xi}^{-1}(\theta/2)$. In addition, $\ell'(\rho)$ is non-decreasing as the function of ρ. Thus, we have

$$\psi(\theta,\rho) \geq \ell'(-\ell(0)/2)\frac{\theta}{2}\bar{\xi}^{-1}(\frac{\theta}{2})$$

for all $\rho \geq -\ell(0)/2$ and $0 \leq \theta < 1$. Then, we can choose

$$\widetilde{\psi}(\theta) = \ell'(-\ell(0)/2)\frac{\theta}{2}\bar{\xi}^{-1}(\frac{\theta}{2}).$$

It is straightforward to confirm that the conditions of Assumption 4 are satisfied. ■

We show some examples in which the existence of $\widetilde{\psi}$ is confirmed from above lemmas.

Example 9 *For the truncated quadratic loss $\ell(z) = (\max\{z+1,0\})^2$, the first condition in Assumption 3 and the first condition in Assumption 4 hold. The inequality $-\ell(0)/2 = -1/2 > \sup\{z : \ell'(z) = 0\} = -1$ in the sufficient condition of Lemma 10 holds. For $z > -1$, it is easy to see that $\ell(z)$ is second-order differentiable and that $\ell''(z) > 0$ holds. In addition, for $z > -1$, $1/\ell'(z)$ is equal to $1/(2z+2)$ which is convex on $(-1,\infty)$. Therefore, the function $\widetilde{\psi}(\theta) = \psi(\theta,-1/2)$ satisfies the second condition in Assumption 4.*

Example 10 *For the exponential loss $\ell(z) = e^z$, we have $1/\ell'(z) = e^{-z}$. Hence, due to Lemma 10, $\psi(\theta,\rho)$ is non-decreasing in ρ. Indeed, we have $\psi(\theta,\rho) = (1-\sqrt{1-\theta^2})e^\rho$.*

Example 11 *In Example 6, we presented the uncertainty set with estimation errors. The uncertainty sets are defined based on the revised function $\bar{\ell}(z)$ in* (23)*. Here, we use a similar function defined by*

$$\bar{\ell}^*(\alpha) = \begin{cases} (|\alpha w - 1| + h)^2 - (1+h)^2, & \alpha \geq 0, \\ \infty, & \alpha < 0, \end{cases} \tag{44}$$

for the construction of uncertainty sets. The function of the form (44) *is derived by setting $\boldsymbol{\mu}_p^T \Sigma_p^{-1} \boldsymbol{\mu}_p = 1$ and $\boldsymbol{\mu}_n = \mathbf{0}$ in* (23)*. Here, w and h are positive constants, and we suppose $w > 1/2$. The corresponding loss function is given as $\bar{\ell}(z)$. Then we have $\bar{\ell}(z) = u(z/w)$ defined in* (24)*. For $w > 1/2$, we can confirm that $\sup\{z : \bar{\ell}'(z) = 0\} < -\bar{\ell}(0)/2$ holds. Since $u(z)$ is not strictly convex, Lemma 10 does not work. Hence, we apply Lemma 11. A simple calculation yields that $\bar{\ell}'(-\bar{\ell}(0)/2) \geq (4w-1)/(4w^2) > 0$ for any $h \geq 0$. Note that $\bar{\ell}(z)$ is differentiable on $\mathbb{R}$. Thus, the monotonicity of $\bar{\ell}'$ for the convex function leads to*

$$\xi(z,\rho) = \frac{1}{\bar{\ell}'(\rho)}\left(\frac{\bar{\ell}(\rho+z)-\bar{\ell}(\rho)}{z} - \frac{\bar{\ell}(\rho)-\bar{\ell}(\rho-z)}{z}\right) \leq \frac{\bar{\ell}'(\rho+z)-\bar{\ell}'(\rho-z)}{\bar{\ell}'(\rho)}.$$

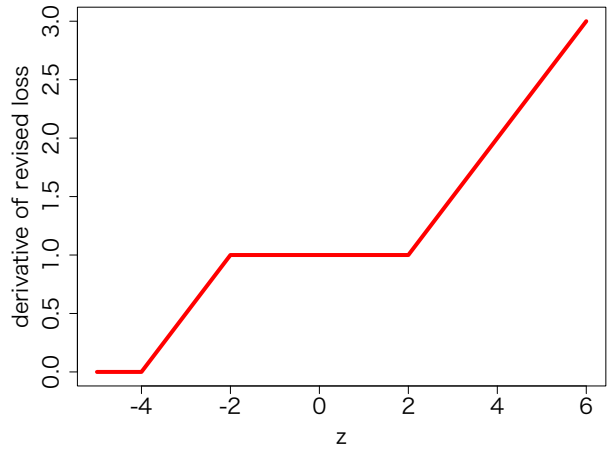


Figure 6: The derivative of the loss function corresponding to the revised uncertainty set with the estimation error.

Figure 6 depicts the derivative of $\bar{\ell}$ *with* $h = 1$ *and* $w = 1$. *Since the derivative* $\bar{\ell}'(z)$ *is Lipschitz continuous and the Lipschitz constant is equal to* $1/(2w)$, *we have* $\bar{\ell}'(\rho + z) - \bar{\ell}'(\rho - z) \leq z/w$. *Therefore, the inequality*

$$\sup_{\rho \geq -\bar{\ell}(0)/2} \xi(z, \rho) \leq \sup_{\rho \geq -\bar{\ell}(0)/2} \frac{z/w}{\bar{\ell}'(\rho)} = \frac{z/w}{\bar{\ell}'(-\bar{\ell}(0)/2)} \leq \frac{4w}{4w-1} z \leq 2z$$

holds. We see that $\bar{\xi}(z) = 2z$ *satisfies the sufficient condition of Lemma 11. The inequality*

$$\bar{\ell}'(-\bar{\ell}(0)/2) \frac{\theta}{2} \bar{\xi}^{-1}(\frac{\theta}{2}) \geq \frac{4w-1}{32w^2} \theta^2$$

ensures that $\widetilde{\psi}(\theta) = \frac{4w-1}{32w^2}\theta^2$ *is a valid choice. Therefore, the loss function corresponding to the revised uncertainty set in Example 6 satisfies the sufficient conditions for the statistical consistency.*

References

S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *J. Comput. Syst. Sci.*, 54(2):317–331, 1997.

P. L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, 8:775–790, April 2007.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

A. Ben-Tal and A. Nemirovski. Robust optimization - methodology and applications. *Math. Program.*, 92(3):453–480, 2002.

A. Ben-Tal, L. El-Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, 2009.

K. P. Bennett and E. J. Bredensteiner. Duality and geometry in SVM classifiers. In *Proceedings of International Conference on Machine Learning*, pages 57–64, 2000.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, 2004.

D. Bertsekas, A. Nedic, and A. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Belmont, MA, 2003.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

D. J. Crisp and C. J. C. Burges. A geometric interpretation of ν-SVM classifiers. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 244–250. MIT Press, 2000.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. *Laboratory, Massachusetts Institute of Technology*, 1999.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, aug 1997.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998.

T. Hastie, R. Tibishirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

K. Huang, H. Yang, I. King, M. R. Lyu, and Laiwan Chan. The minimum error minimax probability machine. *Journal of Machine Learning Research*, 5:1253–1286, 2004.

A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2003.

M. E. Mavroforakis and S. Theodoridis. A geometric approach to support vector machine (svm) classification. *IEEE Transactions on Neural Networks*, 17(3):671–682, 2006.

J. S. Nath and C. Bhattacharyya. Maximum margin classifiers with specified false positive and false negative error rates. In C. Apte, B. Liu, S. Parthasarathy, and D. Skillicorn, editors, *Proceedings of the seventh SIAM International Conference on Data mining*, pages 35–46. SIAM, 2007.

G. Rätsch, B. Schölkopf, A.J. Smola, S. Mika, T. Onoda, and K.-R. Müller. *Robust Ensemble Learning.*, pages 207–220. MIT Press, Cambridge, MA, 2000.

G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine Learning*, 42(3): 287–320, 2001.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.

R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1472, 2002.

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

I. Steinwart. On the optimal parameter choice for ν-support vector machines. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1274–1284, 2003.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008.

S. Sun and J. Shawe-Taylor. Sparse semi-supervised learning using conjugate functions. *Journal of Machine Learning Research*, 11:2423–2455, 2010.

A. Takeda and M. Sugiyama. ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1056–1063, 2008.

A. Takeda, H. Mitsugi, and T. Kanamori. A unified classification model based on robust optimization. *Neural Computation*, 25(3):759–804, 2013.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statitics*, 32(1):56–85, 2004.

D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.