JMLR: Workshop and Conference Proceedings vol 23 (2012) 29.1–29.23 25th Annual Conference on Learning Theory

A Conjugate Property between Loss Functions and Uncertainty Sets in Classification Problems

Takafumi Kanamori KANAMORI@IS.NAGOYA-U.AC.JP
Nagoya University
Furocho, Chikusaku, Nagoya 464-8603, Japan

Akiko Takeda TAKEDA@AE.KEIO.AC.JP
Keio University
3-14-1 Hiyoshi, Kouhoku, Yokohama, Kanagawa 223-8522, Japan

Taiji Suzuki S-TAIJI@STAT.T.U-TOKYO.AC.JP
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

In binary classification problems, mainly two approaches have been proposed; one is loss function approach and the other is minimum distance approach. The loss function approach is applied to major learning algorithms such as support vector machine (SVM) and boosting methods. The loss function represents the penalty of the decision function on the training samples. In the learning algorithm, the empirical mean of the loss function is minimized to obtain the classifier. Against a backdrop of the development of mathematical programming, nowadays learning algorithms based on loss functions are widely applied to real-world data analysis. In addition, statistical properties of such learning algorithms are well-understood based on a lots of theoretical works. On the other hand, some learning methods such as ν-SVM, mini-max probability machine (MPM) can be formulated as minimum distance problems. In the minimum distance approach, firstly, the so-called uncertainty set is defined for each binary label based on the training samples. Then, the best separating hyperplane between the two uncertainty sets is employed as the decision function. This is regarded as an extension of the maximum-margin approach. The minimum distance approach is considered to be useful to construct the statistical models with an intuitive geometric interpretation, and the interpretation is helpful to develop the learning algorithms. However, the statistical properties of the minimum distance approach have not been intensively studied. In this paper, we consider the relation between the above two approaches. We point out that the uncertainty set in the minimum distance approach is described by using the level set of the conjugate of the loss function. Based on such relation, we study statistical properties of the minimum distance approach.

Keywords: loss function; minimum distance problem; uncertainty set; Legendre transformation; consistency.

1. Introduction

We study binary classification problems. We define $\mathcal{X}$ as the input space and $\{+1, -1\}$ as the set of the output 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 label associated with the input x under the probability distribution P. The composite function of the sign function and the decision function, $\mathrm{sign}(f(x))$, is referred to as classifier.

In binary classification problems, the prediction accuracy of the decision function f is measured by the 0-1 loss $[\![yf(x) \leq 0]\!]$, where $[\![A]\!]$ is the indicator function i.e., $[\![A]\!]$ equals 1 if A holds and 0 otherwise. The average prediction performance of the decision function f is evaluated by the expected 0-1 loss, $\mathcal{E}(f) = \mathbb{E}[[\![yf(x) \leq 0]\!]]$. The Bayes risk $\mathcal{E}^*$ is defined as the minimum value of the expected 0-1 loss over all the measurable functions on $\mathcal{X}$, i.e., $\mathcal{E}^* = \inf\{\mathcal{E}(f) : f \in L_0\}$, where L_0 is the set of all measurable functions on $\mathcal{X}$. The Bayes risk is the lowest achievable error rate under the probability P.

Many learning algorithms have been proposed to attack binary classification problems. Here, we introduce ν-support vector machine (SVM) (Schölkopf et al., 2000) as a popular method for classification problems. Based on ν-SVM, we explain the two aspects in the statistical learning, i.e., the loss function approach and the minimum distance approach. Suppose that the input space $\mathcal{X}$ is a subset of the Euclidean space $\mathbb{R}^d$. We consider 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 to be estimated from the training samples. In ν-SVM, the estimator is given by the optimal solution of the optimization problem,

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

where $\|\boldsymbol{w}\|$ denotes the Euclidean norm of $\boldsymbol{w}$. In the above, the parameter $\nu \in (0, 1)$ is a prespecified constant which has the role of the regularization parameter. As Schölkopf et al. (2000) pointed out, the parameter ν controls the number of margin errors and the number of support vectors. In ν-SVM, a variant of the hinge loss, $\max\{\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b),\, 0\}$, is used. In the original formulation of ν-SVM, the non-negativity constraint, $\rho \geq 0$, is introduced for the parameter ρ. We can confirm that for $\nu > 0$, the optimal value of ρ in (1) is non-negative, even when the non-negativity constraint is dropped (Crisp and Burges, 2000).

As pointed out by Crisp and Burges (2000) and Bennett and Bredensteiner (2000), the dual problem of (1) is given as

$$\inf_{\boldsymbol{z}_p,\boldsymbol{z}_n} \|\boldsymbol{z}_p - \boldsymbol{z}_n\| \quad \text{subject to } \boldsymbol{z}_p \in \mathcal{U}_+,\ \boldsymbol{z}_n \in \mathcal{U}_-, \tag{2}$$

where $\mathcal{U}_+$ and $\mathcal{U}_-$ are the reduced convex hulls of the input vectors, i.e., $\mathcal{U}_\pm = \{\sum_{i\in M_\pm} \alpha_i \boldsymbol{x}_i : \sum_{i\in M_\pm} \alpha_i = 1,\ 0 \leq \alpha_i \leq \frac{2}{m\nu}, i \in M_\pm\}$ and $M_+(\text{resp.}M_-) = \{i : y_i = +1(\text{resp.} -1), i = 1, \ldots, m\}$. Given the optimal solutions $\widehat{\boldsymbol{z}}_p, \widehat{\boldsymbol{z}}_n$ for the dual problem (2), the optimal solution of $\boldsymbol{w}$ in (1) is proportional to $\widehat{\boldsymbol{z}}_p - \widehat{\boldsymbol{z}}_n$ with a positive proportional constant. The problem (2) is referred to as the *minimum distance problem*. Instead of the reduced convex hulls, the ellipsoidal sets are also used as $\mathcal{U}_\pm$ (Lanckriet et al., 2003; Nath and Bhattacharyya, 2007). In this paper, the subset $\mathcal{U}_\pm$ is called *uncertainty set*. The minimum distance approach using the uncertainty set is considered to be useful to construct the statistical models with an intuitive geometric interpretation. The interpretation is helpful to develop the learning algorithms (Mavroforakis and Theodoridis, 2006).

The main purpose of this paper is to study the relation between the loss function approach and the minimum distance approach. Up to our knowledge, statistical properties of the minimum distance approach have not been intensively studied. The study of the relation between two approaches enables us to understand learning algorithms using uncertainty sets. We point out that in general

the minimum distance problem with a fixed uncertainty set does not provide an accurate decision function. We need to introduce an uncertainty set having a one-dimensional parameter which specifies the size of the uncertainty set. In this paper, we present some examples of the parametrized uncertainty sets. For a wide class of learning algorithms using uncertainty sets, we show that a revised minimum distance problem with the parametrized uncertainty set recovers the statistical consistency.

The paper is organizes as follows. In Section 2, we present the relation between loss functions and uncertainty sets. In Section 3, we propose a kernel-based learning algorithm using uncertainty sets. Section 4 is devoted to study the statistical properties of the proposed algorithm. Section 5 is the concluding remarks.

We summarize some notations to be used throughout the paper. For a set S in a linear space, the convex-hull of S is denoted as $\mathrm{conv}S$ or $\mathrm{conv}(S)$. For a finite set S, the cardinality of S is denoted as $|S|$. The expectation of the random variable Z is described as $\mathbb{E}[Z]$. The set of all measurable functions on $\mathcal{X}$ is denoted by L_0. The supremum norm of $f \in L_0$ is denoted as $\|f\|_\infty$. For the reproducing kernel Hilbert space $\mathcal{H}$, $\|f\|_{\mathcal{H}}$ is the norm of $f \in \mathcal{H}$ defined from the inner product on $\mathcal{H}$.

2. Relation between loss functions and uncertainty sets

We study the relation between the loss function and the uncertainty set.

2.1. From loss functions to uncertainty sets

Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex and non-decreasing function. For the training samples, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$, we propose a learning method in which the linear decision function, $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$, is estimated by solving

$$\inf_{\boldsymbol{w},b,\rho} -2\rho + \frac{1}{m}\sum_{i=1}^{m} \ell(\rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) \ \text{ subject to } \ \|\boldsymbol{w}\|^2 \le \lambda^2,\ b \in \mathbb{R},\ \rho \in \mathbb{R}. \tag{3}$$

The regularization effect is introduced as the constraint $\|\boldsymbol{w}\|^2 \le \lambda^2$, where λ is the regularization parameter which may depend on the sample size. The statistical learning using (3) is regarded as an extension of ν-SVM. To see this, we define $\ell(z) = \max\{2z/\nu, 0\}$. Let $\widehat{\boldsymbol{w}}, \widehat{b}, \widehat{\rho}$ be an optimal solution of (1) for a fixed $\nu \in (0, 1)$. By comparing the optimality conditions of (1) and (3), we can confirm that (3) with $\lambda = \|\widehat{\boldsymbol{w}}\|$ has the same optimal solution as ν-SVM.

In a similar way as ν-SVM, we derive the dual problem of (3), and obtain the uncertainty set associated with the loss function ℓ in (3). The detailed calculation is presented in Appendix A. We define the conjugate function of $\ell(z)$ as $\ell^*(x) = \sup_{z\in\mathbb{R}}\{xz - \ell(z)\}$, and the constraint $\boldsymbol{\alpha} \in \Delta$ denotes that the vector $\boldsymbol{\alpha}$ satisfies $\sum_{i\in M_+}\alpha_i = \sum_{i\in M_-}\alpha_i = 1, \alpha_i \ge 0$. For each binary label, we define the parametrized uncertainty set by

$$\mathcal{U}_\pm[c] = \left\{ \sum_{i\in M_\pm} \alpha_i\boldsymbol{x}_i : \boldsymbol{\alpha} \in \Delta,\ \frac{1}{m}\sum_{i\in M_\pm} \ell^*(m\alpha_i) \le c \right\} \subset \mathcal{X}, \quad c \in \mathbb{R}, \tag{4}$$

i.e., $\mathcal{U}_+[c]$ for $y = +1$ and $\mathcal{U}_-[c]$ for $y = -1$. Then, the dual problem of (3) is represented as

$$\inf_{c_p,c_n,\boldsymbol{z}_p,\boldsymbol{z}_n} c_p + c_n + \lambda\|\boldsymbol{z}_p - \boldsymbol{z}_n\| \quad \text{subject to } \ \boldsymbol{z}_p \in \mathcal{U}_+[c_p],\ \boldsymbol{z}_n \in \mathcal{U}_-[c_n],\ c_p, c_n \in \mathbb{R}. \tag{5}$$

For all feasible solutions, the uncertainty sets $\mathcal{U}_+[c_p]$ and $\mathcal{U}_-[c_n]$ are not empty. Let $\widehat{\boldsymbol{z}}_p$ and $\widehat{\boldsymbol{z}}_n$ be the optimal solution, then, the optimal solution of $\boldsymbol{w}$ in (3) is equal to $\widehat{\boldsymbol{w}} = \lambda(\widehat{\boldsymbol{z}}_p - \widehat{\boldsymbol{z}}_n)/\|\widehat{\boldsymbol{z}}_p - \widehat{\boldsymbol{z}}_n\|$ for $\widehat{\boldsymbol{z}}_p \neq \widehat{\boldsymbol{z}}_n$ and $\widehat{\boldsymbol{w}} = \boldsymbol{0}$ for $\widehat{\boldsymbol{z}}_p = \widehat{\boldsymbol{z}}_n$. The relation between the loss function and the uncertainty set is given by (4). The estimation of the bias term b is considered in Section 3.

Example 1 (Truncated quadratic loss) *Now consider $\ell(z) = (\max\{1+z, 0\})^2$. This loss function is used in L_2-SVM (Schölkopf and Smola, 2001). The conjugate function is $\ell^*(\alpha) = -\alpha + \alpha^2/4$ for $\alpha \geq 0$ and $\ell^*(\alpha) = \infty$ for $\alpha < 0$. We define $\bar{\boldsymbol{x}}_\pm$ and $\widehat{\Sigma}_\pm$ as the empirical mean and the empirical covariance matrix of the samples $\{\boldsymbol{x}_i : i \in M_\pm\}$, i.e., $\bar{\boldsymbol{x}}_\pm = \frac{1}{m_\pm}\sum_{i\in M_\pm} \boldsymbol{x}_i$ and $\widehat{\Sigma}_\pm = \frac{1}{m_\pm}\sum_{i\in M_\pm}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_\pm)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_\pm)^T$, where m_+ and m_- are defined as $m_\pm = |M_\pm|$. Suppose that $\widehat{\Sigma}_\pm$ is invertible. Then, the uncertainty set corresponding to the truncated quadratic loss is given as $\mathcal{U}_\pm[c] = \left\{\sum_{i\in M_\pm}\alpha_i\boldsymbol{x}_i : \boldsymbol{\alpha} \in \Delta, \sum_{i\in M_\pm}\alpha_i^2 \leq 4(c+1)/m\right\} = \left\{\boldsymbol{z} \in \mathrm{conv}\{\boldsymbol{x}_i : i \in M_\pm\} : (\boldsymbol{z} - \bar{\boldsymbol{x}}_\pm)^T\widehat{\Sigma}_\pm^{-1}(\boldsymbol{z} - \bar{\boldsymbol{x}}_\pm) \leq 4(c+1)m_\pm/m\right\}$. 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 $\boldsymbol{z} \in \mathrm{conv}\{\boldsymbol{x}_i : i \in M_\pm\}$ is not imposed therein.*

2.2. From uncertainty sets to loss functions

We derived parametrized uncertainty sets associated with convex loss functions. Inversely, if the uncertainty set is represented as the form of (4), there exists the corresponding loss function. In general, however, the problem (5) with general uncertainty set does not lead to the minimization problem of the expected loss function under the empirical distribution. This section is devoted to study a way of revising the uncertainty set so as to possess the corresponding loss function.

Suppose that the parametrized uncertainty sets are defined as

$$\mathcal{U}_\pm[c] = \left\{\sum_{i\in M_\pm}\alpha_i\boldsymbol{x}_i : L_\pm^*(\boldsymbol{\alpha}_\pm) \leq c\right\} \subset \mathcal{X}, \tag{6}$$

where L_+^* (L_-^*) is the conjugate of a convex function L_+ (L_-), and the arguments $\boldsymbol{\alpha}_+$ and $\boldsymbol{\alpha}_-$ are defined as $\boldsymbol{\alpha}_\pm = (\alpha_i)_{i\in M_\pm}$. In Example 1, the function $L_\pm^*(\boldsymbol{\alpha}_\pm) = \frac{m}{4}\sum_{i\in M_\pm}\alpha_i^2 - 1$ is employed with the constraint $\boldsymbol{\alpha} \in \Delta$. Here, we consider the following optimization problem,

$$\min_{c_p,c_n,\boldsymbol{z}_p,\boldsymbol{z}_n} c_p + c_n + \lambda\|\boldsymbol{z}_p - \boldsymbol{z}_n\| \ \text{ subject to } \ c_p, c_n \in \mathbb{R},$$
$$\boldsymbol{z}_p \in \mathcal{U}_+[c_p] \cap \mathrm{conv}\{\boldsymbol{x}_i : i \in M_+\},\ \boldsymbol{z}_n \in \mathcal{U}_-[c_n] \cap \mathrm{conv}\{\boldsymbol{x}_i : i \in M_-\}.$$

In the above problem, the constraint defined from the convex-hulls $\mathrm{conv}\{\boldsymbol{x}_i : i \in M_\pm\}$ is added, since the uncertainty set (4) has the same constraint. The dual formulation of the above problem is given as

$$\inf_{\boldsymbol{w},b,\rho,\boldsymbol{\xi}_p,\boldsymbol{\xi}_n} -2\rho + L_+(\boldsymbol{\xi}_+) + L_-(\boldsymbol{\xi}_-) \ \text{ subject to } \rho - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq \xi_i, \forall i,\ \|\boldsymbol{w}\|^2 \leq \lambda^2, \tag{7}$$

where $\boldsymbol{\xi}_+ = (\xi_i)_{i\in M_+}$ and $\boldsymbol{\xi}_- = (\xi_i)_{i\in M_-}$. The dual form implies that $L_\pm$ are regarded as the loss functions for the decision function on training samples. When $L_\pm$ are represented as the empirical mean of a loss function, we can use the standard theoretical tools to analyze the statistical properties of the learning algorithm.

To link the uncertainty set with the empirical loss minimization, we revise the uncertainty sets $\mathcal{U}_\pm[c]$ such that the function $L^*_\pm$ has the additive form. Let m_+ and m_- be $m_\pm = |M_\pm|$, and we define $m_\pm$-dimensional vectors $\mathbf{1}_\pm = (1, \ldots, 1)$ and $\mathbf{0}_\pm = (0, \ldots, 0)$.

For convex functions $L^*_\pm : \mathbb{R}^{m_\pm} \to \mathbb{R}$, we define $\bar{\ell}^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ by

$$\bar{\ell}^*(\alpha) = \begin{cases} L^*_+(\frac{\alpha}{m}\mathbf{1}_+) + L^*_-(\frac{\alpha}{m}\mathbf{1}_-) - L^*_+(\mathbf{0}_+) - L^*_-(\mathbf{0}_-) & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases} \tag{8}$$

Then, we define the revised uncertainty set $\bar{\mathcal{U}}_\pm[c]$ by

$$\bar{\mathcal{U}}_\pm[c] = \Big\{ \sum_{i \in M_\pm} \alpha_i \boldsymbol{x}_i : \boldsymbol{\alpha} \in \Delta, \frac{1}{m} \sum_{i \in M_\pm} \bar{\ell}^*(\alpha_i m) \leq c \Big\}. \tag{9}$$

The dual problem of (5) with $\mathcal{U}_\pm[c] = \bar{\mathcal{U}}_\pm[c]$ is given as

$$\inf_{\boldsymbol{w}, b, \rho, \boldsymbol{\xi}} -2\rho + \frac{1}{m} \sum_{i \in M} \bar{\ell}(\xi_i) \quad \text{subject to } \rho - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \leq \xi_i, \forall i, \ \|\boldsymbol{w}\|^2 \leq \lambda^2. \tag{10}$$

The revision of the uncertainty sets leads to the empirical mean of the revised loss function $\bar{\ell}$. When we study statistical properties of the estimator given by the optimal solution of (10), we can apply the standard theoretical tools, since the objective in the primal expression is described by the empirical mean of the revised loss functions.

We explain the reason why the revised uncertainty set is defined as the form of (9). When the function $L^*_+ + L^*_-$ is described in the additive form, the uncertainty set is kept unchanged by the revision (8). Indeed, if there exists a closed, convex, proper function $\ell : \mathbb{R} \to \mathbb{R}$ such that $\ell^*(0) = 0$, $\ell^*(\alpha) = \infty$ for $\alpha < 0$ and $L^*_+(\boldsymbol{\alpha}_+) + L^*_-(\boldsymbol{\alpha}_-) - L^*_+(\mathbf{0}_+) - L^*_-(\mathbf{0}_-) = \frac{1}{m}\sum_{i \in M} \ell^*(\alpha_i m)$ hold, we obtain $\bar{\ell} = \ell$. See Rockafellar (1970) for the definition of closed, proper function.

We consider the other representation of the uncertainty set. Suppose that the uncertainty set is defined by $\mathcal{U}_\pm[c] = \{\sum_{i \in M_\pm} \alpha_i \boldsymbol{x}_i : h^*_\pm\big(\sum_{i \in M_\pm} \alpha_i \boldsymbol{x}_i\big) \leq c\}$, where $h_\pm$ are convex functions on the input space $\mathcal{X}$. Let $\boldsymbol{\mu}_+$ (resp. $\boldsymbol{\mu}_-$) be the mean of the input vector $\boldsymbol{x}$ conditioned on the positive (resp. negative) label. We define $\bar{\ell}^*$ by

$$\bar{\ell}^*(\alpha) = \begin{cases} h^*_+(\alpha\frac{m_+}{m}\boldsymbol{\mu}_+) + h^*_-(\alpha\frac{m_-}{m}\boldsymbol{\mu}_-) - h^*_+(\mathbf{0}) - h^*_-(\mathbf{0}) & \alpha \geq 0, \\ \infty, & \alpha < 0. \end{cases} \tag{11}$$

and the revised uncertainty set is defined by (9) with the above $\bar{\ell}^*$. In Appendix B, we explain the reason why we employ the formula (11) for the revision of the uncertainty set.

We show an example to illustrate how the revision of the uncertainty set works.

Example 2 *We suppose that $\boldsymbol{\mu}_\pm$ are the mean vectors and $\Sigma_\pm$ are the covariance matrices of the input vector conditioned on each label. We define the uncertainty set by $\mathcal{U}_\pm[c] = \{z \in \mathrm{conv}\{\boldsymbol{x}_i : i \in M_\pm\} : (\boldsymbol{z} - \boldsymbol{\mu})^T \Sigma_\pm^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \leq c, \forall \boldsymbol{\mu} \in \mathcal{A}_\pm\}$, where $\mathcal{A}_\pm$ denotes an estimation error of the mean vector $\boldsymbol{\mu}_\pm$. For example, for a fixed radius $r > 0$, $\mathcal{A}_\pm$ is defined as $\mathcal{A}_\pm = \{\boldsymbol{\mu} \in \mathcal{X} : (\boldsymbol{\mu} - \boldsymbol{\mu}_\pm)^T \Sigma_\pm^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_\pm) \leq r^2\}$. The uncertainty set with estimation error is used by Lanckriet et al. (2003) in MPM. The above uncertainty sets will be useful, when the probability in*

the training phase is slightly different from that in the test phase. Brief calculation yields that $\mathcal{U}_{\pm}[c]$ *is represented by the level set of the convex function* $h^*_{\pm}(\boldsymbol{z}) = \max_{\boldsymbol{\mu}\in\mathcal{A}_{\pm}} (\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma_{\pm}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) = (\sqrt{(\boldsymbol{z}-\boldsymbol{\mu}_{\pm})^T\Sigma_{\pm}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_{\pm})}+r)^2$. *The revised uncertainty set* $\bar{\mathcal{U}}_{\pm}[c]$ *is defined by the function* $\bar{\ell}^*$ *derived from* $h^*_{\pm}$. *We suppose that* $\boldsymbol{\mu}_+ \neq \mathbf{0}$ *and* $\boldsymbol{\mu}_- = \mathbf{0}$ *hold. Let* $d = \sqrt{\boldsymbol{\mu}_+^T\Sigma_+^{-1}\boldsymbol{\mu}_+}$ *and* $h = r/d(>0)$. *Then, the corresponding loss function is given as* $\bar{\ell}(z) = \frac{md^2}{m_p}\, u\big(\frac{z}{d^2}\big)$, *where* $u(z)$ *as defined as* $u(z) = 0$ *for* $z \leq -2h-2$, $u(z) = (\frac{z}{2}+1+h)^20$ *for* $-2h-2 \leq z \leq -2h$, $u(z) = z+2h+1$ *for* $-2h \leq z \leq 2h$, *and* $u(z) = \frac{z^2}{4} + z(1-h) + (1+h)^2$ *for* $2h \leq z$. *When* $r = 0$ *holds,* $\bar{\ell}(z)$ *is reduced to the truncated quadratic function in Example 1. For the positive* r, $\bar{\ell}(z)$ *is linear around* $z = 0$. *By introducing the estimation error represented by* $\mathcal{A}_{\pm}$, *the penalty for the misclassification is reduced from quadratic to linear around the decision boundary, though the original uncertainty set* $\mathcal{U}_{\pm}[c]$ *does not correspond any loss function.*

3. Kernel-based learning algorithm using uncertainty set

Based on the argument in the previous section, we present a kernel variant of the minimum distance problem using parametrized uncertainty sets. Suppose that training samples $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{+1, -1\}$ are observed, where $\mathcal{X}$ is not necessarily a subset of the Euclidean space. We define the 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 the book written by Schölkopf and Smola (2001) for the details of the kernel methods in machine learning. We consider the estimator of the decision function with the form of $f(x) + b$, where $f \in \mathcal{H}$, $b \in \mathbb{R}$.

In Figure 1, we describe the learning algorithm. In the learning algorithm, training samples are divided into two disjoint subsets, T_1 and T_2. The main reason that we decompose the set of training samples into two subsets is to simplify the analysis of the learning algorithm. The training samples in T_1 are used for the estimation of the function part $f \in \mathcal{H}$ in the decision function. We solve the problem (12) which is a kernel variant of the problem (5). For the estimation of the bias term, the empirical 0-1 loss on the data set T_2 is minimized with respect to the one-dimensional parameter b.

In the kernel-based algorithm, the parametrized uncertainty set is defined as a convex subset of the convex-hull of $\{k(\cdot, x_i^{(1)}) : i \in M_{\pm}\}$ in $\mathcal{H}$. Moreover, we assume that $\mathcal{U}_{\pm}[c] \subset \mathcal{U}_{\pm}[c']$ holds for $c \leq c'$ such as (6). When the uncertainty sets involve some parameters to be estimated, a prior knowledge or additional samples independent of the training samples $T_1 \cup T_2$ is used for the estimation.

4. Statistical Properties of Kernel-based Learning Algorithm

In this section, we prove that the expected 0-1 loss of the estimator provided in Figure 1 converges to the Bayes risk $\mathcal{E}^*$, when the uncertainty set corresponds to a *classification-calibrated* loss function.

4.1. Definitions and assumptions

We derive the dual representation of the learning algorithm in Figure 1. For a convex function $\ell : \mathbb{R} \to \mathbb{R}$, let ℓ^* be the conjugate function of ℓ. Suppose that the uncertainty sets are described as

Inputs: Decompose the training samples into two disjoint subsets, $T_1 = \{(x_i^{(1)}, y_i^{(1)}) : i = 1, \ldots, m_1\}$ and $T_2 = \{(x_i^{(2)}, y_i^{(2)}) : i = 1, \ldots, m_2\}$. For the set of training samples T_1, let M_+ and M_- be the index sets defined by $M_\pm = \{i : y_i^{(1)} = \pm 1, i = 1, \ldots, m_1\}$. We define the RKHS $\mathcal{H}$ with the kernel function $k(x, x')$. Prepare the parametrized uncertainty sets $\mathcal{U}_\pm[c]$ in $\mathcal{H}$ such that $\mathcal{U}_\pm[c] \subset \mathrm{conv}\{k(\cdot, x_i^{(1)}) : i \in M_\pm\}$. Set a regularization parameter $\lambda > 0$.

Step 1. Solve the optimization problem,

$$\inf_{\substack{c_p, c_n \\ f_p, f_n}} c_p + c_n + \lambda \|f_p - f_n\|_{\mathcal{H}} \text{ subject to } f_p \in \mathcal{U}_+[c_p], f_n \in \mathcal{U}_-[c_n], c_p, c_n \in \mathbb{R}. \quad (12)$$

Let $\widehat{f}_p$ and $\widehat{f}_n$ be optimal solutions of f_p and f_n. Define $\widehat{f}$ by $\widehat{f} = \lambda(\widehat{f}_p - \widehat{f}_n)/\|\widehat{f}_p - \widehat{f}_n\|_{\mathcal{H}}$ for $\widehat{f}_p \neq \widehat{f}_n$ and $\widehat{f} = 0$ for $\widehat{f}_p = \widehat{f}_n$.

Step 2. Solve the one-dimensional optimization problem with respect to the bias term, $\min_{b \in \mathbb{R}} \frac{1}{m_2} \sum_{i=1}^{m_2} [\![y_i^{(2)} (\widehat{f}(x_i^{(2)}) + b) \leq 0]\!]$, which is defined from the estimator $\widehat{f}$ and the data set T_2. The optimal solution is denoted as $\widetilde{b}$.

Output. The estimator of the decision function is given by $\widehat{f}(x) + \widetilde{b}$.

Figure 1: Kernel-based learning algorithm using uncertainty sets.

the form of

$$\mathcal{U}_\pm[c] = \left\{ \sum_{i \in M_\pm} \alpha_i k(\cdot, x_i^{(1)}) \in \mathcal{H} : \boldsymbol{\alpha} \in \Delta, \frac{1}{m} \sum_{i \in M_\pm} \ell^*(m\alpha_i) \leq c \right\}. \quad (13)$$

We can obtain the uncertainty set of the form (13) by applying the revision method proposed in Section 2.2. As shown in Appendix C, we find that the dual representation of (12) with the uncertainty set (13) is given as

$$\min_{f, b, \rho} -2\rho + \frac{1}{m_1} \sum_{i=1}^{m_1} \ell(\rho - y_i^{(1)}(f(x_i^{(1)}) + b)) \text{ subject to } f \in \mathcal{H}, b \in \mathbb{R}, \rho \in \mathbb{R}, \|f\|_{\mathcal{H}}^2 \leq \lambda^2. \quad (14)$$

We define some notations. Let $\widehat{f}, \widehat{b}$ and $\widehat{\rho}$ be an optimal solution of (14). Note that $\widehat{f}$ is obtained from the dual problem as shown in Step 1 of Figure 1. For a measurable function $f : \mathcal{X} \to \mathbb{R}$ 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 $\mathcal{R}(f, \rho) = -2\rho + \mathbb{E}[\ell(\rho - yf(x))]$, $\mathcal{R}_\lambda(f, \rho) = -2\rho + \mathbb{E}[\ell(\rho - yf(x))] + \theta(\|f\|_{\mathcal{H}}^2 \leq \lambda^2)$, where λ is a positive number and $\theta(A)$ equals 0 when A is true and ∞ otherwise. Let $\mathcal{R}^*$ be the infimum of $\mathcal{R}(f, \rho)$, i.e., $\mathcal{R}^* = \inf\{\mathcal{R}(f, \rho) : f \in L_0, \rho \in \mathbb{R}\}$. 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 $\widehat{\mathcal{R}}_T(f, \rho) = -2\rho + \frac{1}{m} \sum_{i=1}^m \ell(\rho - y_i f(x_i))$, and $\widehat{\mathcal{R}}_{T,\lambda}(f, \rho) = -2\rho + \frac{1}{m} \sum_{i=1}^m \ell(\rho - y_i f(x_i)) + \theta(\|f\|_{\mathcal{H}}^2 \leq \lambda^2)$, respectively. For the observed training samples

$T_1 = \{(x_i^{(1)}, y_i^{(1)}) : i = 1, \ldots, , m_1\}$, clearly the problem (14) is identical to the minimization of $\widehat{\mathcal{R}}_{T_1,\lambda}(f, \rho)$. For the index sets M_+ and M_- in Figure 1, we define $m_\pm = |M_\pm|$.

We introduce the following 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, i.e., 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) *There exists a positive constant $\varepsilon > 0$ such that $P(\{x \in \mathcal{X} : \varepsilon \leq P(+1|x) \leq 1 - \varepsilon\}) > 0$ holds, where $P(y|x)$ is the conditional probability of the label y for given input x.*

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

1. *ℓ is a non-decreasing, convex function, and satisfies the non-negativity condition, i.e., $\ell(z) \geq 0$ for all $z \in \mathbb{R}$. In addition, $\ell(z)$ is not a constant function, i.e., $\lim_{z\to\infty} \ell(z) = \infty$ holds.*
2. *Let $\partial\ell(z)$ be the subdifferential of the loss function ℓ at $z \in \mathbb{R}$ (Rockafellar, 1970, Chap. 23). For any $M > 0$, there exists z_0 such that for all $z \geq z_0$ and all $g \in \partial\ell(z)$, the inequality $g \geq M$ holds. In other word, $\lim_{z\to\infty} \partial\ell(z) = \infty$ holds.*

The hinge loss $\ell(z) = \max\{z, 0\}$ used in ν-SVM and the logistic loss $\ell(z) = \log(1 + e^z)$ do not satisfy the basic assumption above, since the derivative does not go to infinity. On the other hand, the truncated quadratic loss and the exponential loss meet the basic assumption.

Assumption 4 (modified classification-caliblated loss)

1. *$\ell(z)$ is first order differentiable for $z \geq -\ell(0)/2$, and $\ell'(z) > 0$ holds for $z \geq -\ell(0)/2$.*
2. *Let $\psi(\theta, \rho)$ be the function defined as $\psi(\theta, \rho) = \ell(\rho) - \inf_{z\in\mathbb{R}} \left\{ \frac{1+\theta}{2}\ell(\rho - z) + \frac{1-\theta}{2}\ell(\rho + z) \right\}$ for $0 \leq \theta \leq 1, \rho \in \mathbb{R}$. There exist a function $\widetilde{\psi}(\theta)$ and a positive real $\varepsilon > 0$ such that the following conditions are satisfied: (a) $\widetilde{\psi}(0) = 0$ and $\widetilde{\psi}(\theta) > 0$ for $0 < \theta \leq \varepsilon$. (b) $\widetilde{\psi}(\theta)$ is 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 $0 \leq \theta \leq \varepsilon$.*

In Section 4.3, we shall give some sufficient conditions for existence of the function $\widetilde{\psi}$ in Assumption 4.

In the following, we prove the convergence of the error rate to the Bayes risk $\mathcal{E}^*$. The proof consists of two parts. In Section 4.2, 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 $\mathcal{R}^*$. Here, we apply the technique developed by Steinwart (2005). Then, we prove the convergence of the error rate $\mathcal{E}(\widehat{f} + \widetilde{b})$ to the Bayes risk $\mathcal{E}^*$.

In this proof, the concept of the classification-calibrated loss (Bartlett et al., 2006) plays an important role.

4.2. Convergence to Bayes Risk

In Appendix D, we prove that $\lim_{\lambda\to\infty}\inf\{\mathcal{R}_\lambda(f,\rho) : f\in\mathcal{H},\, \rho\in\mathbb{R}\} = \mathcal{R}^* > -\infty$ holds under Assumption 1, 2 and 3. We derive an upper bound of the norm of optimal solutions.

Lemma 1 *Let λ_{m_1} be the regularization parameter depending on m_1. Under Assumption 1, 2 and 3, there are positive constants c, C and a natural number M such that the optimal solutions $\widehat{f},\widehat{b}$ and $\widehat{\rho}$ satisfy*

$$\|\widehat{f}\|_{\mathcal{H}} \le \lambda_{m_1}, \quad |\widehat{b}| \le C\lambda_{m_1}, \quad |\widehat{\rho}| \le C\lambda_{m_1} \tag{15}$$

with the probability greater than $1-e^{-cm_1}$ for $m_1 \ge M$.

Proof We show an idea of the proof. A rigorous proof is shown in Appendix E. Comparing the objective value $\widehat{\mathcal{R}}_{T_1,\lambda_{m_1}}(f+b,\rho)$ at the optimal solution $(\widehat{f},\widehat{b},\widehat{\rho})$ and that at a feasible solution $(f,b,\rho) = (0,0,0)$, we have $\widehat{\rho} \ge -\ell(0)/2$. The optimality condition w.r.t. $\widehat{\rho}$ leads to $2 \in \frac{1}{m_1}\sum_{i=1}^{m_1}\partial\ell(\widehat{\rho}-y_i^{(1)}(\widehat{f}(x_i^{(1)})+\widehat{b})) \ge \frac{1}{m_1}\sum_{i=1}^{m_+}\partial\ell(\widehat{\rho}-\widehat{b}-K\lambda_{m_1})+\frac{1}{m_1}\sum_{i=1}^{m_-}\partial\ell(\widehat{\rho}+\widehat{b}-K\lambda_{m_1})$. The inequalities above and the monotonicity of the subdifferential lead to the fact that there exists a constant $\bar{z}$ such that $|\widehat{\rho}| \le K\lambda_{m_1}+\bar{z}$ and $|\widehat{b}| \le K\lambda_{m_1}+\bar{z}$ hold with high probability. Here, $\bar{z}$ is determined from the marginal probability $P(Y=\pm1)$ and the loss function ℓ. ■

Let us define the covering number of a metric space.

Definition 2 (covering number) *For a metric space $\mathcal{G}$, the covering number of $\mathcal{G}$ is defined as $\mathcal{N}(\mathcal{G},\varepsilon) = \min\{n\in\mathbb{N} : g_1,\dots,g_n \in \mathcal{G}$ such that $\mathcal{G}\subset\bigcup_{i=1}^n B(g_i,\varepsilon)\}$, where $B(g,\varepsilon)$ denotes the closed ball with center g and radius ε.*

Due to Lemma 1, we see that the optimal solution, $(\widehat{f},\widehat{b},\widehat{\rho})$, is included in the set $\mathcal{G}_{m_1} = \{(f,b,\rho)\in\mathcal{H}\times\mathbb{R}^2 : \|f\|_{\mathcal{H}} \le \lambda_{m_1}, |b| \le C\lambda_{m_1}, |\rho| \le C\lambda_{m_1}\}$ with high probability. Suppose that the norm $\|f\|_\infty + |b| + |\rho|$ is introduced on $\mathcal{G}_{m_1}$. We define the function $L(x,y;f,b,\rho) = -2\rho + \ell(\rho - y(f(x)+b))$, and the function set $\mathcal{L}_{m_1} = \{L(x,y;f,b,\rho) : (f,b,\rho)\in\mathcal{G}_{m_1}\}$. Since $\ell : \mathbb{R}\to\mathbb{R}$ is a finite-valued convex function, ℓ is locally Lipschitz continuous. Then, for any sample size m_1, there exists a constant κ_{m_1} depending on m_1 such that $|\ell(z)-\ell(z')| \le \kappa_{m_1}|z-z'|$ holds for all z and z' satisfying $|z|,|z'| \le (K+2C)\lambda_{m_1}$. Then, for any $(f,b,\rho),(f',b',\rho')\in\mathcal{G}_{m_1}$, we have $|L(x,y;f,b,\rho) - L(x,y;f',b',\rho')| \le 2|\rho-\rho'| + \kappa_{m_1}(|\rho-\rho'| + |b-b'| + \|f-f'\|_\infty) \le (2+\kappa_{m_1})(|\rho-\rho'| + |b-b'| + \|f-f'\|_\infty)$. The covering number of $\mathcal{L}_{m_1}$ is evaluated by $\mathcal{N}(\mathcal{L}_{m_1},\varepsilon) \le \mathcal{N}\big(\mathcal{G}_{m_1}, \frac{\varepsilon}{2+\kappa_{m_1}}\big)$, in which the supremum norm is defined on $\mathcal{L}_{m_1}$. Let the metric space $\mathcal{F}_{m_1}$ be $\mathcal{F}_{m_1} = \{f\in\mathcal{H} : \|f\|_{\mathcal{H}} \le \lambda_{m_1}\}$ endowed with the supremum norm, then, we also have $\mathcal{N}\big(\mathcal{G}_{m_1}, \frac{\varepsilon}{2+\kappa_{m_1}}\big) \le \mathcal{N}\big(\mathcal{F}_{m_1}, \frac{\varepsilon}{3(2+\kappa_{m_1})}\big)\big(\frac{6C\lambda_{m_1}(2+\kappa_{m_1})}{\varepsilon}\big)^2$. An upper bound of the covering number of $\mathcal{F}_{m_1}$ is given by Cucker and Smale (2002) and Zhou (2002).

Lemma 3 *Let b_{m_1} be $b_{m_1} = 4C\lambda_{m_1} + \ell((K+2C)\lambda_{m_1})$ in which C is the positive constant defined in Lemma 1. Under Assumption 1 and 3, the following inequality holds:*

$$\begin{aligned} &P\left(\sup_{(f,b,\rho)\in\mathcal{G}_{m_1}} |\widehat{\mathcal{R}}(f+b,\rho) - \mathcal{R}(f+b,\rho)| \ge \varepsilon\right) \\ &\le 2\mathcal{N}\left(\mathcal{F}_{m_1}, \frac{\varepsilon}{9(2+\kappa_{m_1})}\right)\left(\frac{18C\lambda_{m_1}(2+\kappa_{m_1})}{\varepsilon}\right)^2 \exp\left\{-\frac{2m_1\varepsilon^2}{9b_{m_1}^2}\right\}. \end{aligned} \tag{16}$$

Proof We show an idea of the proof. Note that $\|f\|_\infty \leq K\lambda_{m_1}$ holds for $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} \leq \lambda_{m_1}$. A brief calculation yields that $\sup_{\substack{(x,y)\in\mathcal{X}\times\{+1,-1\}\\(f,b,\rho)\in\mathcal{G}_{m_1}}} L(x,y;f,b,\rho) - \inf_{\substack{(x,y)\in\mathcal{X}\times\{+1,-1\}\\(f,b,\rho)\in\mathcal{G}_{m_1}}} L(x,y;f,b,\rho) \leq b_{m_1}$. In the same way as the proof of Lemma 3.4 in Steinwart (2005), the upper bound is derived from Hoeffding's inequality and the inequality $\mathcal{N}\big(\mathcal{G}_{m_1}, \frac{\varepsilon}{2+\kappa_{m_1}}\big) \leq \mathcal{N}\big(\mathcal{F}_{m_1}, \frac{\varepsilon}{3(2+\kappa_{m_1})}\big)\big(\frac{6C\lambda_{m_1}(2+\kappa_{m_1})}{\varepsilon}\big)^2$. ■

We present the main theorem of this section.

Theorem 4 *We suppose that the regularization parameter $\lambda = \lambda_{m_1}$ satisfies $\lim_{m_1\to\infty}\lambda_{m_1} = \infty$, and that Assumption 1, 2 and 3 hold. Moreover we assume that (16) converges to zero for any $\varepsilon > 0$, when the sample size m_1 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_1.*

Proof We show a sketch of the proof. A rigorous proof is shown in Appendix F. Now, we have the convergence $\inf_{f,b,\rho}\mathcal{R}_{\lambda_{m_1}}(f+b,\rho) \to \mathcal{R}^*$ and the uniform convergence

$$\sup_{(f,b,\rho)\in\mathcal{G}_{m_1}} |\widehat{\mathcal{R}}_{T_1}(f+b,\rho) - \mathcal{R}(f+b,\rho)| \to 0$$

in probability, when m_1 tends to infinity. We apply the standard argument on the uniform convergence, we obtain the probabilistic convergence of $\mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho})$ to $\mathcal{R}^*$. ■

We show the order of λ_{m_1} admitting the assumption in Theorem 4.

Example 3 *The Gaussian kernel is universal on $\mathcal{X} = [0,1]^n \subset \mathbb{R}^n$; see Corollary 4.58 of Steinwart and Christmann (2008). According to Zhou (2002), the covering number of the Gaussian RKHS is $\log\mathcal{N}(\mathcal{F}_{m_1},\varepsilon/(18+9\kappa_{m_1})) = O\big(\big(\log(\lambda_{m_1}\kappa_{m_1})\big)^{n+1}\big)$. For any $\varepsilon > 0$, (16) is bounded above by $\exp\{O(-m_1/b_{m_1}^2 + (\log(\lambda_{m_1}\kappa_{m_1}))^{n+1})\}$. For the truncated quadratic loss, we have $\kappa_{m_1} \leq 2((K+2C)\lambda_{m_1}+1) = O(\lambda_{m_1})$ and $b_{m_1} \leq 4C\lambda_{m_1} + ((K+2C)\lambda_{m_1}+1)^2 = O(\lambda_{m_1}^2)$. Let us define $\lambda_{m_1} = m_1^\alpha$ with $0 < \alpha < 1/4$. Then, for any $\varepsilon > 0$, (16) converges to zero when m_1 tends to infinity. In the same way, for the exponential loss we obtain $\kappa_{m_1} = O(e^{(K+2C)\lambda_{m_1}})$ and $b_{m_1} = O(e^{(K+2C)\lambda_{m_1}})$. Hence, $\lambda_{m_1} = (\log m_1)^\alpha$ with $0 < \alpha < 1$ ensures the convergence of (16).*

In this section, we prove that the expected 0-1 loss $\mathcal{E}(\widehat{f}+\widetilde{b})$ converges to the Bayes risk $\mathcal{E}^*$ in the large sample limit. The proof also ensures the convergence of $\mathcal{E}(\widehat{f}+\widehat{b})$ to the Bayes risk. Hence, if the explicit form of the loss function $\ell(z)$ is obtained from the uncertainty set, solving (14) can be another promising method for classification problems.

Theorem 5 *Suppose that $\mathcal{R}(\widehat{f}+\widehat{b},\widehat{\rho})$ converges to $\mathcal{R}^*$ in probability, when the sample size of T_1, i.e., m_1, tends to infinity. For the RKHS $\mathcal{H}$ and the loss function ℓ, we assume Assumption 1, 3 and 4. Then, $\mathcal{E}(\widehat{f}+\widetilde{b})$ converges to $\mathcal{E}^*$ in probability, when the sample sizes of T_1 and T_2 tend to infinity.*

A rigorous proof of Theorem 5 is shown in Appendix G. As a result, we find that the prediction error rate of $\widehat{f}+\widetilde{b}$ converges to the Bayes risk under Assumption 1, 2, 3 and 4.

4.3. Sufficient Conditions of Modified Classification-calibrated Loss

We present some sufficient conditions for existence of the function $\widetilde{\psi}$ in Assumption 4. The proofs of the following lemmas are presented in Appendix H.

Lemma 6 *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. We assume the following conditions: (a) $d < -\ell(0)/2$; (b) $\ell(z)$ is second-order continuously differentiable on the open interval (d, ∞); (c) $\ell''(z) > 0$ holds on (d, ∞); (d) $1/\ell'(z)$ 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 conditions in Lemma 6 are 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. The lemma above works for the truncated quadratic loss $\ell(z) = (\max\{1 + z, 0\})^2$ and the exponential loss $\ell(z) = e^z$. See Example 4 and Example 5.

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

Lemma 7 *Suppose that the first condition in Assumption 3 and the first condition in Assumption 4 hold. Let d be $d = \sup\{z \in \mathbb{R} : \partial\ell(z) = \{0\}\}$. 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 $\xi(z, \rho) = \{\ell(\rho + z) + \ell(\rho - z) - 2\ell(\rho)\}/(z\ell'(\rho))$ for $z > 0$ and $\xi(z, \rho) = 0$ for $z = 0$. Suppose that there exists a function $\bar{\xi}(z)$ for $z \geq 0$ such that the following conditions hold: (a) $\bar{\xi}(z)$ is continuous and strictly increasing on $z \geq 0$, and satisfies $\bar{\xi}(0) = 0$ and $\lim_{z\to\infty} \bar{\xi}(z) > 1$; (b) $\sup_{\rho \geq -\ell(0)/2} \xi(z, \rho) \leq \bar{\xi}(z)$ holds. Then, there exists a function $\widetilde{\psi}$ defined in the second condition of Assumption 4.*

Note that Lemma 7 does not require the second order differentiability of the loss function.

Example 4 *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 6 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 5 *For the exponential loss $\ell(z) = e^z$, we have $1/\ell'(z) = e^{-z}$. Hence, due to Lemma 6, $\psi(\theta, \rho)$ is non-decreasing in ρ. Indeed, we have $\psi(\theta, \rho) = (1 - \sqrt{1 - \theta^2})e^{\rho}$.*

Example 6 *In Example 2, we presented the uncertainty set with estimation errors. We define $\bar{\ell}^*(\alpha)$ by $\bar{\ell}^*(\alpha) = (|\alpha w - 1| + h)^2 - (1 + h)^2$ for $\alpha \geq 0$ and $\bar{\ell}^*(\alpha) = \infty$ for $\alpha < 0$, where w and h are positive constants. Then, the revised uncertainty set is described by $\bar{\ell}^*$. Here, we suppose $w > 1/2$. For the function $\bar{\ell}^*$ defined above, the corresponding loss function is given as $\bar{\ell}(z) = u(z/w)$, where $u(z)$ is equal to the function $u_+(z)$ with $h_+ = h$ defined in Example 2. 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 around $z = 0$, Lemma 6 does not work. Hence, we apply Lemma 7. A simple calculation yields that $\bar{\ell}'(-\bar{\ell}(0)/2) \geq (4w - 1)/(4w^2) > 0$ holds 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 $\xi(z, \rho) = \frac{1}{\bar{\ell}'(\rho)}\left(\frac{\bar{\ell}(\rho+z)-\bar{\ell}(\rho)}{z} - \frac{\bar{\ell}(\rho)-\bar{\ell}(\rho-z)}{z}\right) \leq \frac{\bar{\ell}'(\rho+z)-\bar{\ell}'(\rho-z)}{\bar{\ell}'(\rho)}$. 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 conditions in Lemma 7. Hence, the loss function corresponding to the revised uncertainty set satisfies the conditions for statistical consistency, though the original uncertainty set with the estimation error does not correspond to the empirical mean of a loss function.

5. Conclusion

In this paper, we studied the relation between the loss function approach and the minimum distance approach in binary classification problems. We proposed the learning algorithm based on the revised minimum distance problem, and proved the statistical consistency. In our proof, the hinge loss used in ν-SVM is excluded, though Steinwart (2003) proved the statistical consistency of ν-SVM with a nice choice of the regularization parameter. A future work is to relax the assumptions of our theoretical result so as to include the hinge loss function and other popular loss functions such as the logistic loss. Also, it is important to derive the convergence rate of the proposed learning method. Developing an optimization algorithm is needed for practical data analysis by the statistical learning with uncertainty sets.

References

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.

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.

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

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.

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.

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

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

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 v-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.

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

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

Appendix A. Derivation of (5)

We introduce the slack variables $\xi_i, i = 1, \ldots, m$ satisfying the inequalities $\xi_i \geq \rho - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)$, $i = 1, \ldots, m$. The Lagrangian function of the problem (3) is given as

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

where $\alpha_1, \ldots, \alpha_m$ and μ are the non-negative Lagrange multipliers. The optimality conditions,

$$\frac{\partial L}{\partial \rho} = 0, \quad \frac{\partial L}{\partial b} = 0,$$

and the non-negativity of α_i lead to the constraint on Lagrange multipliers, $\sum_{i \in M_+} \alpha_i = \sum_{i \in M_-} \alpha_i = 1, \alpha_i \geq 0$. In the following, the constraints $\sum_{i \in M_+} \alpha_i = \sum_{i \in M_-} \alpha_i = 1, \alpha_i \geq 0$ are denoted by $\boldsymbol{\alpha} \in \Delta$. We define the conjugate function of $\ell(z)$ as $\ell^*(x) = \sup_{z \in \mathbb{R}}\{xz - \ell(z)\}$. Then, the min-max theorem yields the dual problem of (3),

$$\begin{aligned}
&\sup_{\boldsymbol{\alpha} \geq \boldsymbol{0}, \mu \geq 0} \inf_{\boldsymbol{w}, b, \rho, \boldsymbol{\xi}} L(\boldsymbol{w}, b, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mu) \\
&= -\inf_{\boldsymbol{\alpha}, \mu \geq 0} \sup_{\boldsymbol{w}, \boldsymbol{\xi}} \left\{ \frac{1}{m}\sum_{i=1}^{m}(m\alpha_i\xi_i - \ell(\xi_i)) + \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{w} - \mu(\|\boldsymbol{w}\|^2 - \lambda^2) : \boldsymbol{\alpha} \in \Delta \right\} \\
&= -\inf_{\boldsymbol{\alpha}} \left\{ \frac{1}{m}\sum_{i=1}^{m} \ell^*(m\alpha_i) + \lambda \Big\| \sum_{i \in M_+} \alpha_i \boldsymbol{x}_i - \sum_{i \in M_-} \alpha_i \boldsymbol{x}_i \Big\| : \boldsymbol{\alpha} \in \Delta \right\} \\
&= -\inf_{\boldsymbol{\alpha}, c_p, c_n} \Bigg\{ c_p + c_n + \lambda \Big\| \sum_{i \in M_+} \alpha_i \boldsymbol{x}_i - \sum_{i \in M_-} \alpha_i \boldsymbol{x}_i \Big\| \\
&\qquad : \boldsymbol{\alpha} \in \Delta, \frac{1}{m}\sum_{i \in M_+} \ell^*(m\alpha_i) \leq c_p, \frac{1}{m}\sum_{i \in M_-} \ell^*(m\alpha_i) \leq c_n \Bigg\}. \qquad (17)
\end{aligned}$$

By using the uncertainty set (4), the problem (17) is represented as (5). In Appendix C, we present a rigorous proof that under some assumptions on the loss function ℓ, the min-max theorem works in the above Lagrangian function, i.e., there is no duality gap.

Appendix B. Revision of Uncertainty Sets

We explain a validity of the formula (11). We want to find a function $\bar{\ell}^*(\alpha)$ such that $h_+^*(\sum_{i\in M_+}\alpha_i\boldsymbol{x}_i)+h_-^*(\sum_{i\in M_-}\alpha_i\boldsymbol{x}_i)-h_+^*(\mathbf{0})-h_-^*(\mathbf{0})$ is close to $\frac{1}{m}\sum_{i=1}^m\bar{\ell}^*(m\alpha_i)$ in some sense. We substitute $\alpha_i=\alpha/m$ into $h_\pm^*(\sum_{i\in M_\pm}\alpha_i\boldsymbol{x}_i)$. In the large sample limit, $h_\pm^*(\sum_{i\in M_\pm}\alpha/m\,\boldsymbol{x}_i)$ is approximated by $h_\pm^*(\alpha\frac{m_\pm}{m}\boldsymbol{\mu}_\pm)$. Suppose that $h_+^*(\alpha\frac{m_+}{m}\boldsymbol{\mu}_+)+h_-^*(\alpha\frac{m_-}{m}\boldsymbol{\mu}_-)-h_+^*(\mathbf{0})-h_+^*(\mathbf{0})$ is represented as $\frac{1}{m}\sum_{i=1}^m\bar{\ell}^*(\frac{\alpha}{m}m)=\bar{\ell}^*(\alpha)$. Then, we obtain (11).

Appendix C. Duality between (12) and (14)

Lemma 8 *Suppose that $m_\pm=|M_\pm|$ are positive. Under Assumption 1 and 3, there exists an optimal solution of (14). Moreover, the dual problem of (14) yields the problem (12) with the uncertainty set (13).*

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

$$f(x)=\sum_{j=1}^{m_1}\alpha_j k(x,x_j^{(1)}).$$

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

$$\begin{aligned}&\min_{\boldsymbol{\alpha},b,\rho}\ -2\rho+\frac{1}{m_1}\sum_{i=1}^{m_1}\ell\Bigg(\rho-y_i^{(1)}\Bigg(\sum_{j=1}^{m_1}\alpha_j k(x_i^{(1)},x_j^{(1)})+b\Bigg)\Bigg)\\&\text{subject to}\ \sum_{i,j=1}^{m_1}\alpha_i\alpha_j k(x_i^{(1)},x_j^{(1)})\le\lambda^2.\end{aligned}\tag{18}$$

Let $\zeta_0(\boldsymbol{\alpha},b,\rho)$ be the objective function of (18). Let us define $\mathcal{S}$ be the linear subspace in $\mathbb{R}^{m_1}$ spanned by the column vectors of the gram matrix $(k(x_i^{(1)},x_j^{(1)}))_{i,j=1}^{m_1}$. We can impose the constraint $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{m_1})\in\mathcal{S}$, since the orthogonal complement of $\mathcal{S}$ does not affect the objective and the constraints in (18). We see that Assumption 1 and the reproducing property yield the inequality $\|y_i^{(1)}\sum_{i=1}^{m_1}\alpha_i k(\cdot,x_i^{(1)})\|_\infty\le K\lambda$. Due to this inequality and the assumptions on the function ℓ, the objective function $\zeta_0(\boldsymbol{\alpha},b,\rho)$ is bounded below by

$$\zeta_1(b,\rho)=-2\rho+\frac{m_+}{m_1}\ell(\rho-b-K\lambda)+\frac{m_-}{m_1}\ell(\rho+b-K\lambda).\tag{19}$$

Hence, for any real number c, the inclusion relation

$$\begin{aligned}&\left\{(\boldsymbol{\alpha},b,\rho)\ :\ \zeta_0(\boldsymbol{\alpha},b,\rho)\le c,\ \sum_{i,j=1}^{m_1}\alpha_i\alpha_j k(x_i^{(1)},x_j^{(1)})\le\lambda^2,\ \boldsymbol{\alpha}\in\mathcal{S}\right\}\\&\subset\left\{(\boldsymbol{\alpha},b,\rho)\ :\ \zeta_1(b,\rho)\le c,\ \sum_{i,j=1}^{m_1}\alpha_i\alpha_j k(x_i^{(1)},x_j^{(1)})\le\lambda^2,\ \boldsymbol{\alpha}\in\mathcal{S}\right\}\end{aligned}\tag{20}$$

holds. Note that the vector $\boldsymbol{\alpha}$ satisfying $\sum_{i,j=1}^{m_1} \alpha_i \alpha_j k(x_i^{(1)}, x_j^{(1)}) \leq \lambda^2$ and $\boldsymbol{\alpha} \in \mathcal{S}$ is restricted to a compact subset in $\mathbb{R}^{m_1}$, since the gram matrix $(k(x_i^{(1)}, x_j^{(1)}))_{i,j=1}^{m}$ is positive definite on the subspace $\mathcal{S}$. We shall prove that the subset (20) 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 the convex function $\zeta_2(u_1, u_2)$ defined by

$$\zeta_2(u_1, u_2) = -u_1 + \frac{m_+}{m_1}\ell(u_1 - K\lambda) - u_2 + \frac{m_-}{m_1}\ell(u_2 - K\lambda).$$

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

$$\lim_{|u_1| \to \infty} -u_1 + \frac{m_+}{m_1}\ell(u_1 - K\lambda) = \infty.$$

The same limit holds for $-u_2 + \frac{m_-}{m_1}\ell(u_2 - K\lambda)$. Hence, the level set of $\zeta_2(u_1, u_2)$ is closed and bounded, i.e., compact. As a result, the level set of $\zeta_1(b, \rho)$ is also compact. Therefore, the subset (20) is also compact in $\mathbb{R}^{m_1+2}$. This implies that (18) has an optimal solution.

Next, we prove the duality between (12) and (14). Since (18) has an optimal solution, the problem using the slack variables $\xi_i, i = 1, \ldots, m_1$,

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, b, \rho, \boldsymbol{\xi}} -2\rho + \frac{1}{m_1}\sum_{i=1}^{m_1} \ell(\xi_i) \\ &\text{subject to } \sum_{i,j=1}^{m_1} \alpha_i \alpha_j k(x_i^{(1)}, x_j^{(1)}) \leq \lambda^2,\ \rho - y_i^{(1)}(\sum_{j=1}^{m_1} \alpha_i k(x_i^{(1)}, x_j^{(1)}) + b) \leq \xi_i,\ i = 1, \ldots, m_1. \end{aligned}$$

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 6.2.4). Indeed, at the feasible solution, $\boldsymbol{\alpha} = \mathbf{0}, b = 0, \rho = 0$ and $\xi_i = 1, i = 1, \ldots, m_1$, 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, i.e., there is no duality gap. ■

Appendix D. Convergence of Expected Loss

Lemma 9 *Under Assumption 2 and Assumption 3, we have* $\mathcal{R}^* > -\infty$.

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

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

For given η satisfying $\varepsilon \leq \eta \leq 1-\varepsilon$, 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

$$\partial\xi(f,\rho) = \left\{ (0, -\frac{2}{P(s)})^T + u\eta(-1,1)^T + v(1-\eta)(1,1)^T : u \in \partial\ell(\rho - f),\ v \in \partial\ell(\rho + f) \right\}.$$

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 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

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

holds for all $f,\rho \in \mathbb{R}$ and all η such that $\varepsilon \leq \eta \leq 1-\varepsilon$. 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 10 *Under Assumption 1, 2 and 3, we have*

$$\lim_{\lambda\to\infty} \inf\{\mathcal{R}_\lambda(f,\rho) : f \in \mathcal{H},\ \rho \in \mathbb{R}\} = \mathcal{R}^*. \tag{21}$$

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

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

holds for any $\rho \in \mathbb{R}$. Thus, we have $\inf\{\mathcal{R}(f,\rho) : f \in \mathcal{H}\} = \inf\{\mathcal{R}(f,\rho) : f \in L_0\}$ for any $\rho \in \mathbb{R}$. Then, the equality

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

holds. Under Assumption 2 and Assumption 3, we have $\mathcal{R}^* > -\infty$ due to Lemma 9. Then, for any $\varepsilon > 0$, there exist $\lambda_\varepsilon > 0$, $f_\varepsilon \in \mathcal{H}$ and $\rho_\varepsilon \in \mathbb{R}$ such that $\|f_\varepsilon\|_{\mathcal{H}} \leq \lambda_\varepsilon$ and $\mathcal{R}(f_\varepsilon, \rho_\varepsilon) \leq \mathcal{R}^* + \varepsilon$ hold. For all $\lambda \geq \lambda_\varepsilon$ we have

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

On the other hand, it is clear that the inequality $\mathcal{R}^* \leq \inf\{\mathcal{R}_\lambda(f, \rho) : f \in \mathcal{H}, \rho \in \mathbb{R}\}$ holds. Hence, Eq.(21) holds. ■

Appendix E. Proof of Lemma 1

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

$$\frac{1}{2}P(Y = +1) < \frac{m_+}{m_1}, \quad \frac{1}{2}P(Y = -1) < \frac{m_-}{m_1} \tag{22}$$

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 (22) holds with the probability higher than $1 - e^{-cm_1}$.

Lemma 8 in Appendix C ensures that the problem (14) has optimal solutions $\widehat{f}, \widehat{b}, \widehat{\rho}$. The first inequality in (15), i.e., $\|\widehat{f}\|_{\mathcal{H}} \leq \lambda_{m_1}$, is clearly satisfied. Then, we have $\|\widehat{f}\|_\infty \leq K\lambda_{m_1}$ from the reproducing property of the RKHSs. The definition of the estimator and the non-negativity of ℓ yield that

$$-2\widehat{\rho} \leq -2\widehat{\rho} + \frac{1}{m_1}\sum_{i=1}^{m_1} \ell(\widehat{\rho} - y_i^{(1)}(\widehat{f}(x_i^{(1)}) + \widehat{b})) \leq \ell(0).$$

Then, we have

$$\widehat{\rho} \geq -\frac{\ell(0)}{2}. \tag{23}$$

Next, we consider the optimality condition of the problem (14). The Lagrangian of the optimization problems is given as

$$L(f, b, \rho, \mu) = -2\rho + \frac{1}{m_1}\sum_{i=1}^{m_1} \ell(\rho - y_i^{(1)}(f(x_i^{(1)}) + b)) + \mu(\|f\|_{\mathcal{H}}^2 - \lambda_{m_1}^2),$$

where $\mu \geq 0$ is the Lagrange multiplier of the inequality constraint $\|f\|_{\mathcal{H}}^2 \leq \lambda_{m_1}^2$. According to the calculus of subdifferential introduced in Section 23 of Rockafellar (1970), the derivative of L with respect to ρ leads to an optimality condition,

$$0 \in -2 + \frac{1}{m_1}\sum_{i=1}^{m_1} \partial\ell(\widehat{\rho} - y_i^{(1)}(\widehat{f}(x_i^{(1)}) + \widehat{b})).$$

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

$$\begin{aligned} 2 &\geq \frac{1}{m_1}\sum_{i=1}^{m_1} \partial\ell(\widehat{\rho} - y_i^{(1)}\widehat{b} - K\lambda_{m_1}) \\ &= \frac{1}{m_1}\sum_{i=1}^{m_+} \partial\ell(\widehat{\rho} - \widehat{b} - K\lambda_{m_1}) + \frac{1}{m_1}\sum_{j=1}^{m_-} \partial\ell(\widehat{\rho} + \widehat{b} - K\lambda_{m_1}) \\ &\geq \frac{1}{m_1}\sum_{i=1}^{m_+} \partial\ell(\widehat{\rho} - \widehat{b} - K\lambda_{m_1}). \end{aligned}$$

In the above expressions, $\sum_{i=1}^{m_+}\partial\ell$ denotes the m_+-fold sum of the set $\partial\ell$. Let z_p be a real number satisfying $\frac{2m_1}{m_+} < \partial\ell(z_p)$, i.e., all elements in $\partial\ell(z_p)$ are greater than $\frac{2m_1}{m_+}$. Then, $\widehat{\rho} - \widehat{b} - K\lambda_{m_1}$ should be less than z_p. In the same way, for z_n satisfying $\frac{2m_1}{m_-} < \partial\ell(z_n)$, we have $\widehat{\rho} + \widehat{b} - K\lambda_{m_1} < z_n$. Hence, the inequalities

$$\begin{aligned} \widehat{\rho} &\leq K\lambda_{m_1} + \max\{z_p, z_n\}, \\ |\widehat{b}| &\leq \frac{\ell(0)}{2} + K\lambda_{m_1} + \max\{z_p, z_n\} \end{aligned}$$

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

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

Inequalities in (22) lead to

$$\max\left\{\frac{2m_1}{m_+}, \frac{2m_1}{m_-}\right\} < \max\left\{\frac{4}{P(Y=+1)}, \frac{4}{P(Y=-1)}\right\}.$$

Hence, we can choose $\bar{z}$ satisfying $\max\{z_p, z_n\} < \bar{z}$. Suppose that $\ell(0)/2 \leq K\lambda_{m_1} + \bar{z}$ holds for $m_1 \geq M$. Then, the inequalities

$$|\widehat{\rho}| \leq K\lambda_{m_1} + \bar{z}, \quad |\widehat{b}| \leq \frac{\ell(0)}{2} + K\lambda_{m_1} + \bar{z},$$

hold with the probability higher than $1 - e^{-cm_1}$ for $m_1 \geq M$. By choosing an appropriate positive constant $C > 0$, we obtain (15). ■

Appendix F. Proof of Theorem 4

Proof Lemma 10 in Appendix D assures that, for any $\gamma > 0$, there exists sufficiently large M_1 such that

$$|\inf\{\mathcal{R}_{\lambda_{m_1}}(f+b, \rho) : f \in \mathcal{H},\, b, \rho \in \mathbb{R}\} - \mathcal{R}^*| \leq \gamma$$

holds for all $m_1 \geq M_1$. Thus, there exist f_γ, b_γ and ρ_γ such that

$$|\mathcal{R}_{\lambda_{m_1}}(f_\gamma + b_\gamma, \rho_\gamma) - \mathcal{R}^*| \leq 2\gamma$$

and $\|f_\gamma\|_{\mathcal{H}} \leq \lambda_{m_1}$ hold for $m_1 \geq M_1$. Due to the law of large numbers, the inequality

$$|\widehat{\mathcal{R}}_{T_1}(f_\gamma + b_\gamma, \rho_\gamma) - \mathcal{R}(f_\gamma + b_\gamma, \rho_\gamma)| \leq \gamma$$

holds with high probability, say $1 - \delta_{m_1}$, for $m_1 \geq M_2$. The boundedness property in Lemma 1 leads to

$$P((\widehat{f}, \widehat{b}, \widehat{\rho}) \in \mathcal{G}_{m_1}) \geq 1 - e^{-cm_1}$$

for $m_1 \geq M_3$. In addition, by the uniform bound shown in Lemma 3, the inequality

$$\sup_{(f,b,\rho)\in\mathcal{G}_{m_1}} |\widehat{\mathcal{R}}_{T_1}(f + b, \rho) - \mathcal{R}(f + b, \rho)| \leq \gamma$$

holds with probability $1 - \delta'_{m_1}$. Hence, the probability such that the inequality

$$|\widehat{\mathcal{R}}_{T_1}(\widehat{f} + \widehat{b}, \widehat{\rho}) - \mathcal{R}(\widehat{f} + \widehat{b}, \widehat{\rho})| \leq \gamma$$

holds is higher than $1 - e^{-cm_1} - \delta'_{m_1}$ for $m_1 \geq M_3$. Let M_0 be $M_0 = \max\{M_1, M_2, M_3\}$. We have the inequalities

$$\widehat{\mathcal{R}}_{T_1}(\widehat{f} + \widehat{b}, \widehat{\rho}) = \widehat{\mathcal{R}}_{T_1,\lambda_{m_1}}(\widehat{f} + \widehat{b}, \widehat{\rho}) \leq \widehat{\mathcal{R}}_{T_1,\lambda_{m_1}}(f_\gamma + b_\gamma, \rho_\gamma) = \widehat{\mathcal{R}}_{T_1}(f_\gamma + b_\gamma, \rho_\gamma).$$

Then, for any $\gamma > 0$, the following inequalities hold with probability higher than $1 - e^{-cm_1} - \delta'_{m_1} - \delta_{m_1}$ for $m_1 \geq M_0$,

$$\begin{aligned}
\mathcal{R}(\widehat{f} + \widehat{b}, \widehat{\rho}) &\leq \widehat{\mathcal{R}}_{T_1}(\widehat{f} + \widehat{b}, \widehat{\rho}) + \gamma \\
&\leq \widehat{\mathcal{R}}_{T_1}(f_\gamma + b_\gamma, \rho_\gamma) + \gamma \\
&\leq \mathcal{R}(f_\gamma + b_\gamma, \rho_\gamma) + 2\gamma \\
&= \mathcal{R}_{\lambda_{m_1}}(f_\gamma + b_\gamma, \rho_\gamma) + 2\gamma \\
&\leq \mathcal{R}^* + 4\gamma.
\end{aligned}$$

■

Appendix G. Proof of Theorem 5

Proof For a fixed ρ such that $\rho \geq -\ell(0)/2$, the loss function $\ell(\rho - z)$ is classification-calibrated (Bartlett et al., 2006), since $\ell'(\rho) > 0$ holds. Hence $\psi(\theta, \rho)$ in Assumption 4 satisfies $\psi(0, \rho) = 0$, $\psi(\theta, \rho) > 0$ for $0 < \theta \leq 1$, and $\psi(\theta, \rho)$ is continuous and strictly increasing for $\theta \in [0, 1]$. 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 for the classification-calibrated loss. 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

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

since $\widehat{\rho} \ge -\ell(0)/2$ holds due to (23). We assumed that $\mathcal{R}(\widehat{f}+\widehat{b}, \widehat{\rho})$ converges to $\mathcal{R}^*$ in probability. Then, for any $\varepsilon > 0$, the inequality

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

holds with high probability for sufficiently large m_1. Thus, $\psi(\mathcal{E}(\widehat{f}+\widehat{b}) - \mathcal{E}^*, \widehat{\rho})$ converges to zero in probability. The inequality

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

and the assumption on the function $\widetilde{\psi}$ ensure that $\mathcal{E}(\widehat{f}+\widehat{b})$ converges to $\mathcal{E}^*$ in probability, when m_1 tends to infinity. As a result, for any $\gamma > 0$,

$$|\mathcal{E}(\widehat{f}+\widehat{b}) - \mathcal{E}^*| \le \gamma \tag{24}$$

holds with probability higher than $1 - \delta_{m_1,\gamma}$ with respect to the probability distribution of T_1, where $\delta_{m_1,\gamma}$ satisfies $\lim_{m_1\to\infty} \delta_{m_1,\gamma} = 0$ for any $\gamma > 0$.

Next, we study the relation between $\widehat{f}+\widehat{b}$ and $\widehat{f}+\widetilde{b}$. The sample size of T_2 is m_2. For any fixed $f \in \mathcal{H}$, we define the set of 0-1 valued functions, $\mathcal{S}_f = \{[\![f(x)+b]\!] : b \in \mathbb{R}\}$. The VC-dimension of $\mathcal{S}_f$ equals to one[1]. Indeed, for two distinct points $x, x' \in \mathcal{X}$ such that $f(x) \ge f(x')$, the event such that $[\![f(x)+b]\!] = 0$ and $[\![f(x')+b]\!] = 1$ is impossible. Hence, for any $\varepsilon > 0$ and any $f \in \mathcal{H}$, the inequality

$$\sup_{b\in\mathbb{R}} |\widehat{\mathcal{E}}_{T_2}(f+b) - \mathcal{E}(f+b)| \le \gamma \tag{25}$$

holds with probability higher than $1 - \delta''_{m_2,\gamma}$ with respect to the joint probability of training sample T_2. Note that $\delta''_{m_2,\gamma}$ depends only on m_2, γ and the VC-dimension of $\mathcal{S}_f$. Thus, δ''_{m_2} is independent of the choice of $f \in \mathcal{H}$. Remember that $\widehat{f}+\widehat{b}$ depends only on the data set T_1. Due to the law of large numbers, the inequality

$$|\widehat{\mathcal{E}}_{T_2}(\widehat{f}+\widehat{b}) - \mathcal{E}(\widehat{f}+\widehat{b})| \le \gamma$$

holds with probability higher than $1 - \delta'_{m_2,\gamma}$ with respect to the probability distribution of T_2 conditioned on T_1. Since the 0-1 loss is bounded, it is possible to choose $\delta'_{m_2,\gamma}$ independent of $\widehat{f}$. From the uniform convergence property (25), the following inequality also holds

$$|\widehat{\mathcal{E}}_{T_2}(\widehat{f}+\widetilde{b}) - \mathcal{E}(\widehat{f}+\widetilde{b})| \le \gamma$$

1. See Vapnik (1998) for the definition of the VC dimension.

with probability higher than $1 - \delta''_{m_2,\gamma}$ with respect to the probability distribution of T_2 conditioned on the observation of T_1. In addition, we have

$$\widehat{\mathcal{E}}_{T_2}(\widehat{f} + \widetilde{b}) \leq \widehat{\mathcal{E}}_{T_2}(\widehat{f} + \widehat{b}).$$

Given the training samples T_1 satisfying (24), the inequalities

$$\mathcal{E}(\widehat{f} + \widetilde{b}) \leq \widehat{\mathcal{E}}_{T_2}(\widehat{f} + \widetilde{b}) + \gamma \leq \widehat{\mathcal{E}}_{T_2}(\widehat{f} + \widehat{b}) + \gamma \leq \mathcal{E}(\widehat{f} + \widehat{b}) + 2\gamma \leq \mathcal{E}^* + 3\gamma$$

hold with probability higher than $1 - \delta'_{m_2,\gamma} - \delta''_{m_2,\gamma}$ with respect to the probability distribution of T_2 conditioned on the observation of T_1. Hence, as for the conditional probability, we have

$$P(\{T_2 : \mathcal{E}(\widehat{f} + \widetilde{b}) \leq \mathcal{E}^* + 3\gamma\} \,|\, T_1) \geq 1 - \delta'_{m_2,\gamma} - \delta''_{m_2,\gamma}.$$

Remember that $\delta'_{m_2,\gamma}$ and $\delta''_{m_2,\gamma}$ do not depend on T_1. Hence, as for the joint probability of T_1 and T_2, we have

$$P(\{T_1, T_2 : \mathcal{E}(\widehat{f} + \widetilde{b}) \leq \mathcal{E}^* + 3\gamma\}) \geq (1 - \delta'_{m_2,\gamma} - \delta''_{m_2,\gamma})(1 - \delta_{m_1,\gamma}).$$

The above inequality implies that $\mathcal{E}(\widehat{f} + \widetilde{b})$ converges to $\mathcal{E}^*$ in probability, when m_1 and m_2 tend to infinity. ■

Appendix H. Proofs of Lemma 6 and Lemma 7

First, we show the proof of Lemma 6.

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 $\psi(\theta, \rho)$,

$$\inf_{z \in \mathbb{R}} \frac{1+\theta}{2}\ell(\rho - z) + \frac{1-\theta}{2}\ell(\rho + z). \tag{26}$$

The objective function is a finite-valued convex function on $\mathbb{R}$, and diverges to infinity when z tends to $\pm\infty$. Hence, there exists an optimal solution. Let $z^* \in \mathbb{R}$ be an optimal solution of (26). The optimality condition is given as

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

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 $\rho + z^* \leq d$. These inequalities contradict $\rho > d$. Hence, we have $\rho - z^* > d$ and $\rho + z^* > d$, i.e., $|z^*| < \rho - d$. 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 (26) with respect to z leads to 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, i.e., $z^* = z(\rho)$. 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. ■

Next, we show the proof of Lemma 7.

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

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

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

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

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