Convex Formulation for Learning from Positive and Unlabeled Data

Marthinus Christoffel du Plessis CHRISTO@MS.K.U-TOKYO.AC.JP
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Gang Niu NIUGANG@BAIDU.CN
Baidu Inc., Beijing, 100085, China

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

We discuss binary classification from only positive and unlabeled data (*PU classification*), which is conceivable in various real-world machine learning problems. Since unlabeled data consists of both positive and negative data, simply separating positive and unlabeled data yields a biased solution. Recently, it was shown that the bias can be canceled by using a particular *non-convex* loss such as the ramp loss. However, classifier training with a non-convex loss is not straightforward in practice. In this paper, we discuss a *convex* formulation for PU classification that can still cancel the bias. The key idea is to use different loss functions for positive and unlabeled samples. However, in this setup, the hinge loss is not permissible. As an alternative, we propose the double hinge loss. Theoretically, we prove that the estimators converge to the optimal solutions at the optimal parametric rate. Experimentally, we demonstrate that PU classification with the double hinge loss performs as accurate as the non-convex method, with a much lower computational cost.

1. Introduction

Let us consider the problem of learning a classifier only from positive and unlabeled data. This problem, which refer to as *PU classification*, arises in various practical situations under different guises. For example:

- The goal of *identification* is to find samples in an unlabeled dataset that are similar to the samples provided by a user. Such a situation occurs, e.g., in automatic face tagging: a user provides a set of images of himself, and the task is to automatically tag photos in the user's photo album.
- *Inlier-based outlier detection* is aimed at identifying outliers in an unlabeled dataset based on another dataset that consists only of inliers (Hido et al., 2008; Smola et al., 2009). Thanks to the information brought by the inlier dataset, this inlier-based approach is more powerful than the conventional completely unsupervised approach. This problem is also known as *semi-supervised novelty detection* (Scott & Blanchard, 2009; Blanchard et al., 2010).
- If the negative class is *too diverse*, it is difficult to collect negative data in a representative way. Such a situation is typically observable in "one-vs-rest" classification. For example, when classifying land cover images into urban and non-urban regions (Li et al., 2011), it is easy to obtain urban samples, but it is difficult to representatively collect diverse non-urban samples.
- The *negative-class dataset shift* changes the probability distributions of negative samples between the time when the training data is collected and when the classifier is applied to the test data. Solving this problem with an ordinary classifier would require constant generation of the negative dataset to keep up with the changing distributions. On the other hand, PU classification only requires to update the unlabeled dataset, which is much less costly. Negative-class dataset shift may occur in spam detection, where adversarial spammers may change the tendency of negative samples ('spam') to defeat the existing classifier, while the positive class ('non-spam') is expected to remain stable over time.

Given its wide applicability as described above, PU classification is gathering a great deal of attention these days. A naive approach to PU classification is to train a classi-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

fier to separate positive and unlabeled samples. However, such a naive approach yields a poor solution since the unlabeled dataset consists of both positive and negative data. Although using a loss function weighted according to the class prior of the unlabeled dataset[1] was shown to produce a better solution (Blanchard et al., 2010; Scott & Blanchard, 2009), the PU classifier trained in this way still has a systematic estimation bias.

Recently, it was shown that using a loss function $\ell(z)$ such that $\ell(z) + \ell(-z) = 1$ can cancel the bias completely. For example, the *ramp loss*, which is used in the *robust support vector machine* (Collobert et al., 2006; Wu & Liu, 2007), satisfies this condition[2]. Classifier training with the ramp loss can be performed, e.g,. via the *convex-concave procedure* (CCCP) (Yuille & Rangarajan, 2002). However, non-convex optimization is computationally expensive and only a sub-optimal local solution may be obtained. Another non-convex formulation was proposed in Smola et al. (2009).

To overcome this weakness of the non-convex formulation, we analyze a formulation of PU classification that is *convex* but can still cancel the bias. The key idea is to use different loss functions for positive and unlabeled samples: an ordinary convex loss function $\ell(z)$ for unlabeled samples and a composite loss function $\ell(z) - \ell(-z)$ for positive samples[3]. If $\ell(z) - \ell(-z)$ is a convex function, the entire objective function becomes convex and thus the global solution can be obtained efficiently. The *logistic loss* and the *squared loss* immediately yield convex composite loss functions. On the other hand, the composite loss derived from the *hinge loss* is not convex, but the modified hinge loss with an extra kink (which we call the *double hinge loss*) yields a convex composite loss.

Theoretically, we prove that the estimators converge to the optimal solutions at the optimal parametric rate. Experimentally, the superior accuracy and computational advantage of the proposed double hinge loss is illustrated on benchmark datasets.

2. Non-convex PU classification

In this section, we formulate the problem of PU classification and review the non-convex PU classification method proposed in du Plessis et al. (2014).

[1]Several methods have been introduced to estimate this class prior (e.g., Scott & Blanchard, 2009; du Plessis & Sugiyama, 2014).

[2]Loss function $\ell(z)$ that satisfies $\ell(z) + \ell(-z) = 1$ is always *non-convex*.

[3]Note that this idea has been previously shown in the context of learning from noisy labels (Natarajan et al., 2013). For correct parameter choice, PU learning can be interpreted as a special case of learning with noisy labels.

Formulation of PU classification: Let $\boldsymbol{x} \in \mathbb{R}^d$ be a d-dimensional pattern and $y \in \{1, -1\}$ be a class label. We assume that we have a positive dataset $\mathcal{X}$, and an unlabeled dataset $\mathcal{X}'$ i.i.d. as

$$\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^{n} \sim p(\boldsymbol{x}|y=1), \quad \mathcal{X}' := \{\boldsymbol{x}'_j\}_{j=1}^{n'} \sim p(\boldsymbol{x}),$$

where $p(\boldsymbol{x}|y)$ is the class-conditional density of patterns and $p(\boldsymbol{x})$ is the marginal density of patterns. Since the unlabeled dataset $\mathcal{X}'$ consists of positive and negative samples, the marginal density is $p(\boldsymbol{x}) := \pi p(\boldsymbol{x}|y=1) + (1-\pi)p(\boldsymbol{x}|y=-1)$. The goal is to learn a classifier $g(\boldsymbol{x})$ that assigns a label $\widehat{y}$ to a new pattern $\boldsymbol{x}$ as $\widehat{y} = \operatorname{sign}(g(\boldsymbol{x}))$.

The optimal classifier g^* is given by $g^* = \arg\min_{g \in \mathcal{G}} J_{0\text{-}1}(g)$, where $J_{0\text{-}1}(g)$ is the expected misclassification rate when the classifier $g(\boldsymbol{x})$ is applied to unlabeled samples distributed according to $p(\boldsymbol{x})$:

$$J_{0\text{-}1}(g) = \pi \mathbb{E}_1 \left[\ell_{0\text{-}1}(g(X))\right] + (1-\pi)\mathbb{E}_{-1}\left[\ell_{0\text{-}1}(-g(X))\right], \tag{1}$$

where the zero-one loss is $\ell_{0\text{-}1}(z) = \frac{1}{2}\operatorname{sign}(z) + \frac{1}{2}$.

PU classification by non-convex loss minimization: In the ordinary classification setting where positive and negative samples are available for classifier training, the expectations $\mathbb{E}_1$ and $\mathbb{E}_{-1}$ in Eq. (1) can be estimated by corresponding sample averages. In the PU classification setting, however, no labeled samples from the negative class is available and therefore $\mathbb{E}_{-1}$ cannot be estimated directly.

This problem can be avoided by expressing $J_{0\text{-}1}(g)$ as follows (du Plessis et al., 2014):

$$J_{0\text{-}1}(g) = 2\pi \mathbb{E}_1 \left[\ell_{0\text{-}1}(g(X))\right] + \mathbb{E}_X \left[\ell_{0\text{-}1}\left(-g(X)\right)\right] - \pi, \tag{2}$$

where $\mathbb{E}_X$ denotes the expectation over $p(\boldsymbol{x})$. This comes from

$$\begin{aligned}
&\mathbb{E}_X \left[\ell_{0\text{-}1}(-g(X))\right] \\
&\quad = \pi \mathbb{E}_1 \left[\ell_{0\text{-}1}(-g(X))\right] + (1-\pi)\mathbb{E}_{-1}\left[\ell_{0\text{-}1}(-g(X))\right] \\
&\quad = \pi \left(1 - \mathbb{E}_1 \left[\ell_{0\text{-}1}(g(X))\right]\right) + (1-\pi)\mathbb{E}_{-1}\left[\ell_{0\text{-}1}(-g(X))\right],
\end{aligned}$$

where the last line is due to $\ell_{0\text{-}1}(-z) = 1 - \ell_{0\text{-}1}(z)$. Note that Eq. (2) corresponds to the *cost-sensitive classification* (Elkan, 2001) with weights $2\pi/\eta$ and $1/(1-\eta)$, where η is the proportion of positive samples to unlabeled samples.

In practice, minimizing Eq. (2) is problematic since the subgradient of the zero-one loss is zero everywhere except when $z = 0$. For this reason, the zero-one loss is often

substituted with a *surrogate* loss function[4] $\ell(z)$:

$$J_{\mathrm{PU}}(g) = 2\pi\mathbb{E}_1\left[\ell(g(X))\right] + \Big[\pi\mathbb{E}_1\left[\ell(-g(X))\right] + (1-\pi)\mathbb{E}_{-1}\left[\ell(-g(X))\right]\Big] - \pi$$
$$= \underbrace{\pi\mathbb{E}_1\left[\ell\left(g(X)\right)\right] + (1-\pi)\mathbb{E}_{-1}\left[\ell(-g(X))\right]}_{\text{Ordinary error term}} + \underbrace{\pi\mathbb{E}_1\left[\ell(g(X)) + \ell(-g(X))\right]}_{\text{Superfluous penalty}} - \pi.$$

The first and second terms correspond to the ordinary classification loss, while the third term is a *superfluous* term that is specific to the PU classification setting. Due to the superfluous term, a systematic estimation bias is incurred by naive surrogate loss minimization.

However, as shown in du Plessis et al. (2014), the superfluous term can be canceled when the loss function satisfies $\ell(z) + \ell(-z) = 1$. Note that this condition is met only by non-convex loss functions such as the ramp loss[5].

Non-convex loss functions are, however, often problematic in practice due to the difficulty of non-convex optimization and the existence of local sub-optimal solutions. In the next section, we explore an alternative way to remove the superfluous penalty.

3. Convex PU classification

In this section, we give the formulation for convex PU classification.

Formulation: Let us consider another expression of $J_{0\text{-}1}(g)$ based on

$$(1-\pi)\mathbb{E}_{-1}\left[\ell_{0\text{-}1}(-g(X))\right] = \mathbb{E}_X\left[\ell_{0\text{-}1}(-g(X))\right] - \pi\mathbb{E}_1\left[\ell_{0\text{-}1}(-g(X))\right].$$

Substituting this into Eq. (1), we obtain

$$J_{0\text{-}1}(g) = \pi\mathbb{E}_1\left[\ell_{0\text{-}1}(g(X)) - \ell_{0\text{-}1}(-g(X))\right] + \mathbb{E}_X\left[\ell_{0\text{-}1}(-g(X)\right].$$

If the zero-one loss is replaced with a surrogate loss $\ell(z)$, we have

$$J(g) = \pi\mathbb{E}_1\left[\widetilde{\ell}(g(X))\right] + \mathbb{E}_X\left[\ell(-g(X))\right], \qquad (3)$$

where $\widetilde{\ell}(z)$ is the *composite loss*: $\widetilde{\ell}(z) = \ell(z) - \ell(-z)$. Eq.(3) corresponds to using an ordinary loss for unlabeled samples and a composite loss for positive samples.

[4]Examples of surrogate loss functions are illustrated in Fig. 1. Many surrogate loss functions are convex, which results in convex optimization problems.

[5]The same condition was also proved in Ghosh et al. (2014) for learning with label noise.

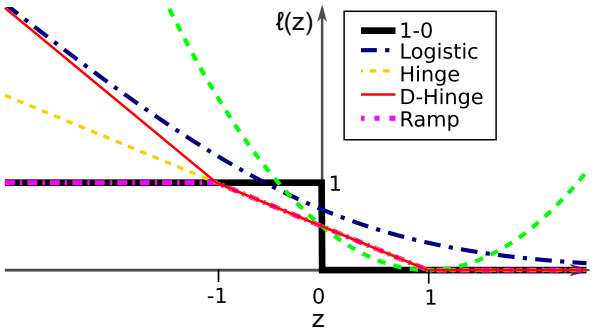

Figure 1. Selected loss functions.

When $\ell(z)$ is convex, the composite loss $\widetilde{\ell}(z)$ is the difference between two convex functions. The key question is whether the composite loss can be convex, which makes Eq.(3) a convex function. The following simple theorem (proven in Appendix A.1) positively answer the question.

Theorem 1. *If the composite loss $\widetilde{\ell}(z)$ is convex, it is linear.*

Various losses are illustrated in Fig. 1 (definitions are in Table 1). A simple calculation shows that some losses, such as the Hinge loss, do not result in a linear composite loss.

For simplicity, let us always normalize the losses so that the composite loss is $\widetilde{\ell}(z) = -z$. This results in an objective function of

$$J(g) = \pi\mathbb{E}_1\left[-g(X)\right] + \mathbb{E}_X\left[\ell\left(-g(X)\right)\right]. \qquad (4)$$

Note that the above is a special case of the previously proposed estimator in Natarajan et al. (2013) for learning from label noise. For appropriate parameter choices, the learning with label noise problem is reduced to PU learning.

Empirical version: In practice, we use a linear-in-parameter model for function $g(\boldsymbol{x})$:

$$g(\boldsymbol{x}) = \boldsymbol{\alpha}^\top\boldsymbol{\varphi}(\boldsymbol{x}) + b, \qquad (5)$$

where $\boldsymbol{\varphi}(\boldsymbol{x}) = \begin{bmatrix}\varphi_1(\boldsymbol{x}) & \ldots & \varphi_m(\boldsymbol{x})\end{bmatrix}^\top$ is a set of basis functions. For basis functions, we may use, e.g., the Gaussian functions centered around sample points $\varphi_\ell(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x}-\boldsymbol{c}_\ell\|^2/(2\sigma^2)\right)$, where $\{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_m\} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{x}'_1,\ldots,\boldsymbol{x}'_{n'}\}$, and $m = n+n'$. Alternatively, linear or polynomial functions can be used as basis functions. Using this model, Eq. (3) can be empirically estimated as

$$\widehat{J}(\boldsymbol{\alpha},b) = -\frac{\pi}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}^\top\boldsymbol{\varphi}(\boldsymbol{x}_i) - \pi b + \frac{1}{n'}\sum_{j=1}^{n'}\ell\left(-\boldsymbol{\alpha}^\top\boldsymbol{\varphi}(\boldsymbol{x}'_j) - b\right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha},$$

where the last term is for regularization. In Eq. (4), the first two terms are always positive. However, it may happen in degenerate cases, due to inadequate regularization, that the first two terms in the above empirical criterion are not bounded below by zero. To avoid numerical difficulties, we may in practice constrain these two terms to be non-negative.

The last remaining choice to obtain a practical algorithm is the choice of the loss function $\ell(z)$. We will discuss several choices in the following section.

4. Convex loss functions for PU classification

In this section, various practical choices of convex loss functions are explored.

Squared loss: The squared loss, defined as $\ell_{\mathrm{S}}(z) = \frac{1}{4}(z-1)^2$, results in the following objective function:

$$J_{\mathrm{S}}(g) = -\pi\mathbb{E}_1\left[g(\boldsymbol{x})\right] + \frac{1}{4}\mathbb{E}_X\left[\left(g(X)+1\right)^2\right] \qquad (6)$$
$$= \frac{1}{4}\int g(\boldsymbol{x})^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int g(\boldsymbol{x})[2\pi p_1(\boldsymbol{x}) - p(\boldsymbol{x})]\mathrm{d}\boldsymbol{x} + C,$$

where C is an irrelevant constant. Let us assign a class label y for $\boldsymbol{x}$ according to the difference of class-posteriors:

$$\begin{aligned} r(\boldsymbol{x}) &= p(y=1|\boldsymbol{x}) - p(y=-1|\boldsymbol{x}) \\ &= \left[p(\boldsymbol{x}|y=1)\pi - p(\boldsymbol{x}|y=-1)(1-\pi)\right]/p(\boldsymbol{x}) \\ &= \left[2\pi p(\boldsymbol{x}|y=1) - p(\boldsymbol{x})\right]/p(\boldsymbol{x}). \end{aligned}$$

Then our objective function corresponds to the least-squares fitting of the difference of posteriors $r(\boldsymbol{x})$ to a model $g(\boldsymbol{x})$ up to an irrelevant constant:

$$\frac{1}{4}\int\left(g(\boldsymbol{x}) - \frac{2\pi p_1(\boldsymbol{x}) - p(\boldsymbol{x})}{p(\boldsymbol{x})}\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

An advantage of this squared-loss formulation is that it can be analytically solved. For example, when b is omitted from the model Eq. (5), the objective function with the ℓ_2-regularizer becomes

$$\widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}) = \frac{1}{4n'}\boldsymbol{\alpha}^\top\Phi_{\mathrm{U}}^\top\Phi_{\mathrm{U}}\boldsymbol{\alpha} + \frac{1}{2n'}\mathbf{1}^\top\Phi_{\mathrm{U}}\boldsymbol{\alpha} - \frac{\pi}{n}\mathbf{1}^\top\Phi_{\mathrm{P}}\boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha},$$

where $[\Phi_{\mathrm{P}}]_{i\ell} = \varphi_\ell(\boldsymbol{x}_i)$, and $[\Phi_{\mathrm{U}}]_{j\ell} = \varphi_\ell(\boldsymbol{x}'_j)$. The minimizer of this objective function can be analytically obtained as

$$\boldsymbol{\alpha} = \left(\frac{1}{2n}\Phi_{\mathrm{U}}^\top\Phi_{\mathrm{U}} + \lambda I\right)^{-1}\left[\frac{\pi}{n}\Phi_{\mathrm{P}}^\top\mathbf{1} - \frac{1}{2n'}\Phi_{\mathrm{U}}^\top\mathbf{1}\right].$$

However, a drawback of the squared loss function is that the function increases as $z > 1$. This is undesirable, since a model that correctly classifies the sample when $z > 1$ is penalized.

Logistic loss: The logistic loss is defined as $\ell_{\mathrm{LL}}(z) = \log(1+\exp(-z))$. We therefore wish to minimize the objective function:

$$J_{\mathrm{LL}}(g) = -\pi\mathbb{E}_1\left[g(X)\right] + \mathbb{E}_X\left[\log(1+\exp(g(X)))\right]. \qquad (7)$$

The logistic loss is monotone decreasing when $z > 1$, so in this sense it is preferable over the squared loss for classification.

The proposed method can be related to *ordinary logistic regression*. The objective function for ordinary logistic regression is

$$\begin{aligned} \mathbb{E}_X\left[\ell_{\mathrm{LL}}(g(X))\right] &= \pi\mathbb{E}_1\left[\log\left(1+\exp(-g(X))\right)\right] \\ &\quad + (1-\pi)\mathbb{E}_{-1}\left[\log\left(1+\exp(g(X))\right)\right]. \end{aligned}$$

We use the identity $\log(1+\exp(-z)) = -z + \log(1+\exp(z))$ in the first term to get

$$\begin{aligned} \mathbb{E}_X\left[\ell_{\mathrm{LL}}(g(X))\right] &= -\pi\mathbb{E}_1\left[g(X)\right] \\ &\quad + \pi\mathbb{E}_1\left[\log(1+\exp(g(X)))\right] \\ &\quad + (1-\pi)\mathbb{E}_{-1}\left[\log(1+\exp(g(X))\right]. \end{aligned}$$

By collecting the last two terms into $\mathbb{E}_X$, we see that this is equivalent to Eq. (7). This implies that ordinary logistic regression can be exactly performed in the PU classification setup.

The regularized empirical approximation for the objective function in Eq. (7) is

$$\widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}, b) = -\frac{\pi}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}^\top\boldsymbol{\varphi}(\boldsymbol{x}_i) - \pi b + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + \frac{1}{n'}\sum_{j=1}^{n'}\ell_{\mathrm{LL}}\left(-\boldsymbol{\alpha}^\top\boldsymbol{\varphi}(\boldsymbol{x}'_j) - b\right). \qquad (8)$$

This function is continuous and differentiable, therefore optimization can be performed using a quasi-Newton method (see Appendix B for details).

Hinge and double hinge losses: The *hinge* loss is defined as $\ell_{\mathrm{H}}(z) = \frac{1}{2}\max(0, 1-z)$. From Table 1 we see that the composite loss $\widetilde{\ell}_{\mathrm{H}}(z)$ is not a linear function. This non-convex composite loss would lead to an undesirable non-convex optimization problem.

We can, however, obtain a convex objective function if the loss is slightly modified as $\ell_{\mathrm{DH}}(z) = \max\left(-z, \max\left(0, \frac{1}{2}-\frac{1}{2}z\right)\right)$. Since this loss function has an extra kink at $z=-1$, we refer to it as the *double hinge loss* (see Fig. 1). For this loss function, the composite loss term is $\widetilde{\ell}_{\mathrm{DH}}(z) = -z$, which is a convex function.

Table 1. A selected list of loss functions and their composite losses. Losses with a linear composite loss function was normalized so that $\widetilde{\ell}(z) = -z$.

Loss name	**Notation**	$\ell(z)$	$\widetilde{\ell}(z)$	Notes
Square loss	$\ell_{\mathrm{S}}(z)$	$\frac{1}{4}(z-1)^2 - \frac{1}{4}$	$-z$	**Convex**
Modified Huber loss	$\ell_{\mathrm{MH}}(z)$	$\begin{cases} \frac{1}{4}\max(0, 1-z)^2 & z \geq -1 \\ -z & z < -1 \end{cases}$	$-z$	**Convex**
Logistic loss	$\ell_{\mathrm{LL}}(z)$	$\log(1+\exp(-z))$	$-z$	**Convex**
Hinge loss	$\ell_{\mathrm{H}}(z)$	$\frac{1}{2}\max(0, 1-z)$	$\begin{cases} \frac{1}{2}(1-z) & z \leq -1, \\ -z & -1 \leq z \leq 1, \\ \frac{1}{2}(-1-z) & z \geq 1. \end{cases}$	**Non-convex**
Double hinge loss	$\ell_{\mathrm{DH}}(z)$	$\max(-z, \max(0, \frac{1}{2} - \frac{1}{2}z))$	$-z$	**Convex**
Perceptron loss	$\ell_{\mathrm{P}}(z)$	$\max(-z, 0)$	$-z$	**Convex**
Boosting loss	$\ell_{\mathrm{EXP}}(z)$	$\exp(-z)$	$\exp(-z) - \exp(z)$	**Non-convex**

The empirical optimization problem for the double hinge loss is

$$\widehat{J}_{\mathrm{DH}}(\boldsymbol{\alpha}, b) = -\frac{\pi}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}^\top\boldsymbol{\varphi}(\boldsymbol{x}_i) - \pi b + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + \frac{1}{n'}\sum_{j=1}^{n'}\ell_{\mathrm{DH}}\left(-\boldsymbol{\alpha}^\top\varphi(\boldsymbol{x}'_j) - b\right). \quad (9)$$

As in the standard support vector machines, we may rewrite the minimization of the above criterion as a quadratic program by using slack variables $\boldsymbol{\xi}$ to bound the max operators:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} & -\frac{\pi}{n}\mathbf{1}^\top\Phi_{\mathrm{P}}\boldsymbol{\alpha} - \pi b + \frac{1}{n'}\mathbf{1}^\top\boldsymbol{\xi} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} \\ \text{s.t.} & \boldsymbol{\xi} \geq \mathbf{0}, \\ & \boldsymbol{\xi} \geq \frac{1}{2}\mathbf{1} + \frac{1}{2}\Phi_{\mathrm{U}}\boldsymbol{\alpha} + \frac{1}{2}b\mathbf{1}, \\ & \boldsymbol{\xi} \geq \Phi_{\mathrm{U}}\boldsymbol{\alpha} + b\mathbf{1}, \end{array}$$

where $\geq$ is applied element-wise on vectors.

5. Discussion

In this section, we discuss the relation between PU classification and *inlier-based outlier detection* (Hido et al., 2008; Smola et al., 2009).

The objective of inlier-based outlier detection is to find outliers in an unlabeled dataset based on an inlier dataset. Regarding inliers as samples from the positive class, we can show that the class-posterior $p(y = 1|\boldsymbol{x})$ is proportional to the *ratio* of the densities between the positive samples and unlabeled samples:

$$p(y = 1|\boldsymbol{x}) \propto \frac{p(\boldsymbol{x}|y = 1)}{p(\boldsymbol{x})} = r(\boldsymbol{x}).$$

Since the density ratio $r(\boldsymbol{x})$ will tend to take large values for inliers and small values for outliers, it can be used for outlier detection.

The density ratio $r(\boldsymbol{x})$ can be naively estimated by first estimating the densities $p(\boldsymbol{x}|y = 1)$ and $p(\boldsymbol{x})$ from the positive and unlabeled datasets separately and then computing the ratio of estimated densities. However, this two-step procedure is undesirable since high-dimensional density estimation is often unreliable and taking their ratio can further magnify the estimation error. To cope with this problem, in Keziou (2003) and Nguyen et al. (2007), the following objective function for density ratio estimation was introduced:

$$\sup_{g}\int g(\boldsymbol{x})p(\boldsymbol{x}|y = 1)\mathrm{d}\boldsymbol{x} - \int f^*(g(\boldsymbol{x}))p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad (10)$$

where $f(t)$ is a convex function such that $f(1) = 0$, and $f^*(z) = \sup_t\ tz - f(t)$ denotes its Fenchel dual. Eq.(10) is actually a lower bound of the f-divergence from $p(\boldsymbol{x}|y = 1)$ to $p(\boldsymbol{x})$ (Ali & Silvey, 1966):

$$\int f\left(\frac{p(\boldsymbol{x}|y = 1)}{p(\boldsymbol{x})}\right)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Eq.(10) is maximized at $g = r$ if $r(\boldsymbol{x}) \in \partial f^*(g(\boldsymbol{x}))$ (i.e., the solution is a function of the density ratio). This estimator has been used for inlier-based outlier detection under the Kullback-Leibler divergence (Kullback & Leibler, 1951) in Smola et al. (2009) and under the Pearson divergence (Pearson, 1900) in Hido et al. (2008).

On the other hand, in PU classification, we are interested in the sign of the difference of class-posterior probabilities:

$$p(y = 1|\boldsymbol{x}) - p(y = -1|\boldsymbol{x}) \propto \frac{\pi p(\boldsymbol{x}|y = 1)}{p(\boldsymbol{x})} - \frac{1}{2}, \quad (11)$$

which also includes the density ratio $r(\boldsymbol{x})$. Thus, inlier-based outlier detection and PU classification are highly related to each other. However, an important difference is that inlier-based outlier detection requires an outlier score for evaluating the outlyingness of samples, while PU classification requires the threshold between the positive and negative classes. Because of this difference, Eq.(11) also contains the class prior probability $p(y = 1) = \pi$.

Nevertheless, we can utilize the density-ratio framework of f-divergence estimation in PU classification. Indeed,

Table 2. Conjugates and the corresponding loss function for PU learning (cf. Table 1).

$\bar{f}(t)$	$\bar{f}^*(z)$	Loss
$(t-\frac{1}{2})^2$	$\frac{1}{4}(z+1)^2-\frac{1}{4}$	$\ell_S(z)$
$(t-\frac{1}{2})^2, 0\leq t\leq 1$	$\begin{cases}\frac{1}{4}\max(0,1+z)^2-\frac{1}{4} & z\leq 1\\ z-\frac{1}{4} & z>1\end{cases}$	$\ell_{\mathrm{MH}}(z)$
$\lvert 2t-1\rvert, 0\leq t\leq 1$	$\max\left(z,\max\left(0,\frac{1}{2}+\frac{1}{2}z\right)\right)-\frac{1}{2}$	$\ell_{\mathrm{DH}}(z)$
$(1-t)\ln(1-t)$ $+t\ln(t)$	$\ln(1+\exp(z))$	$\ell_{\mathrm{LL}}(z)$

Eq. (4) can be expressed as

$$\sup_g \pi\int g(\boldsymbol{x})p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}-\int \bar{f}^*(g(\boldsymbol{x}))p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{12}$$

where $\bar{f}^*(t)$ corresponds to the loss function. Note that Eq.(12) is an upper bound of

$$\int \bar{f}\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

For different $\bar{f}(t)$, we can recover the squared loss, modified Huber loss, double hinge loss and logistic loss, as shown in Table 2.

6. Theoretical Analysis

In this section, we establish convergence results for the proposed methods. Assume that the number of basis functions m is a constant independent of n and n', i.e., $g(\boldsymbol{x})$ is *parametric*, and the bias b is ignored for simplicity: $g(\boldsymbol{x})=\sum_{j=1}^m \alpha_j\varphi_j(\boldsymbol{x})=\boldsymbol{\alpha}^\top\boldsymbol{\varphi}(\boldsymbol{x})$. Assume that the ideal estimates are given by

$$\begin{aligned}\boldsymbol{\alpha}^*_{\mathrm{S}}&=\arg\min J_{\mathrm{S}}(\boldsymbol{\alpha}),\\ \boldsymbol{\alpha}^*_{\mathrm{LL}}&=\arg\min J_{\mathrm{LL}}(\boldsymbol{\alpha}),\\ \boldsymbol{\alpha}^*_{\mathrm{DH}}&=\arg\min J_{\mathrm{DH}}(\boldsymbol{\alpha}),\end{aligned}$$

respectively, where we plug $g(\boldsymbol{x})$ into the original objectives $J_{\mathrm{S}}(g)$, $J_{\mathrm{LL}}(g)$ and $J_{\mathrm{DH}}(g)$. We also assume that the empirical estimates are given by

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}&=\arg\min \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}),\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}&=\arg\min \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}),\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{DH}}&=\arg\min \widehat{J}_{\mathrm{DH}}(\boldsymbol{\alpha}).\end{aligned}$$

We then derive convergence rates of the empirical estimates and the empirical objectives based on a theory known as *perturbation analysis of optimization problems* (see Bonnans & Shapiro, 1998; Bonnans & Cominetti, 1996, and references therein).

Our main idea is to regard three empirical objectives as perturbed optimizations of three expected objectives, and to establish *Lipschitzian behavior* of optimal solutions to the perturbed optimizations. Without loss of generality, assume that $0\leq\varphi_j(\boldsymbol{x})\leq 1$ for all $j=1,\ldots,m$ and $\boldsymbol{x}\in\mathbb{R}$, and that there exists a constant M_α such that $\|\widehat{\boldsymbol{\alpha}}\|_2\leq M_\alpha$ for the optimal solution $\widehat{\boldsymbol{\alpha}}$ to any optimization if $\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$ is regularized. To begin with, we have the following second-order growth conditions.

Lemma 2. *It holds that*

$$\begin{aligned}J_S(\boldsymbol{\alpha})&\geq J_S(\boldsymbol{\alpha}^*_S)+\lambda\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^*_S\|_2^2,\\ J_{LL}(\boldsymbol{\alpha})&\geq J_{LL}(\boldsymbol{\alpha}^*_{LL})+\lambda\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^*_{LL}\|_2^2,\\ J_{DH}(\boldsymbol{\alpha})&\geq J_{LL}(\boldsymbol{\alpha}^*_{DH})+\lambda\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^*_{LL}\|_2^2.\end{aligned}$$

First, consider the squared loss. Let $\boldsymbol{u}=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{u}_3\mid \boldsymbol{u}_1\in\mathcal{S}^m_+,\boldsymbol{u}_2\in\mathbb{R}^m,\boldsymbol{u}_3\in\mathbb{R}^m\}$ be a set of perturbation parameters, where $\mathcal{S}^m_+\subset\mathbb{R}^{m\times m}$ is the cone of m-by-m positive semi-definite matrices. Define our perturbed objective function by

$$\begin{aligned}J_{\mathrm{S}}(\boldsymbol{\alpha},\boldsymbol{u})=&\frac{1}{4}\boldsymbol{\alpha}^\top\left(\int\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\boldsymbol{u}_1\right)\boldsymbol{\alpha}\\ &+\frac{1}{2}\left(\int\boldsymbol{\varphi}(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\boldsymbol{u}_2\right)^\top\boldsymbol{\alpha}\\ &-\pi\left(\int\boldsymbol{\varphi}(\boldsymbol{x})p_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\boldsymbol{u}_3\right)^\top\boldsymbol{\alpha}+\frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha},\end{aligned}$$

$$\boldsymbol{\alpha}_{\mathrm{S}}(\boldsymbol{u})=\arg\min_{\boldsymbol{\alpha}} J_{\mathrm{S}}(\boldsymbol{\alpha},\boldsymbol{u}).$$

It is obvious that $J_{\mathrm{S}}(\boldsymbol{\alpha})=J_{\mathrm{S}}(\boldsymbol{\alpha},\boldsymbol{0})$, and $\widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha})=J_{\mathrm{S}}(\boldsymbol{\alpha},\boldsymbol{u})$, where

$$\begin{aligned}\boldsymbol{u}_1&=\frac{1}{n'}\sum_{i=1}^{n'}\boldsymbol{\varphi}(\boldsymbol{x}'_i)\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top-\int\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\ \boldsymbol{u}_2&=\frac{1}{n'}\sum_{i=1}^{n'}\boldsymbol{\varphi}(\boldsymbol{x}'_i)-\int\boldsymbol{\varphi}(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\ \boldsymbol{u}_3&=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}(\boldsymbol{x}_i)-\int\boldsymbol{\varphi}(\boldsymbol{x})p_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{13}$$

Lemma 3. *The difference function $J_S(\cdot,\boldsymbol{u})-J_S(\cdot)$ is Lipschitz continuous modulus $\omega(\boldsymbol{u})=\mathcal{O}(\|\boldsymbol{u}_1\|_{\mathrm{Fro}}+\|\boldsymbol{u}_2\|_2+\|\boldsymbol{u}_3\|_2)$ on a sufficiently small neighborhood of $\boldsymbol{\alpha}^*_S$.*

Theorem 4. *As $n,n'\to\infty$, we have*

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_S-\boldsymbol{\alpha}^*_S\|_2&=\mathcal{O}_p(n^{-1/2}+n'^{-1/2}),\\ |\widehat{J}_S(\widehat{\boldsymbol{\alpha}}_S)-J_S(\boldsymbol{\alpha}^*_S)|&=\mathcal{O}_p(n^{-1/2}+n'^{-1/2}).\end{aligned}$$

Second, consider the logistic loss. Let $\mathcal{U}$ be a Banach space of Lipschitz continuous functions $u:\mathbb{R}^m\mapsto\mathbb{R}$ equipped

with a sup-norm $\|u\|_\infty = \sup_{\boldsymbol{\alpha}} |u(\boldsymbol{\alpha})|$. Let $\boldsymbol{u} = \{\boldsymbol{u}_3, u_4 \mid \boldsymbol{u}_3 \in \mathbb{R}^m, u_4 \in \mathcal{U}\}$ be a set of perturbation parameters, and define our perturbed objective functional by

$$J_{\mathrm{LL}}(\boldsymbol{\alpha}, \boldsymbol{u}) = -\pi \left(\int \boldsymbol{\varphi}(\boldsymbol{x}) p_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \boldsymbol{u}_3 \right)^\top \boldsymbol{\alpha} + \int \ln(1 + \exp(\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha})) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + u_4(\boldsymbol{\alpha}) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha},$$

$\boldsymbol{\alpha}_{\mathrm{LL}}(\boldsymbol{u}) = \arg\min_{\boldsymbol{\alpha}} J_{\mathrm{LL}}(\boldsymbol{\alpha}, \boldsymbol{u})$.

It is not difficult to see that $J_{\mathrm{LL}}(\boldsymbol{\alpha}) = J_{\mathrm{LL}}(\boldsymbol{\alpha}, \boldsymbol{0})$ where $0 \in \mathcal{U}$, and $\widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}) = J_{\mathrm{LL}}(\boldsymbol{\alpha}, \boldsymbol{u})$ where

$$\begin{aligned} \boldsymbol{u}_3 &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i) - \int \boldsymbol{\varphi}(\boldsymbol{x}) p_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \\ u_4(\boldsymbol{\alpha}) &= \frac{1}{n'} \sum_{i=1}^{n'} \ln(1 + \exp(\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha})) \\ &\quad - \int \ln(1 + \exp(\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha})) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{14}$$

Lemma 5. *The difference function $J_{LL}(\cdot, \boldsymbol{u}) - J_{LL}(\cdot)$ is Lipschitz continuous modulus $\omega(\boldsymbol{u}) = \mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_4))$ on a sufficiently small neighborhood of $\boldsymbol{\alpha}^*_{LL}$, where $\mathrm{Lip}(u_4)$ is the best Lipschitz constant of u_4.*

Theorem 6. *As $n, n' \to \infty$, we have*

$$\begin{aligned} \|\widehat{\boldsymbol{\alpha}}_{LL} - \boldsymbol{\alpha}^*_{LL}\|_2 &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}), \\ |\widehat{J}_{LL}(\widehat{\boldsymbol{\alpha}}_{LL}) - J_{LL}(\boldsymbol{\alpha}^*_{LL})| &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}). \end{aligned}$$

Finally, consider the double hinge loss. Here we use a similar set of perturbation parameters $\boldsymbol{u} = \{\boldsymbol{u}_3, u_5\}$ as the logistic loss, and define our perturbed objective functional by

$$\begin{aligned} J_{\mathrm{DH}}(\boldsymbol{\alpha}, \boldsymbol{u}) &= -\pi \left(\int \boldsymbol{\varphi}(\boldsymbol{x}) p_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \boldsymbol{u}_3 \right)^\top \boldsymbol{\alpha} \\ &\quad + \int \max\left\{0, \frac{1 + \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}}{2}, \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}\right\} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\quad + u_5(\boldsymbol{\alpha}) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha}, \\ \boldsymbol{\alpha}_{\mathrm{DH}}(\boldsymbol{u}) &= \arg\min_{\boldsymbol{\alpha}} J_{\mathrm{DH}}(\boldsymbol{\alpha}, \boldsymbol{u}). \end{aligned}$$

It is easy to see that $J_{\mathrm{DH}}(\boldsymbol{\alpha}) = J_{\mathrm{DH}}(\boldsymbol{\alpha}, \boldsymbol{0})$ where $0 \in \mathcal{U}$, and $\widehat{J}_{\mathrm{DH}}(\boldsymbol{\alpha}) = J_{\mathrm{DH}}(\boldsymbol{\alpha}, \boldsymbol{u})$ where

$$\begin{aligned} \boldsymbol{u}_3 &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i) - \int \boldsymbol{\varphi}(\boldsymbol{x}) p_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \\ u_5(\boldsymbol{\alpha}) &= \frac{1}{n'} \sum_{i=1}^{n'} \max\left\{0, \frac{1 + \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}}{2}, \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}\right\} \\ &\quad - \int \max\left\{0, \frac{1 + \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}}{2}, \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}\right\} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{15}$$

Lemma 7. *The difference function $J_{DH}(\cdot, \boldsymbol{u}) - J_{DH}(\cdot)$ is Lipschitz continuous modulus $\omega(\boldsymbol{u}) = \mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_5))$ on a sufficiently small neighborhood of $\boldsymbol{\alpha}^*_{DH}$.*

Theorem 8. *As $n, n' \to \infty$, we have*

$$\begin{aligned} \|\widehat{\boldsymbol{\alpha}}_{DH} - \boldsymbol{\alpha}^*_{DH}\|_2 &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}), \\ |\widehat{J}_{DH}(\widehat{\boldsymbol{\alpha}}_{DH}) - J_{DH}(\boldsymbol{\alpha}^*_{DH})| &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}). \end{aligned}$$

To sum up, the empirical estimates and the empirical objectives converge in $\mathcal{O}_p(n^{-1/2} + n'^{-1/2})$ to the corresponding targets in all of three cases. This is the optimal convergence rate, since it is of order $\mathcal{O}_p(n^{-1/2})$ when approximating an expectation by an empirical average based on n data. Note that there is a generalization error bound in du Plessis et al. (2014) that is also of order $\mathcal{O}_p(n^{-1/2} + n'^{-1/2})$ under the problem setting of PU classification. Nevertheless, their proposed method is non-convex and has no convergence analysis. As a consequence, our proposed methods are advantageous because they possess both bounds of the convergence rate and generalization error.

7. Experiments

In this section we report experimental results.

Numerical illustrations: We numerically illustrate the effect of multiple local minima for the ramp loss on a simple numerical problem. The two class-conditional distributions are

$$p(\boldsymbol{x}|y=1) = \mathcal{N}_x\left(2, 1^2\right) \text{ and } p(x|y=-1) = \mathcal{N}_x\left(-2, 1^2\right),$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the univariate normal distribution with mean μ and variance σ^2. We generate 10 positive samples and, using a class prior of $\pi = 0.5$, generate 20 unlabeled samples. Using a model $g(x) = wx + b$, and choosing $\lambda = 10^{-3}$ we plot the value of the objective function w.r.t. to w and b in Fig. 2. Even in this simple problem, we see that the ramp loss function has multiple minima. We see in Fig. 3 that the bad local minimum leads to a worse classification result.

Next, we illustrate the failure of applying ordinary classifiers in the PU learning problem due to the superfluous penalty term. The positive and negative classes are distributed as $\mathcal{U}(0.1, 1)$ and $\mathcal{U}(-1.1, -0.1)$, where $\mathcal{U}(a, b)$ is the uniform density between a and b. This dataset is trivially separable and the offset $-b/w$ of a classifier should always be in the range $[-0.1, 0.1]$. Drawing different datasets and training classifiers on it gives the results in Fig. 4. In this simple example, we see that directly using the hinge loss and logistic loss result in a wrong classification boundary – even in fully separable datasets. Our proposed method and the ramp loss yield correct solutions.

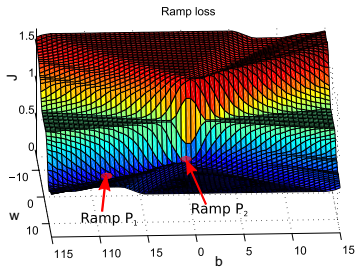

Figure 2. Objective value of the ramp loss w.r.t. w and b, illustrating multiple local minima. The objective value for P_1 is higher than P_2.

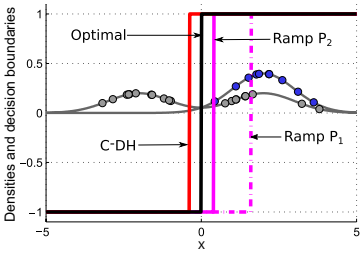

Figure 3. Resulting discriminant boundary for the ramp loss minima P_1 and P_2 and the double hinge loss. P_1 (with a higher value) is an inferior classifier.

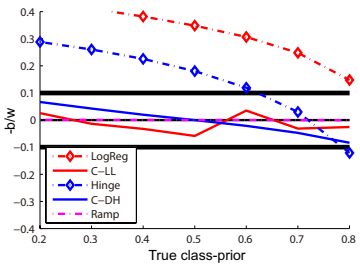

Figure 4. Average offset $-b/w$ for different classifiers trained on the fully-separable problem. Correct classifiers should be between -0.1 and 0.1.

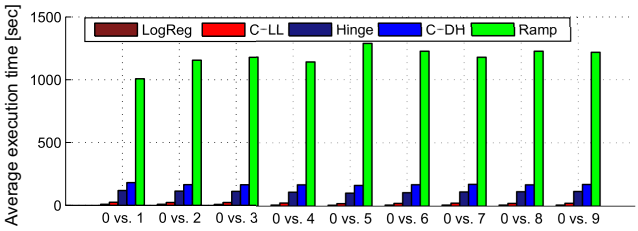

Figure 5. Average execution time of different methods

Benchmark datasets: We performed experiments illustrating the method on the MNIST dataset. The following methods were compared:

- **LogReg**, **Hinge**: Training a weighted classifier with the logistic loss and the hinge loss. These methods are convex, but subject to the superfluous penalty.
- **C-LL**, **C-DH** (proposed): The convex methods proposed in Sec. 4 using the logistic loss and double hinge loss.
- **Ramp**: The method of du Plessis et al. (2014) using the non-convex ramp loss. The objective function was minimized using the convex-concave procedure (Yuille & Rangarajan, 2002; Collobert et al., 2006) (see Appendix B.4 for a detailed discussion).

All methods used the same model. Hyperparameters were selected via cross-validation on the zero-one loss objective in Eq. (2). The "0" digit was used for the positive class, and another digit was used for the negative class (i.e., one dataset for each digit "1"…"9"). Dimensionality was reduced to 2 via principal component analysis and 200 positive samples and 400 unlabeled samples were drawn. The class prior was varied across experiments, but it is assumed that the class prior is known at training time[6].

The classification accuracy is given in Table 3. From this we, see that the ramp-loss and the proposed double hinge loss give accurate results. The comparison in average computational time (Fig. 5) shows however that our proposed double hinge loss method is significantly faster.

[6]In practice, it may be estimated with methods such as (Blanchard et al., 2010; du Plessis & Sugiyama, 2014).

Table 3. Classification accuracy (in percent) of the proposed methods. Best and equivalent methods (under 5% t-test) are bold.

Dataset	π	LogReg	C-LL	Hinge	C-DH	Ramp
0 vs 1	0.1	3.1%	0.8%	0.7%	**0.5%**	**0.5%**
	0.4	8.9%	1.3%	1.6%	**0.9%**	**0.8%**
	0.7	8.8%	1.5%	1.8%	**0.6%**	1.0%
0 vs 2	0.1	4.2%	3.0%	3.1%	**2.8%**	**2.7%**
	0.4	11.8%	6.0%	7.4%	**5.3%**	**5.3%**
	0.7	11.2%	6.9%	7.6%	**5.1%**	5.4%
0 vs. 3	0.1	4.1%	2.7%	2.9%	**2.5%**	**2.5%**
	0.4	11.9%	5.8%	7.4%	**5.1%**	**5.1%**
	0.7	11.3%	6.9%	7.5%	**5.1%**	5.4%
0 vs 4	0.1	3.7%	1.8%	2.1%	1.6%	**1.4%**
	0.4	10.8%	3.9%	5.0%	2.8%	**2.5%**
	0.7	10.2%	4.4%	5.3%	**2.7%**	**2.8%**
0 vs 5	0.1	5.0%	4.1%	4.4%	4.0%	**3.9%**
	0.4	14.4%	**9.4%**	11.5%	**9.4%**	**9.3%**
	0.7	13.4%	11.1%	13.2%	10.5%	**10.0%**
0 vs 6	0.1	4.1%	3.1%	3.1%	**2.9%**	**2.8%**
	0.4	11.6%	6.4%	7.8%	**5.9%**	**5.8%**
	0.7	11.6%	7.1%	7.8%	**6.0%**	**6.1%**
0 vs 7	0.1	3.7%	2.0%	2.2%	1.7%	**1.5%**
	0.4	10.4%	4.0%	4.9%	**3.0%**	**2.8%**
	0.7	10.2%	5.0%	5.0%	**2.8%**	3.1%
0 vs 8	0.1	4.0%	2.8%	2.8%	**2.6%**	**2.6%**
	0.4	11.4%	5.8%	6.9%	**5.1%**	**5.0%**
	0.7	10.8%	6.4%	7.9%	**5.3%**	**5.3%**
0 vs 9	0.1	4.0%	2.7%	2.9%	**2.5%**	**2.4%**
	0.4	11.1%	4.8%	5.7%	**4.1%**	**4.0%**
	0.7	10.5%	5.3%	5.8%	**3.8%**	**3.9%**

8. Conclusion

We discussed a convex framework for learning from positive and unlabeled data. Theoretically, it was shown that the proposed estimators converge to the optimal solutions at the optimal parametric rate. Experimentally it was shown that PU classification with the proposed double hinge loss perform as accurate as the non-convex ramp-loss method, but with a much lower computational burden. Furthermore, we related the convex PU learning framework to several f-divergence estimation based PU learning methods.

Acknowledgements MCdP and MS were supported by JST CREST.

References

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

Blanchard, G., Lee, G., and Scott, C. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

Bonnans, F. and Cominetti, R. Perturbed optimization in banach spaces I: A general theory based on a weak directional constraint qualification; II: A theory based on a strong directional qualification condition; III: Semi-infinite optimization. *SIAM Journal on Control and Optimization*, 34:1151–1171, 1172–1189, and 1555–1567, 1996.

Bonnans, F. and Shapiro, A. Optimization problems with perturbations, a guided tour. *SIAM Review*, 40(2):228–264, 1998.

Collobert, R., Sinz, F. H., Weston, J., and Bottou, L. Trading convexity for scalability. In *Proceedings of 23rd International Conference on Machine Learning*, pp. 201–208, 2006.

du Plessis, M. C. and Sugiyama, M. Class prior estimation from positive and unlabeled data. *IEICE Transactions on Information and Systems*, E97-D, 2014.

du Plessis, M. C, Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 27*, pp. 703–711. 2014.

Elkan, C. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001.

Ghosh, A., Manwani, N., and Sastry, P. S. Making risk minimization tolerant to label noise. *CoRR*, abs/1403.3610, 2014.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., and Kanamori, T. Inlier-based outlier detection via direct density ratio estimation. In *Proceedings of IEEE International Conference on Data Mining*, pp. 223–232, 2008.

Keziou, A. Dual representation of ϕ-divergences and applications. *Comptes Rendus Mathématique*, 336(10):857–862, 2003.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

Li, W., Guo, Q., and Elkan, C. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):717–725, 2011.

Natarajan, N., Dhillon, I.S., Ravikumar, P.K., and Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pp. 1196–1204, 2013.

Nguyen, X.-L., Wainwright, M. J., and Jordan, M. I. Nonparametric estimation of the likelihood ratio and divergence functionals. In *Proceedings of IEEE International Symposium on Information Theory*, pp. 2016–2020, june 2007.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

Scott, C. and Blanchard, G. Novelty detection: Unlabeled data definitely help. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pp. 464–471, 2009.

Smola, A. J., Song, L., and Teo, C. H. Relative novelty detection. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pp. 536–543, 2009.

Wu, Y. and Liu, Y. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

Yuille, A. L. and Rangarajan, A. The concave-convex procedure (CCCP). In *Advances in Neural Information Processing Systems 14*, pp. 1033–1040. MIT Press, 2002.

A. Proofs

A.1. Proof of Theorem 1

If the composite loss $\widetilde{\ell}(z)$ is convex, it is linear.

Proof: The composite loss is an odd function:

$$\widetilde{\ell}(-z) = \ell(-z) - \ell(z) = -\widetilde{\ell}(z),$$

Therefore, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) = -\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(-z)$. If the composite loss $\widetilde{\ell}(z)$ is convex, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) \geq 0$ holds for all z. Since the convexity of $\widetilde{\ell}(z)$ implies the convexity of $\widetilde{\ell}(-z)$, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(-z) \geq 0$ should also hold for all z. However, if $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) > 0$, then $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(-z) < 0$ holds, which is contradictory to the convexity of $\widetilde{\ell}(-z)$. Therefore, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) = 0$ should hold, which is satisfied only when $\widetilde{\ell}(z)$ is linear. □

A.2. Proof of Lemma 2

$J_{\mathrm{S}}(\boldsymbol{\alpha})$ is strongly convex in $\boldsymbol{\alpha}$ with parameter at least λ, and thus

$$\begin{aligned} J_{\mathrm{S}}(\boldsymbol{\alpha}) &\geq J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) + \nabla J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}) + \lambda\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2^2 \\ &\geq J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) + \lambda\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2^2, \end{aligned}$$

where we use the optimality condition $\nabla J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) = \mathbf{0}$. Similarly, we can prove the other two inequalities. □

A.3. Proof of Lemma 3

The difference function can be written as

$$J_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{S}}(\boldsymbol{\alpha}) = \frac{1}{4}\boldsymbol{\alpha}^\top \boldsymbol{u}_1 \boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{u}_2^\top \boldsymbol{\alpha} - \pi \boldsymbol{u}_3^\top \boldsymbol{\alpha},$$

with a partial gradient

$$\frac{\partial}{\partial \boldsymbol{\alpha}}(J_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{S}}(\boldsymbol{\alpha})) = \frac{1}{2}\boldsymbol{u}_1\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{u}_2 - \pi \boldsymbol{u}_3.$$

Given the δ-ball of $\boldsymbol{\alpha}^*_{\mathrm{S}}$, i.e., $B_\delta(\boldsymbol{\alpha}^*_{\mathrm{S}}) = \{\boldsymbol{\alpha} \mid \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 \leq \delta\}$, it is easy to see that for any $\boldsymbol{\alpha} \in B_\delta(\boldsymbol{\alpha}^*_{\mathrm{S}})$,

$$\|\boldsymbol{\alpha}\|_2 \leq \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 + \|\boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 \leq 1 + M_\alpha,$$

and then

$$\left\| \frac{\partial}{\partial \boldsymbol{\alpha}}(J_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{S}}(\boldsymbol{\alpha})) \right\|_2 \leq \frac{1}{2}(1 + M_\alpha)\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \frac{1}{2}\|\boldsymbol{u}_2\|_2 + \pi\|\boldsymbol{u}_3\|_2.$$

This means that $J_{\mathrm{S}}(\cdot, \boldsymbol{u}) - J_{\mathrm{S}}(\cdot)$ is Lipschitz continuous on $B_\delta(\boldsymbol{\alpha}^*_{\mathrm{S}})$ with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \|\boldsymbol{u}_2\|_2 + \|\boldsymbol{u}_3\|_2)$. □

A.4. Proof of Lemma 5

The difference function can be written as

$$J_{\mathrm{LL}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}) = -\pi \boldsymbol{u}_3^\top \boldsymbol{\alpha} + u_4(\boldsymbol{\alpha}).$$

Given $\boldsymbol{\alpha} \in B_\delta(\boldsymbol{\alpha}^*_{\mathrm{LL}})$, we have known that $-\pi \boldsymbol{u}_3^\top \boldsymbol{\alpha}$ is Lipschitz continuous with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}_3\|_2)$ in the proof of Lemma 3. Consequently, $J_{\mathrm{LL}}(\cdot, \boldsymbol{u}) - J_{\mathrm{LL}}(\cdot)$ is Lipschitz continuous on $B_\delta(\boldsymbol{\alpha}^*_{\mathrm{LL}})$ with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_4))$. □

A.5. Proof of Lemma 7

Same as the proof of Lemma 5. □

A.6. Proof of Theorem 4

Let $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ be defined as in Eq. (13). According to the *central limit theorem*,

$$\|\boldsymbol{u}_1\|_{\mathrm{Fro}} = \mathcal{O}_p(n'^{-1/2}), \quad \|\boldsymbol{u}_2\|_2 = \mathcal{O}_p(n'^{-1/2}), \quad \|\boldsymbol{u}_3\|_2 = \mathcal{O}_p(n^{-1/2}),$$

as $n, n' \to \infty$. Thus, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 &\leq \lambda^{-1}\omega(\boldsymbol{u}) \\ &= \mathcal{O}(\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \|\boldsymbol{u}_2\|_2 + \|\boldsymbol{u}_3\|_2) \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2})\end{aligned}$$

by Lemma 2, Lemma 3, and Proposition 6.1 in Bonnans & Shapiro (1998, p. 19).

On the other hand,

$$|\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| \leq |\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| + |\widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})|,$$

in which

$$\begin{aligned}\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) &= (\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} + \boldsymbol{\alpha}^*_{\mathrm{S}})^\top \left(\frac{1}{4n'}\sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i)\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top + \frac{\lambda}{2} I_m\right)(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}) \\ &\quad + \left(\frac{1}{2n'}\sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i)\right)^\top (\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}) - \pi\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i)\right)^\top (\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}), \\ \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) &= \frac{1}{4}\boldsymbol{\alpha}^{*\top}_{\mathrm{S}}\boldsymbol{u}_1\boldsymbol{\alpha}^*_{\mathrm{S}} + \frac{1}{2}\boldsymbol{u}_2\boldsymbol{\alpha}^*_{\mathrm{S}} - \pi\boldsymbol{u}_3\boldsymbol{\alpha}^*_{\mathrm{S}}.\end{aligned}$$

Since $0 \leq \varphi_j(\boldsymbol{x}) \leq 1$, $\|\boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 \leq M_\alpha$ and $\|\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}\|_2 \leq M_\alpha$,

$$\begin{aligned}|\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| &\leq |\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| + |\widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| \\ &\leq \mathcal{O}_p(\|\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2) + \mathcal{O}_p(\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \|\boldsymbol{u}_2\|_2 + \|\boldsymbol{u}_3\|_2) \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}),\end{aligned}$$

which completes the proof. □

A.7. Proof of Theorem 6

Let $\boldsymbol{u}_3$ and $u_4(\boldsymbol{\alpha})$ be defined as in Eq. (14). The gradient of u_4 is given by

$$\nabla u_4(\boldsymbol{\alpha}) = \frac{1}{n'}\sum_{i=1}^{n'} \frac{\boldsymbol{\varphi}(\boldsymbol{x}'_i)}{1 + \exp(-\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top\boldsymbol{\alpha})} - \int \frac{\boldsymbol{\varphi}(\boldsymbol{x})}{1 + \exp(-\boldsymbol{\varphi}(\boldsymbol{x})^\top\boldsymbol{\alpha})} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

According to the central limit theorem,

$$\|\boldsymbol{u}_3\|_2 = \mathcal{O}_p(n^{-1/2}), \quad \mathrm{Lip}(u_4) = \mathcal{O}_p(n'^{-1/2}),$$

as $n, n' \to \infty$, since $\mathrm{Lip}(u_4) = \sup_{\boldsymbol{\alpha}} \|\nabla u_4(\boldsymbol{\alpha})\|_2$ and

$$\sup_{\boldsymbol{\alpha}\in\mathbb{R}^m, \boldsymbol{x}\in\mathbb{R}^d} \left\| \frac{\boldsymbol{\varphi}(\boldsymbol{x})}{1 + \exp(-\boldsymbol{\varphi}(\boldsymbol{x})^\top\boldsymbol{\alpha})} \right\|_2 \leq m^{1/2} < \infty.$$

Thus, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}}\|_2 &\leq \lambda^{-1}\omega(\boldsymbol{u}) \\ &= \mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_4))\end{aligned}$$

$$= \mathcal{O}_p(n^{-1/2} + n'^{-1/2})$$

by Lemma 2, Lemma 5, and Proposition 6.1 in Bonnans & Shapiro (1998, p. 19).

On the other hand,

$$|\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \leq |\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| + |\widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})|.$$

For the second term,

$$\begin{aligned}|\widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| &= |-\pi \boldsymbol{u}_3^\top \boldsymbol{\alpha}^*_{\mathrm{LL}} + u_4(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \\ &\leq \pi M_\alpha \|\boldsymbol{u}_3\|_2 + |u_4(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2})\end{aligned}$$

according to the central limit theorem. For the first term, it is a bit more complex:

$$\begin{aligned}|\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \leq \left|\frac{\lambda}{2}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} + \boldsymbol{\alpha}^*_{\mathrm{LL}})^\top(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}})\right| + \left|\pi\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{\varphi}(\boldsymbol{x}_i)\right)^\top(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}})\right| \\ + \frac{1}{n'}\sum_{i=1}^{n'} |\ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})) - \ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}))|.\end{aligned}$$

Let $f(z,t) = \ln(1+\exp(z+t))$, then $\lim_{t\to 0} f(z,t) = f(z,0)$ and

$$\lim_{t\to 0}\frac{f(z,t)-f(z,0)}{t} = \lim_{t\to 0}\frac{\partial}{\partial t}f(z,t) = \frac{1}{1+\exp(-z-t)} < \infty,$$

where we use *L'Hôpital's rule*. In other words, $f(z,t)$ approaches $f(z,0)$ in $\mathcal{O}(t)$ as $t \to 0$. Subsequently, for any $\boldsymbol{x} \in \mathbb{R}^d$, by $z = \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}$ and $t = \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}$ we can obtain

$$\begin{aligned}|\ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})) - \ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}))| &= \mathcal{O}(|\boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}|) \\ &= \mathcal{O}(m^{1/2}\|\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}}\|_2),\end{aligned}$$

which results in $|\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| = \mathcal{O}_p(n^{-1/2} + n'^{-1/2})$. □

A.8. Proof of Theorem 8

The proof goes along the same line as that of Theorem 6. Let $\boldsymbol{u}_3$ and $u_5(\boldsymbol{\alpha})$ be defined as in Eq. (15). Note that the function $\max\{0, (1+z)/2, z\}$ is piecewise linear in z, differentiable almost everywhere, and $0 \leq (\mathrm{d}/\mathrm{d}z)\max\{0,(1+z)/2,z\} \leq 1$. As a result,

$$\|\boldsymbol{u}_3\|_2 = \mathcal{O}_p(n^{-1/2}), \quad \mathrm{Lip}(u_5) = \mathcal{O}_p(n'^{-1/2}),$$

as $n, n' \to \infty$, and

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_{\mathrm{DH}} - \boldsymbol{\alpha}^*_{\mathrm{DH}}\|_2 &\leq \lambda^{-1}\omega(\boldsymbol{u}) \\ &= \mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_5)) \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2})\end{aligned}$$

by Lemma 2, Lemma 7, and Proposition 6.1 in Bonnans & Shapiro (1998, p. 19).

On the other hand,

$$\begin{aligned}|\widehat{J}_{\mathrm{DH}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DH}}) - J_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}})| &\leq |\widehat{J}_{\mathrm{DH}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DH}}) - \widehat{J}_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}})| + |\widehat{J}_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}}) - J_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}})| \\ &\leq \frac{1}{n'}\sum_{i=1}^{n'} |\max\{0, (1+\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}\}\end{aligned}$$

$$- \max\{0, (1 + \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}^*_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}\}| + \mathcal{O}_p(n^{-1/2} + n'^{-1/2}).$$

Let $f(z,t) = \max\{0, (1+z+t)/2, z+t\}$, then $\lim_{t\to 0} f(z,t) = f(z,0)$ and for $z \in \mathbb{R} \setminus \{0,1\}$,

$$\lim_{t\to 0} \frac{f(z,t) - f(z,0)}{t} = \lim_{t\to 0} \frac{\partial}{\partial t} f(z,t) \in \left\{0, \frac{1}{2}, 1\right\}.$$

In other words, $f(z,t)$ approaches $f(z,0)$ in $\mathcal{O}(t)$ as $t \to 0$ almost surely. Subsequently, for any $\boldsymbol{x} \in \mathbb{R}^d$, by $z = \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{DH}}$ and $t = \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{DH}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{DH}}$ we can obtain

$$\begin{aligned}
|\max\{0, (1 + \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}\} - \max\{0, (1 + \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}\}| &= \mathcal{O}(|\boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}|) \\
&= \mathcal{O}(m^{1/2} \|\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}}\|_2) \\
&= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}),
\end{aligned}$$

which completes the proof. □

B. Optimization problems

In this section, we give exact optimization problems for the optimization methods presented in the paper. The logistic regression and logistic loss method is solved with a quasi-Newton method, and therefore we provide the derivatives in Sec. B.1.

The Hinge loss and Double Hinge loss result in quadratic problems. The ramp-loss is solved via a sequence of quadratic problems. All quadratic problems are expressed in the form

$$\begin{array}{ll} \min_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^\top H \boldsymbol{\alpha} + \boldsymbol{f}^\top \boldsymbol{\alpha} \\ \text{s.t.} & L\boldsymbol{\alpha} \preceq \boldsymbol{k} \\ & \boldsymbol{l} \preceq \boldsymbol{\alpha} \end{array}$$

This standard form can then just be plugged into an off-the-shelf optimization package such as Gurobi, IBM CPLEX or MATLAB's internal 'quadprog' function.

B.1. Logistic loss

The gradient for the objective function in Eq. (8) is

$$\begin{aligned}
\frac{\partial \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}, b)}{\partial \boldsymbol{\alpha}} = &-\frac{\pi}{n} \Phi_{\mathrm{P}}^\top \mathbf{1} + \lambda \boldsymbol{\alpha} \\
&- \frac{1}{n'} \sum_{j=1}^{n'} \ell'_{\mathrm{LL}} \left(-\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}'_j) - b\right) \boldsymbol{\varphi}(\boldsymbol{x}'_j),
\end{aligned}$$

where $\ell'_{\mathrm{LL}}(z)$ is the derivative of $\ell_{\mathrm{LL}}(z)$:

$$\ell'_{\mathrm{LL}}(z) = -\frac{\exp(-z)}{1 + \exp(-z)}.$$

The derivative with respect to the unregularized constant b is

$$\frac{\partial \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}, b)}{\partial b} = -\pi - \frac{1}{n'} \sum_{j=1}^{n'} \ell'_{\mathrm{LL}} \left(-\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}'_j) - b\right).$$

B.2. Double Hinge Loss - PU Learning

The objective function can be expressed as

$$-\frac{\pi}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i) + \frac{1}{n'} \sum_{j=1}^{n'} \max\left(0, \max\left(g(\boldsymbol{x}'_j), \frac{1}{2} + \frac{1}{2} g(\boldsymbol{x}'_j)\right)\right) + \frac{\lambda}{2} \|g\|_2^2$$

$$= -\frac{\pi}{n}\sum_{i=1}^{n}\left(\sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}_i)+b\right) + \frac{1}{n'}\sum_{j=1}^{n'}\max\left(0, \max\left(\sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_j)+b, \frac{1}{2}+\frac{1}{2}\left(\sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_j)+b\right)\right)\right) + \frac{\lambda}{2}\sum_{\ell=1}^{m}\alpha_\ell^2$$

The objective function can then be expressed as

$$\begin{array}{ll}\min\limits_{\boldsymbol{\alpha},b,\boldsymbol{\xi}} & -\frac{\pi}{n}\mathbf{1}^\top\Phi_{\mathrm{P}}\boldsymbol{\alpha} - \pi b + \frac{1}{n'}\mathbf{1}^\top\boldsymbol{\xi} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} \\ \text{s.t.} & \boldsymbol{\xi} \succeq \mathbf{0}, \\ & \boldsymbol{\xi} \succeq \frac{1}{2}\mathbf{1} + \frac{1}{2}\Phi_{\mathrm{U}}\boldsymbol{\alpha} + \frac{1}{2}b\mathbf{1}, \\ & \boldsymbol{\xi} \succeq \Phi_{\mathrm{U}}\boldsymbol{\alpha} + b\mathbf{1},\end{array}$$

Let

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\alpha}_{b\times 1} \\ b \\ \boldsymbol{\xi}_{n'\times 1} \end{bmatrix}.$$

Then H is defined as

$$H = \begin{bmatrix} \lambda I_{m\times m} & O_{m\times 1} & O_{m'\times n'} \\ O_{1\times m} & 0 & O_{1\times n'} \\ O_{n'\times m} & O_{n'\times 1} & O_{n'\times n'} \end{bmatrix},$$

where $O_{n\times m}$ is a zero matrix of n rows and m columns. The linear part of the objective is

$$\boldsymbol{f} = \begin{bmatrix} -\frac{\pi}{n}\Phi_{\mathrm{P}}^\top\mathbf{1} \\ -\pi \\ \frac{1}{n'}\mathbf{1}_{n'\times 1} \end{bmatrix}$$

The lower-bound is

$$\boldsymbol{l} = \begin{bmatrix} -\infty_{m\times 1} \\ -\infty \\ \mathbf{0}_{n'\times 1} \end{bmatrix}.$$

The first linear constraint is

$$\begin{aligned} \boldsymbol{\xi} &\succeq \frac{1}{2}\mathbf{1} + \frac{1}{2}\Phi_{\mathrm{U}}\boldsymbol{\alpha} + \frac{1}{2}b\mathbf{1} \\ \frac{1}{2}\Phi_{\mathrm{U}}\boldsymbol{\alpha} + \frac{1}{2}b\mathbf{1} - \boldsymbol{\xi} &\preceq -\frac{1}{2}\mathbf{1} \\ \begin{bmatrix} \frac{1}{2}\Phi_{\mathrm{U}} & \frac{1}{2}\mathbf{1}_{n'\times 1} & -I_{n'\times n'} \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ u \\ \boldsymbol{\xi} \end{bmatrix} &\preceq -\frac{1}{2}\mathbf{1}_{n'\times 1}. \end{aligned}$$

The second linear constraint is

$$\begin{aligned} \boldsymbol{\xi} &\succeq \Phi_{\mathrm{U}}\boldsymbol{\alpha} + b\mathbf{1} \\ \Phi_{\mathrm{U}}\boldsymbol{\alpha} + b\mathbf{1} - \boldsymbol{\xi} &\preceq \mathbf{0}_{n'\times 1} \\ \begin{bmatrix} \Phi_{\mathrm{U}} & \mathbf{1}_{n'\times 1} & -I_{n'\times n'} \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ b \\ \boldsymbol{\xi} \end{bmatrix} &\preceq \mathbf{0}_{n'\times 1}. \end{aligned}$$

Combining the two sets of inequalities, we get

$$L = \begin{bmatrix} \frac{1}{2}\Phi_{\mathrm{U}} & \frac{1}{2}\mathbf{1}_{n'\times 1} & -I_{n'\times n'} \\ \Phi_{\mathrm{U}} & \mathbf{1}_{n'\times 1} & -I_{n'\times n'} \end{bmatrix},$$

and

$$\boldsymbol{k} = \begin{bmatrix} -\frac{1}{2}\mathbf{1}_{n'\times 1} \\ \mathbf{0}_{n'\times 1} \end{bmatrix}.$$

B.3. Weighted hinge loss classifier

We want a cost-sensitive classifier with a per-sample weighting. Using the model

$$g(\boldsymbol{x}) = \sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_) + b,$$

where

$$\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m\} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\},$$

we wish to minimize

$$\begin{aligned} J(g) &= \frac{1}{n} \sum_{i=1}^{b} w_i \ell_{\mathrm{H}} \left(y_i \sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha}, \\ &= \frac{1}{2n} \sum_{i=1}^{n} w_i \max \left(0, 1 - y_i \sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha}. \end{aligned}$$

This gives a QP of

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} & \frac{1}{2n} \boldsymbol{w}^\top \boldsymbol{\xi} + \frac{\lambda}{2} \boldsymbol{\alpha}^\top R \boldsymbol{\alpha} \\ \text{s.t.} & \xi_i \geq 0, \qquad \forall i = 1, \ldots, n \\ & \xi_i \geq 1 - y_i \sum_{\ell=1}^{b} \alpha_\ell k(\boldsymbol{x}_i, \boldsymbol{c}_\ell) + u \qquad \forall i = 1, \ldots, n. \end{array}$$

We then set

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\alpha} \\ b \\ \boldsymbol{\xi} \end{bmatrix}.$$

H is then

$$H = \begin{bmatrix} \lambda I & O_{m \times 1} & O_{m \times n} \\ O_{1 \times n} & 0 & O_{1 \times n} \\ O_{n \times n} & O_{n \times 1} & O_{n \times n} \end{bmatrix}.$$

The linear term is

$$\boldsymbol{f} = \begin{bmatrix} \mathbf{0}_{m \times 1} \\ 0 \\ \frac{1}{2n} \boldsymbol{w} \end{bmatrix}$$

The lower bound is

$$\boldsymbol{l} = \begin{bmatrix} -\infty_{m \times 1} \\ -\infty \\ \mathbf{0}_{n \times 1} \end{bmatrix}$$

Define $\bar{\Phi}$ as

$$\bar{\Phi}_{i\ell} = y_i \varphi_\ell(\boldsymbol{x}_i).$$

The constraint can be written in matrix form as

$$\begin{aligned} \boldsymbol{\xi} &\succeq \mathbf{1}_{n \times 1} - \left(\bar{\Phi} \boldsymbol{\alpha} + b \boldsymbol{y} \right) \\ -\bar{\Phi} \boldsymbol{\alpha} - b \boldsymbol{y} - \boldsymbol{\xi} &\preceq -\mathbf{1}_{n \times 1} \end{aligned}$$

The matrix is then

$$L = \begin{bmatrix} -\Phi & -\boldsymbol{y} & -I_{n \times n} \end{bmatrix},$$

and $\boldsymbol{k}$ is

$$\boldsymbol{k} = \left[-\mathbf{1}_{n \times 1} \right].$$

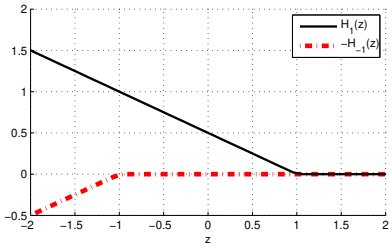

Figure 6. Decomposition of the ramp-loss into convex and concave parts.

B.4. Weighted ramp-loss classifier (CCCP)

Classification with the ramp-loss is difficult, due to the the non-convexity of the loss function. One popular method to perform optimization is to split the non-convex function into a convex and concave part. The concave part is then upper-bounded by a linear function, and optimization is iteratively performed: minimization of the upper-bound, and tightening of the upper-bound around the new minima. We minimize the ramp-loss problem here using this approach. This is a straightforward application of the convex-concave procedure (CCCP) in Yuille & Rangarajan (2002) and is essentially the same as Collobert et al. (2006).

We wish to minimize the following non-convex objective function:

$$J(\boldsymbol{\alpha}, b) = \frac{1}{n}\sum_{i=1}^{n} w_i \ell_{\mathrm{R}}\left(y_i \sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b\right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}, \tag{16}$$

where the ramp loss $\ell_{\mathrm{R}}(z)$ is defined as

$$\ell_{\mathrm{R}}(z) = \max\left(0, \min\left(1, \frac{1}{2} - \frac{1}{2}z\right)\right) = \frac{1}{2}\max\left(0, \min(2, 1-z)\right).$$

By defining the following (slightly more general) hinge loss

$$H_\epsilon(z) = \frac{1}{2}\max(0, \epsilon - z),$$

the ramp loss $\ell_{\mathrm{R}}(z)$ can be decomposed as:

$$\ell_{\mathrm{R}}(z) = H_1(z) - H_{-1}(z).$$

This is illustrated in Fig. 6. The objective Eq. (16) can therefore be decomposed as

$$\begin{aligned} J(\boldsymbol{\alpha}, b) &= J_{\mathrm{vex}}(\boldsymbol{\alpha}, b) + J_{\mathrm{cave}}(\boldsymbol{\alpha}, b), \\ J_{\mathrm{vex}}(\boldsymbol{\alpha}, b) &= \frac{1}{n}\sum_{i=1}^{n} w_i H_1\left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b\right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}, \\ J_{\mathrm{cave}}(\boldsymbol{\alpha}, b) &= -\frac{1}{n}\sum_{i=1}^{n} w_i H_{-1}\left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b\right) \end{aligned}$$

The following self-evident relation can be used to upper-bound the concave part

$$\begin{aligned} tz - f(z) \leq \sup_{y\in\mathbb{R}} yt - f(y) \\ \Rightarrow f(z) \geq tz - f^*(t), \end{aligned} \tag{17}$$

where

$$f^*(t) = \sup_{y\in\mathbb{R}} yt - f(y).$$

The inequality in Eq.(17) is known as the *Fenchel inequality* and the function $f^*(z)$ is known as the *Fenchel dual* or *convex conjugate*. Applying the above inequality to $H_\epsilon(z)$, we can obtain a bound as

$$\begin{aligned} H_\epsilon(z) &\geq zt - H_\epsilon^*(t), \\ -H_\epsilon(z) &\leq H_\epsilon^*(t) - zt, \end{aligned}$$

where $H_\epsilon^*(t)$ is the Fenchel dual of $H_\epsilon(z)$. The Fenchel dual of $H_{-1}(t)$ is (the full calculation is given in Appendix B.4.3)

$$H_{-1}^*(t) = \begin{cases} -t & -\frac{1}{2} \leq t \leq 0, \\ \infty & \text{otherwise.} \end{cases}$$

We can minimize the upper-bound as

$$\arg\min_t H_{-1}^*(t) - tz = \begin{cases} t = 0 & z > -1. \\ t = -\frac{1}{2} & z \leq -1. \end{cases}$$

The concave part is then bounded, with the parameter $\boldsymbol{a}$ as

$$\bar{J}_{\text{cave}}(\boldsymbol{\alpha}, b, \boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n} w_i \left(H_1^*(a_i) - a_i y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) \right),$$

where $J_{\text{cave}}(\boldsymbol{\alpha}, u) \leq \bar{J}_{\text{cave}}(\boldsymbol{\alpha}, b, \boldsymbol{a})$, for any $\boldsymbol{a}$.

B.4.1. TIGHTENING OF THE UPPER-BOUND

The upperbound is minimized (tightened) when

$$a_i = \begin{cases} -\frac{1}{2} & y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b\right) \leq -1, \\ 0 & \text{otherwise.} \end{cases}$$

B.4.2. MINIMIZING THE OBJECTIVE

We wish to minimize the convex part and the upper bound $\bar{J}(\boldsymbol{\alpha}, u, \boldsymbol{a}) = J_{\text{vex}}(\boldsymbol{\alpha}, u) + \bar{J}_{\text{cave}}(\boldsymbol{\alpha}, u, \boldsymbol{a})$ with respect to $\boldsymbol{a}$. This gives an objective of

$$\bar{J}(\boldsymbol{\alpha}, b, \boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n} w_i H_1 \left(y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) \right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{1}{n}\sum_{i=1}^{n} w_i a_i y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right).$$

We define the following matrices:

$$\begin{aligned} \Phi_{i,\ell} &= y_i k(\boldsymbol{x}_i, \boldsymbol{c}_\ell), \\ \bar{\Phi}_{i,\ell} &= w_i a_i y_i k(\boldsymbol{x}_i, \boldsymbol{c}_\ell), \end{aligned}$$

The QP for this is then

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} & \frac{1}{2n}\boldsymbol{w}^\top \boldsymbol{\xi} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{1}{n}\mathbf{1}^\top \bar{\Phi}\boldsymbol{\alpha} - b\frac{1}{n}\sum_{i=1}^{n} w_i a_i y_i. \\ s.t. & \xi_i \geq 0 \qquad \forall i = 1, \ldots, n \\ & \xi_i \geq 1 - y_i \left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) \qquad \forall i = 1, \ldots, n. \end{array}$$

We define again

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\alpha} \\ b \\ \boldsymbol{\xi} \end{bmatrix}.$$

The quadratic term is

$$H = \begin{bmatrix} \lambda I_{m\times m} & O_{m\times 1} & O_{n\times n} \\ O_{1\times n} & 0 & O_{1\times n} \\ O_{n\times n} & O_{n\times 1} & O_{n\times n} \end{bmatrix}.$$

The linear term is

$$\boldsymbol{f} = \begin{bmatrix} -\frac{1}{n}\bar{\Phi}^\top \mathbf{1} \\ -\frac{1}{n}\sum_{i=1}^n w_i a_i y_i \\ \frac{1}{2n}\boldsymbol{w} \end{bmatrix}$$

The lower-bound is

$$lb = \begin{bmatrix} -\infty_{m\times 1} \\ -\infty \\ \mathbf{0}_{n\times 1} \end{bmatrix}.$$

The linear term is

$$-\Phi\boldsymbol{\alpha} - b\boldsymbol{y} - \boldsymbol{\xi} \preceq -\mathbf{1}_{n\times 1}.$$

This gives a matrix of

$$L = \begin{bmatrix} -\Phi & -\boldsymbol{y} & -I_{n\times n} \end{bmatrix},$$

and $\boldsymbol{k}$ is

$$\boldsymbol{k} = \left[-\mathbf{1}_{n\times 1}\right].$$

B.4.3. Calculation of the Fenchel dual of $H_\epsilon(z)$

In this section, we briefly give the derivation of the Fenchel dual of $H_\epsilon(z)$

$$\begin{aligned} H_\epsilon^*(t) &= \sup_v tv - H_\epsilon(v) \\ &= \sup_v tv - \frac{1}{2}\max\left(0, \epsilon - v\right). \end{aligned}$$

To make the above easier, we split the domain of the v:

$$\begin{aligned} H_\epsilon^*(t) &= \max\left(\sup_{v\le\epsilon} tv - \frac{1}{2}\max\left(0,\epsilon - v\right), \sup_{v>\epsilon} tv - \frac{1}{2}\max\left(0,\epsilon - v\right)\right), \\ &= \max\left(\sup_{v\le\epsilon} tv - \frac{1}{2}\left(\epsilon - v\right), \sup_{v>\epsilon} tv\right). \end{aligned}$$

For the first part:

$$\begin{aligned} \sup_{v\le\epsilon} tv - \frac{1}{2}\left(\epsilon - v\right) &= \sup_{v\le\epsilon} v\left(t + \frac{1}{2}\right) - \frac{1}{2}\epsilon, \\ &= \begin{cases} \epsilon t & t \ge -\frac{1}{2}, \\ \infty & t < \frac{1}{2} \end{cases} \end{aligned}$$

The second part is

$$\sup_{t>\epsilon} tv = \begin{cases} \epsilon v & t \le 0, \\ \infty & t > 0. \end{cases}$$

Putting these two together gives:

$$H_\epsilon^*(t) = \begin{cases} \epsilon t & -\frac{1}{2} \le t \le 0, \\ \infty & \text{otherwise.} \end{cases}$$