Neural Networks 53 (2014) 95–108

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Learning using privileged information: SVM+ and weighted SVM

Maksim Lapin [a,*], Matthias Hein [b], Bernt Schiele [a]

[a] *Max Planck Institute for Informatics, Saarbrücken, Germany*

[b] *Saarland University, Saarbrücken, Germany*

ARTICLE INFO

Article history:
Received 7 June 2013
Received in revised form 4 February 2014
Accepted 9 February 2014

Keywords:
SVM
SVM+
Weighted SVM
Importance weighting
Privileged information
Prior knowledge

ABSTRACT

Prior knowledge can be used to improve predictive performance of learning algorithms or reduce the amount of data required for training. The same goal is pursued within the learning using privileged information paradigm which was recently introduced by Vapnik et al. and is aimed at utilizing additional information available only at training time—a framework implemented by SVM+. We relate the privileged information to importance weighting and show that the prior knowledge expressible with privileged features can also be encoded by weights associated with every training example. We show that a weighted SVM can always replicate an SVM+ solution, while the converse is not true and we construct a counterexample highlighting the limitations of SVM+. Finally, we touch on the problem of choosing weights for weighted SVMs when privileged features are not available.

1. Introduction: prior knowledge, privileged information, and instance weights

Classification is a well-studied problem in machine learning, however, learning still remains a challenging task when the amount of training data is limited. Hence, information available *in addition* to the training sample – the prior knowledge – is the crucial factor in achieving further performance improvement.

Prior knowledge comes in different forms and its incorporation into the learning problem depends on a particular setting as well as the algorithm. This paper focuses on introducing prior knowledge into a support vector machine (SVM) for binary classification. Lauer and Bloch (2008) provide a review of different ways to incorporate prior knowledge into SVMs and give a categorization of the reviewed methods based on the *type* of prior knowledge they assume; see also Schölkopf and Smola (2002). We will mainly consider the scenario where the additional information is about the *training data* rather than about the target function. A loosely related setting is the semi-supervised learning approach (Chapelle, Schölkopf, & Zien, 2006), where unlabeled data carries certain information about the marginal distribution in the input space.

Recently, Vapnik and Vashist (2009) introduced the learning using privileged information (LUPI) paradigm which aims at improving predictive performance of learning algorithms and reducing the amount of required training data. The additional information in this framework comes in the form of privileged features, which are available at training time, but not at test time. These features are used to parametrize the upper bound on the loss function and, essentially, are used to estimate the loss of an optimal classifier on the given training sample. Higher loss may be seen as an indication that a given point is likely to be an outlier, and, hence, should be treated differently than a non-outlier. This simple idea has been extensively explored in the literature and we give a few pointers in Section 1.2. The additional information about which training examples are likely to be outliers can be encoded via instance weights, therefore, one can already anticipate a close relation between the LUPI framework and importance weighting which is discussed next.

In the weighted learning scenario, each training example comes with a non-negative weight which is used in the loss function to balance the cost of errors. A typical example where instance weights appear naturally is the cost-sensitive learning (Elkan, 2001). If classes are unbalanced or different misclassification errors incur different penalties, one can encode that prior knowledge in the form of instance weights. Assigning high weight to a data point suggests that the learning algorithm should try to classify that point correctly, possibly at the cost of misclassifying "less important" points. In this paper, however, we do *not* make the cost-sensitive assumption, i.e., we do not assume that different errors incur different costs on the *test* set. Instead, we decouple importance weighting on the training and on the test sets, and we only focus on the former. This allows us, in particular, to also assign a *high* weight to an outlier if that ultimately leads to a better model.

[*] Corresponding author. Tel.: +49 681 9325 2122.
E-mail addresses: mlapin@mpi-inf.mpg.de, maksim.lapin@gmail.com (M. Lapin).

http://dx.doi.org/10.1016/j.neunet.2014.02.002
0893-6080/© 2014 Elsevier Ltd. All rights reserved.

As mentioned above, there are different forms of prior knowledge that can be encoded differently. In this paper, we show that instance weights can express *the same type of prior knowledge* that is encoded via privileged features. In particular, this allows one to interpret the effect of privileged features in terms of the incurred importance weights. Remarkably, the resulting weights *do* emphasize outliers, which also happen to be support vectors in SVMs.

Our focus in this work is on the study of the SVM+ algorithm, which is an extension of the support vector machine to the LUPI framework (Vapnik & Vashist, 2009). Using basic tools of convex analysis, we investigate uniqueness of the SVM+ solution and its relation to solutions of the weighted SVM (WSVM). It turns out there is a simple connection between an SVM+ solution and WSVM instance weights, moreover, that relation can be used to better understand the SVM+ algorithm and to study its limitations. Having realized that instance weights in WSVMs can serve the same purpose as privileged features in SVM+, we also turn to the problem of choosing weights when privileged features are not available.

1.1. Our contributions

Below is a summary of contributions of this work.

- We show that any non-trivial SVM+ solution is unique (in the primal), which is a stronger result than the one available for (W)SVMs, where the offset b may not be unique.
- By reformulating the SVM+ dual optimization problem, we reveal its close connection to the WSVM algorithm. In particular, we show that any SVM+ dual solution can be used to construct weights for the WSVM that will yield the same primal solution up to the non-uniqueness of b. This implies that WSVM with appropriately chosen weights can mimic SVM+ and that it is always possible to go from an SVM+ solution to a WSVM solution.
- We also study whether it is always possible to go in the opposite direction (which would imply that the two algorithms are equivalent). We give the necessary and sufficient condition for such an equivalence to hold and reveal that the SVM+ solutions are a strict subset of the WSVM solutions. We construct a simple counterexample where a WSVM solution cannot be found by SVM+, no matter which privileged features are used or which values the hyper-parameters take.
- Finally, we turn to the problem of choosing weights in the absence of privileged features. We show that the weights can be learned directly from data by minimizing an estimate of risk similar to standard procedures of hyper-parameter tuning. In the idealized setting, where the estimate is computed on a large validation set, we show that the WSVM with learned weights outperforms both the SVM and the SVM+. This highlights the potential of weighted learning and should motivate further work on the choice of weights.

1.2. Related work

We now briefly discuss related work on learning using privileged information and weighted learning.

Since the introduction of the new learning paradigm and the corresponding SVM+ algorithm in Vapnik (2006) and later in Vapnik and Vashist (2009); Vapnik, Vashist, and Pavlovitch (2009), there is a growing body of work on theoretical analysis (Pechyony & Vapnik, 2010), implementation (Pechyony & Vapnik, 2011) and application of the proposed framework to various machine learning settings. Liang and Cherkassky (2008); Liang, Cai, and Cherkassky (2009) study the relation between the SVM+ approach and the multi-task learning scenario, Fouad, Tino, Raychaudhury, and Schneider (2012) apply the SVM+ idea to metric learning, and Chen, Liu, and Lyu (2012) extend it to boosting algorithms. Feyereisl and Aickelin (2012) use privileged information for data clustering and Wolf and Levy (2013) propose an SVM$\ominus$ method to compute similarity scores in video face recognition. Note, however, that the latter method is not related to the SVM− algorithm we have in mind in Section 4.5. In particular, SVM$\ominus$ reduces to SVM with a pre-processing step, similar to Schölkopf, Simard, Smola, and Vapnik (1998), while in our case the optimization problem as well as the motivation are entirely different.

Instance weighting has been widely used in various machine learning settings and the topic is to too vast to cover all of the related work here. We only give a few pointers to papers on cost-sensitive learning (Margineantu, 2002; Zadrozny, Langford, & Abe, 2003), sample bias correction (Cortes, Mansour, & Mohri, 2010; Heckman, 1979), domain adaptation (Shimodaira, 2000; Sugiyama & Müller, 2005), online learning (Dredze, Crammer, & Pereira, 2008), and active learning (Beygelzimer, Dasgupta, & Langford, 2009). Perhaps the most related in terms of the learning algorithm (SVM) and the *interpretation* of instance weights are the works on fuzzy SVM (Lin & Wang, 2002), where each data point has a fuzzy class membership represented by a weight between 0 and 1, weighted margin SVM (Wu & Srihari, 2004), where again each label has a confidence score between 0 and 1, and weighted SVM with an outlier detection pre-processing step (Yang, Song, & Cao, 2005), where a kernel-based clustering algorithm is used to generate instance weights.

1.3. Organization

The rest of the paper is organized as follows. In Section 2 we introduce the SVM+ and the weighted SVM (WSVM) algorithms. In Section 3 we study basic properties of these algorithms, namely, uniqueness of their solutions. In Section 4 we present our main result which consists of four parts. Theorem 3 shows that any SVM+ solution is also a WSVM solution with appropriately chosen weights, Theorem 4 gives the necessary and sufficient condition for equivalence between the SVM+ and WSVM problems, and Section 4.4 presents an example where a WSVM solution cannot be found by SVM+, no matter which privileged features are used. Finally, Section 4.5 discusses whether it is possible to complement SVM+ with an SVM−.

Section 5 is concerned with the problem of choosing weights, where we propose a weight learning method in Section 5.3. Lastly, Section 6 presents experimental results on a number of publicly available data sets and Section 7 gives some concluding remarks.

All proofs are moved to Appendix to enhance readability.

2. Preliminaries

In this section we describe the necessary background. Our results are based on basic notions from convex analysis (Boyd & Vandenberghe, 2004) and, in particular, on the Karush–Kuhn–Tucker (KKT) conditions. For convenience, the latter are provided in Appendix A for both of the optimization problems studied below.

2.1. The setting and notation

We consider a binary classification problem with an instance space $\mathcal{X}$ and the label set $\mathcal{Y} = \{-1, 1\}$. Let $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a training sample drawn i.i.d. from an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$, and ℓ be a convex loss function $\ell : \mathbb{R} \to \mathbb{R}_+$, e.g., the hinge loss $\ell(yf(\boldsymbol{x})) = [1 - yf(\boldsymbol{x})]_+$. The task is to learn $f : \mathcal{X} \to \mathbb{R}$ from a set of hypotheses $\mathcal{H}$, that yields label prediction by sign $f(\boldsymbol{x})$ and achieves the lowest expected loss $L(f) := \mathbb{E}\ell(Yf(X))$.

We use $\tilde{\mathcal{X}}$ to denote the space of privileged information used in the SVM+, while the $\star$ is reserved to indicate a solution to an optimization problem.

In the non-linear setting, the input data is first mapped into a feature space endowed with an inner product. The decision space $\mathcal{X}$ is mapped into $\mathcal{Z}$ via a feature map Φ ($\boldsymbol{x}_i \mapsto \Phi(\boldsymbol{x}_i) = \boldsymbol{z}_i$) and the correcting space $\tilde{\mathcal{X}}$ is mapped into $\tilde{\mathcal{Z}}$ via $\tilde{\Phi}$ ($\tilde{\boldsymbol{x}}_i \mapsto \tilde{\Phi}(\tilde{\boldsymbol{x}}_i) = \tilde{\boldsymbol{z}}_i$). It is known (Schölkopf, Herbrich, & Smola, 2001) that inner products correspond to positive definite kernel functions[1] as follows: $\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle_{\mathcal{Z}} = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle_{\mathcal{Z}} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (and similar for $\tilde{\mathcal{X}}$), which allows to formulate algorithms with general kernels in mind. Since the corresponding space should be clear from the context, we omit the subscripts when dealing with inner products and the induced norms.

Unless transposed with $^\top$, all vectors are column vectors denoted by lower case bold letters, matrices are denoted by capital bold letters, and random variables are denoted by capital letters. We let $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ and $\mathbf{Y} = \operatorname{diag}(\boldsymbol{y})$. The kernel matrices $\mathbf{K}$ and $\tilde{\mathbf{K}}$ are defined entrywise via $\mathbf{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\tilde{\mathbf{K}}_{ij} = \tilde{k}(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)$, where $i, j = 1, \dots, n$. We also introduce the index sets $\mathcal{I}_{\pm} := \{i : y_i \gtrless 0\}$, $\mathcal{I}_0 := \{i : y_i f(\boldsymbol{x}_i) < 1\}$, $\mathcal{I}_1 := \{i : y_i f(\boldsymbol{x}_i) \le 1\}$, and a shorthand $\mathrm{P}(1|\boldsymbol{x}) := \mathrm{P}(Y = 1|X = \boldsymbol{x})$.

Finally, $\mathcal{N}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A})$ stand correspondingly for the null space and the column space of a matrix $\mathbf{A}$, $\boldsymbol{a}^\perp$ is the orthogonal complement of $\boldsymbol{a}$, and $\mathbf{0}$ (respectively $\mathbf{1}$) is the vector of all zeros (ones).

2.2. The SVM+ optimization problem

In the framework of learning using privileged information (LUPI), the decision space $\mathcal{X}$ is augmented with a correcting space $\tilde{\mathcal{X}}$ of privileged features $\tilde{\boldsymbol{x}}$ that are available *at training time only* and are essentially used to estimate the loss $\ell(y_i f^\star(\boldsymbol{x}_i))$ of an optimal classifier $f^\star := \arg\min_{f \in \mathcal{H}} L(f)$ on the given training sample. The SVM+ algorithm (Pechyony & Vapnik, 2011) is a generalization of the support vector machine that implements the LUPI paradigm. The slack variables ξ_i are parametrized as a function of privileged features:

$$\xi_i(\tilde{\boldsymbol{w}}, \tilde{b}) := \langle \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}_i \rangle + \tilde{b},$$

where $(\tilde{\boldsymbol{w}}, \tilde{b})$ are the additional parameters to be learned. The following optimization problem defines the SVM+ algorithm.

$$\begin{aligned} \min_{\boldsymbol{w}, b, \tilde{\boldsymbol{w}}, \tilde{b}} \quad & \frac{1}{2}(\|\boldsymbol{w}\|^2 + \gamma \|\tilde{\boldsymbol{w}}\|^2) + C \sum_{i=1}^n \xi_i(\tilde{\boldsymbol{w}}, \tilde{b}) \\ \text{s.t.} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle + b) \ge 1 - \xi_i(\tilde{\boldsymbol{w}}, \tilde{b}), \qquad \xi_i(\tilde{\boldsymbol{w}}, \tilde{b}) \ge 0. \end{aligned} \tag{1}$$

Note that there are two hyper-parameters, γ and C, that control the trade-off between the three terms of the objective, where the second term limits the capacity of the set of correcting functions $\xi_i(\tilde{\boldsymbol{w}}, \tilde{b})$.

2.3. The WSVM optimization problem

The weighted support vector machine (WSVM) is a well-known generalization of the standard SVM. Each instance $(\boldsymbol{x}_i, y_i)$ is assigned an importance weight $c_i \in \mathbb{R}_+$ and in place of the standard empirical risk estimator $\hat{L}(f) := n^{-1} \sum_{i=1}^n \ell(y_i f(\boldsymbol{x}_i))$ its weighted version is employed:

$$\hat{L}_w(f) := \sum_{i=1}^n c_i \ell(y_i f(\boldsymbol{x}_i)).$$

The WSVM optimization problem is given below.

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^n c_i \xi_i \\ \text{s.t.} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle + b) \ge 1 - \xi_i, \quad \xi_i \ge 0. \end{aligned} \tag{2}$$

At first glance, it may appear that the two generalizations of the SVM are unrelated. As will become clear in the following, however, there is a relation between the two and the solution space of WSVMs includes SVM+ solutions. This is not very surprising as soon as one realizes that re-weighting allows to alter the loss function to a large extent and, in particular, one can mimic the effect of privileged features. The close relationship can already be seen when comparing the corresponding dual problems.

2.4. The dual optimization problems

Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be the Lagrange dual variables of the SVM+ or the WSVM problem corresponding respectively to the first and the second inequality constraints (Schölkopf & Smola, 2002; Vapnik et al., 2009). Define $\tilde{\boldsymbol{\alpha}} := \boldsymbol{\alpha} + \boldsymbol{\beta} - \boldsymbol{c}$, where for the SVM+ we set $\boldsymbol{c} = C\mathbf{1}$, and note that $\boldsymbol{\beta}$ can be eliminated leading to the constraint $\alpha_i \le c_i + \tilde{\alpha}_i$. Let

$$F(\boldsymbol{\alpha}) := \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{YKY}\boldsymbol{\alpha} - \mathbf{1}^\top \boldsymbol{\alpha}, \qquad \tilde{F}(\tilde{\boldsymbol{\alpha}}) := \frac{1}{2}\tilde{\boldsymbol{\alpha}}^\top \tilde{\mathbf{K}} \tilde{\boldsymbol{\alpha}}.$$

It is not hard to see that the following optimization problem is equivalent to the dual of the SVM+ problem (1).

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}}} \quad & F(\boldsymbol{\alpha}) + \frac{1}{\gamma}\tilde{F}(\tilde{\boldsymbol{\alpha}}) \\ \text{s.t.} \quad & \boldsymbol{y}^\top \boldsymbol{\alpha} = 0, \qquad \mathbf{1}^\top \tilde{\boldsymbol{\alpha}} = 0, \quad 0 \le \alpha_i \le C + \tilde{\alpha}_i. \end{aligned} \tag{3}$$

Likewise, the problem below is equivalent to the dual of the WSVM problem (2).

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & F(\boldsymbol{\alpha}) \\ \text{s.t.} \quad & \boldsymbol{y}^\top \boldsymbol{\alpha} = 0, \quad 0 \le \alpha_i \le c_i. \end{aligned} \tag{4}$$

Note that the constraint $\alpha_i \le c_i + \tilde{\alpha}_i$ is the crucial part of the SVM+ problem as it introduces the coupling between the decision space $\mathcal{X}$ and the correcting space $\tilde{\mathcal{X}}$. Recall from the representer theorem (Schölkopf et al., 2001) that an SVM solution has the form $f = \sum_{i=1}^n \alpha_i y_i k(\boldsymbol{x}_i, \cdot)$. Correcting features thus control the maximum influence a data point $(\boldsymbol{x}_i, y_i)$ can have on the resulting classifier, just like the weights in WSVMs.

3. Uniqueness results

The connection between SVM+ and WSVM explored in Section 4 relies on the analysis of uniqueness of their solutions. Effectively, the statements can only be made with respect to the classes of equivalent solutions and equivalent weights, hence, it is imperative to first obtain a better understanding of different sources of non-uniqueness in the aforementioned problems.

In this section, we show that every non-trivial SVM+ solution is unique, unlike WSVM solutions that may have a non-unique offset b. Furthermore, we describe a set of equivalent weights that yield the same WSVM solutions. The latter will be used to prove equivalence between the SVM+ and the WSVM algorithms under additional constraints.

3.1. Uniqueness of WSVM and SVM+ solutions

We begin with a known result due to Burges and Crisp (1999) that characterizes uniqueness of the weighted SVM solution.

[1] A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which for all $n \in \mathbb{N}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathcal{X}$ gives rise to a positive definite kernel matrix $\mathbf{K}$ is called a positive definite kernel.

Essentially, it states that if there is an equilibrium between instance weights of support vectors, then the separating hyperplane can be shifted within a certain range without altering the total cost in the WSVM problem. In that case, a WSVM solver has to choose a value for the offset using some heuristic, e.g., it can choose the middle point in the allowed range of b.

Theorem 1. *The solution to the problem (2) is unique in $\boldsymbol{w}$. It is not unique in b and $\boldsymbol{\xi}$ iff at least one of the following two conditions holds:*

$$\sum_{i\in\mathcal{I}_-\cap\mathcal{I}_0} c_i = \sum_{i\in\mathcal{I}_+\cap\mathcal{I}_1} c_i, \qquad \sum_{i\in\mathcal{I}_+\cap\mathcal{I}_0} c_i = \sum_{i\in\mathcal{I}_-\cap\mathcal{I}_1} c_i.$$

Note that in practice it may happen that one of the two conditions holds and the WSVM problem (2) does not have a unique solution. This is not the case for the SVM+ as shown next.

Theorem 2. *The solution to the problem (1) is unique in $(\boldsymbol{w}, \tilde{\boldsymbol{w}}, \tilde{b})$ for any $C > 0$, $\gamma > 0$. If there is a support vector, then b is unique as well, otherwise:*

$$\max_{i\in\mathcal{I}_+}(1 - \langle\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}_i\rangle - \tilde{b}) \le b \le \min_{i\in\mathcal{I}_-}(\langle\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b} - 1).$$

This result is interesting on its own, since it shows that the SVM+ is formulated in a way that privileged features always give enough information to choose *the* unique solution (if there are no support vectors, then the constant classifier can be given by $b = \pm 1$ depending on the class balance).

Results concerning uniqueness of dual solutions are more technical and are moved to the Appendix.

3.2. Equivalent weights

Apart from the conditions discussed in the previous section, another source of non-uniqueness is that any given WSVM solution corresponds, in general, to multiple weight vectors $\boldsymbol{c}$. In this section, we give a characterization of all such vectors.

Definition 1. A family of equivalent weights $\mathcal{W}$ is defined for a given WSVM solution as follows:

$$\mathcal{W} := \{\boldsymbol{\mu} + \boldsymbol{\nu} \mid \boldsymbol{\mu} \in \mathcal{U},\ \ \boldsymbol{\nu} \in \mathcal{V}\},$$

$$\mathcal{U} := \left\{\boldsymbol{\mu} \in \mathbb{R}^n_+ \mid \sum_i \mu_i y_i \boldsymbol{z}_i = \boldsymbol{w}^\star,\ \sum_i \mu_i y_i = 0, \right.$$
$$\left. \sum_i \mu_i = \sum_i \alpha_i^\star,\ \ \mu_i(\xi_i^\star - h_i) = 0\ \forall i \right\},$$

$$\mathcal{V} := \{\boldsymbol{\nu} \in \mathbb{R}^n_+ \mid \nu_i \xi_i^\star = 0 \quad \forall i\},$$

where $h_i := [1 - y_i(\langle\boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + b^\star)]_+$ is the hinge loss at a point $i = 1, \ldots, n$.

The following simple statement shows that the set $\mathcal{W}$ defined above contains *all* weights that correspond to a given WSVM solution.

Proposition 1. *Let $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the WSVM problem (2). The point $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star)$ is primal optimal for any weight vector $\boldsymbol{c} \in \mathcal{W}$, and all such weights are contained in $\mathcal{W}$.*

Corollary 1. *There always exists a weight vector $\boldsymbol{c}' \in \mathcal{W}$ such that $\boldsymbol{c}' = \boldsymbol{\alpha}' = \boldsymbol{\alpha}^\star$ and $\boldsymbol{\beta}' = \boldsymbol{0}$.*

It is not surprising that *a posteriori* all weight could be concentrated on support vectors as suggested by Corollary 1. As will become clear in the following, this is close to what the SVM+ algorithm is constrained to do.

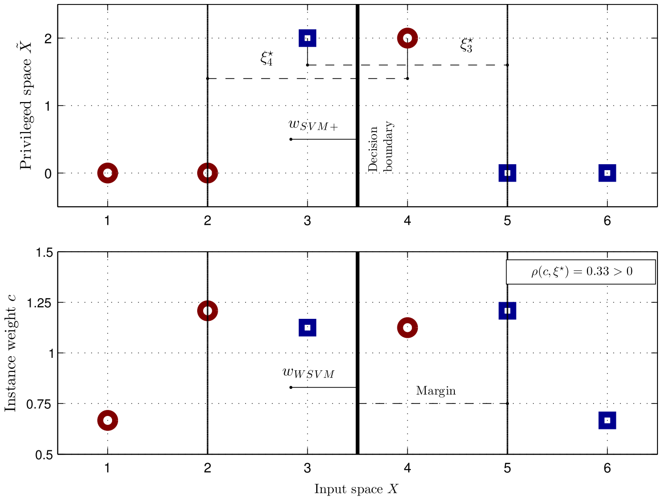

Fig. 1. An example of equivalence between SVM+ (top) and WSVM (bottom). The privileged features coincide with the optimal slack variables $\xi_i^\star$, as motivated by the LUPI paradigm, and instance weights c_i are given by the sum of SVM+ dual variables (Theorem 3). Note that whenever a WSVM solution is constructed from an SVM+ solution, as in this case, the weighted average loss is greater than the non-weighted one, i.e., $\rho(\boldsymbol{c}, \boldsymbol{\xi}^\star) \ge 0$ (Theorem 4).

4. Relation between SVM+ and WSVM

In this section, we present our main theoretical result on the conditions under which the SVM+ and the WSVM are equivalent. Section 4.1 shows that it is always possible to construct weights from an SVM+ solution such that the WSVM will have the same solution. Section 4.2 discusses when it is possible to go in the opposite direction and reveals a fundamental constraint of the SVM+ algorithm. Finally, Section 4.3 states the necessary and sufficient condition for their equivalence. Furthermore, we present a counterexample violating that condition in Section 4.4 and discuss SVM− in Section 4.5.

4.1. SVM+ solutions are also WSVM solutions

The following theorem shows that any SVM+ solution is also a solution to the WSVM problem with appropriately chosen weights and such a choice of weights can always be given by the SVM+ dual variables.

Theorem 3. *Let $(\boldsymbol{w}^\star, b^\star, \tilde{\boldsymbol{w}}^\star, \tilde{b}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the SVM+ problem. There exists a choice of weights $\boldsymbol{c}$, namely $\boldsymbol{c} = \boldsymbol{\alpha}^\star + \boldsymbol{\beta}^\star$, and $\boldsymbol{\xi}^\star$ such that $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ is a primal–dual optimal point for the WSVM problem.*

Note that a direct corollary of this result is that, just like a good choice of privileged features leads to improved predictive performance of the SVM+ (Pechyony & Vapnik, 2010), a good choice of weights leads to improved performance of the WSVM. This claim is verified empirically in the experimental Section 6 when weights are learned in an idealized setting, which is close to the Oracle SVM setting of Vapnik and Vashist (2009).

Fig. 1 shows a toy example where an SVM+ solution is used to compute weights $\boldsymbol{c} = \boldsymbol{\alpha}^\star + \boldsymbol{\beta}^\star$ that force the WSVM to find exactly the same solution. Note that the outliers (points 3 and 4) receive relatively high weight, so that the weighted average loss is greater than the non-weighted one. See Section 4.3 for further details.

4.2. Which WSVM solutions are SVM+ solutions?

We now consider the opposite direction and characterize the SVM+ solutions in terms of the induced instance weights. The following Lemma 1 highlights the bias of the SVM+ algorithm as it establishes that every solution must satisfy a certain relation between the dual variables (respectively the weights) and the loss on the training sample. This is the key to showing that the SVM+ and the WSVM algorithms are not equivalent, and that the latter is strictly more generic as it does not impose that additional constraint.

Lemma 1. *Assume any given $C > 0$, $\gamma \geq 0$ and let $(\boldsymbol{w}^\star, b^\star, \tilde{\boldsymbol{w}}^\star, \tilde{b}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the SVM+ problem (1), then the following holds:*

$$\frac{\sum_{i=1}^{n}(\alpha_i^\star + \beta_i^\star)h_i}{\sum_{i=1}^{n}(\alpha_i^\star + \beta_i^\star)} \geq \frac{1}{n}\sum_{i=1}^{n} h_i, \tag{5}$$

where $h_i := [1 - y_i(\langle \boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + b^\star)]_+$ is the hinge loss at a point $i = 1, \ldots, n$. If $\gamma = 0$, then (5) is satisfied with equality.

Taking into account that the corresponding weights in the WSVM are given by the sum of the SVM+ dual variables, the above inequality can be re-written in a more compact form.

Corollary 2 (*The Necessary Condition*)**.** *Assume the setting of Theorem 3, then $\xi_i^\star = h_i$ and*

$$\langle \mathbf{c} - \bar{c}\mathbf{1}, \boldsymbol{\xi}^\star\rangle \geq 0, \quad \text{where } \bar{c} := n^{-1}\sum_{i=1}^{n} c_i.$$

Proof. Follows from Theorem 3 and Lemma 1. □

Note that this result suggests a simple way to interpret the effect of privileged features—they impose a re-weighting of the input training data. Moreover, at the end of training more emphasis will be on points with positive loss and less on easy points, in particular, the non-support vectors may end up with zero weight.

4.3. SVM+ and WSVM equivalence

We now state the main result of this paper which gives the necessary and sufficient condition for the equivalence between the SVM+ and the WSVM.

Theorem 4. *Let $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}_0^\star, \boldsymbol{\beta}_0^\star)$ be a primal–dual optimal point for the WSVM problem with instance weights $\mathbf{c}_0 \in \mathbb{R}_+^n$, not all zero. There exists a choice of C, γ, and correcting features $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^n$ such that $(\boldsymbol{w}^\star, b^\star)$ is optimal for the SVM+ problem iff:*

$$\exists \mathbf{c} \in \mathcal{W} : \rho(\mathbf{c}, \boldsymbol{\xi}^\star) := \langle \mathbf{c} - \bar{c}\mathbf{1}, \boldsymbol{\xi}^\star\rangle \geq 0, \tag{6}$$

where $\bar{c} := n^{-1}\sum_{i=1}^{n} c_i$. If $\rho(\mathbf{c}, \boldsymbol{\xi}^\star) \geq 0$, one such possible choice is as follows:

$$C = \bar{c}, \qquad \gamma = \rho(\mathbf{c}, \boldsymbol{\xi}^\star), \qquad \tilde{x}_i = \xi_i^\star - \tilde{b}^\star, \quad \forall i \tag{7}$$

moreover, the optimal $\tilde{w}^\star$ and $\tilde{b}^\star$ in that case are:

$$\tilde{w}^\star = 1, \qquad \tilde{b}^\star = \langle \mathbf{c}, \boldsymbol{\xi}^\star\rangle / \langle \mathbf{c}, \mathbf{1}\rangle. \tag{8}$$

Let us make a few remarks. First, condition (6) can be rewritten in terms of averages as

$$\sum_{i=1}^{n} \omega_i \xi_i^\star \geq \frac{1}{n}\sum_{i=1}^{n} \xi_i^\star, \tag{9}$$

where $\omega_i := c_i / \sum_{i=1}^{n} c_i$ is the normalized weight. Hence, any SVM+ solution has an equivalent WSVM setting that puts *more weight on hard examples*, i.e., the points with higher loss.

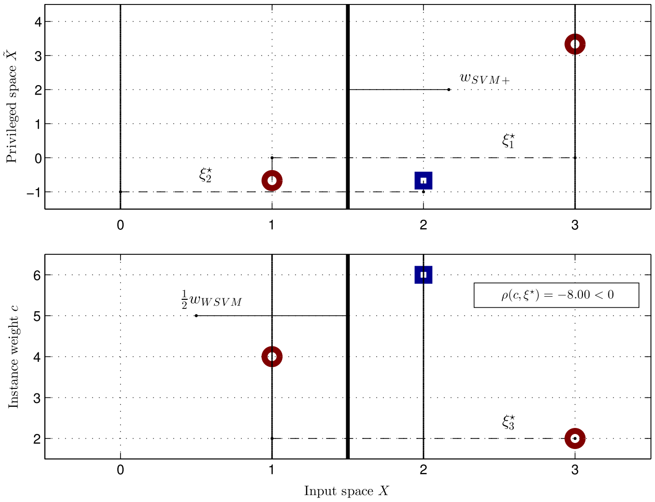

Fig. 2. An example of a WSVM solution (bottom) that cannot be found by SVM+ (top). The instance weights c_i are chosen in a way to avoid a zero norm constant classifier ($f = +1$). The resulting weighted average loss is less than the non-weighted one, hence the SVM+ cannot find this solution. Computing the privileged features as in (7) leads to an SVM+ solution with the opposite prediction and a higher value of the weighted average loss.

Further, it is clear from Definition 1 that the weight of points with $y_i f(\boldsymbol{x}_i) > 1$ can be changed arbitrarily without altering the f since in that case $\xi_i^\star = 0$, $\alpha_i^\star = 0$ and $\beta_i^\star = c_i$, i.e., these points are not support vectors and they have no influence on the final classifier. Hence, their weight – the upper bound on the influence – does not matter.

This reasoning leads us to a condition that is much easier to check in practice than the one in Theorem 4. Note that condition (6) involves the set of equivalent weights and it *is* possible to check it directly using the definition of $\mathcal{W}$ as will be discussed below. However, if the kernel matrix is non-singular, as is often the case with the Gaussian kernel, then one can simply take $\mathbf{c} = \boldsymbol{\alpha}^\star$ and check (6) for that particular weight vector *only*.

Proposition 2. *Let $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the WSVM problem with instance weights $\mathbf{c} \in \mathbb{R}_+^n$, not all zero. If*

$$\mathcal{N}(\mathbf{YKY}) \cap \mathbf{1}^\perp \cap \boldsymbol{y}^\perp = \{\mathbf{0}\},$$

then there exists a choice of C, γ, and $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^n$ such that $(\boldsymbol{w}^\star, b^\star)$ is optimal for the SVM+ problem iff:

$$\rho(\boldsymbol{\alpha}^\star, \boldsymbol{\xi}^\star) = \boldsymbol{\xi}^{\star\top}\big(\mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big)\boldsymbol{\alpha}^\star \geq 0. \tag{10}$$

Intuitively, the SVM+ algorithm maximizes the margin $2\,\|\boldsymbol{w}\|^{-1}$ by minimizing $F(\boldsymbol{\alpha})$, as in the standard SVM, and also gradually shifts focus to hard examples by minimizing $\tilde{F}(\tilde{\boldsymbol{\alpha}})$. As long as there are sufficiently many points on the "right" side of the margin, (9) can be achieved by reducing the weight of such non-support vectors, and so the SVM+ solution space is as rich as that of the WSVM. In general, however, (9) may not be attainable without altering the f as demonstrated by the counter example below.

4.4. WSVM solution not found by SVM+

We now consider the case when misclassified training points have low weight, i.e., $\rho(\mathbf{c}, \boldsymbol{\xi}^\star) < 0$, and give an example where SVM+ fails to find the corresponding WSVM solution.

Consider the training sample below (Fig. 2):

$$S = \{(1, +1), (2, -1), (3, +1)\}, \qquad \mathbf{c} = (4, 6, 2)^\top.$$

The corresponding primal–dual optimal point is

$$w^\star = -2, \qquad b^\star = 3, \qquad \boldsymbol{\xi}^\star = (0, 0, 4)^\top,$$
$$\boldsymbol{\alpha}^\star = (4, 6, 2)^\top, \qquad \boldsymbol{\beta}^\star = (0, 0, 0)^\top.$$

Since $\rho(\boldsymbol{c}, \boldsymbol{\xi}^\star) = -\frac{2}{3} < 0$, this solution does not correspond to any of the SVM+ solutions (Lemma 1). Note that one can easily verify that $\mathcal{N}(\mathbf{YKY}) \cap \mathbf{1}^\perp \cap \boldsymbol{y}^\perp$ contains only $\mathbf{0}$, hence, Proposition 2 already completes the claim. Similarly, one can show using Definition 1 that $\mathcal{U} = \{\boldsymbol{\alpha}^\star\}$ and that other equivalent weights can only increase the weight of points 1 and 2, which would only decrease $\rho(\boldsymbol{c}, \boldsymbol{\xi}^\star)$. Therefore, there is no $\boldsymbol{c}' \in \mathcal{W}$ for which $\rho(\boldsymbol{c}', \boldsymbol{\xi}^\star) \geq 0$ and, by Theorem 4, there is no correcting space that would make $(w^\star, b^\star) = (-2, 3)$ an SVM+ solution.

Fig. 2 shows the learned WSVM and SVM+ models, where we used $\tilde{x}_i = \xi_i^\star - \langle \boldsymbol{c}, \boldsymbol{\xi}^\star \rangle / \langle \boldsymbol{c}, \mathbf{1} \rangle$, $C = \bar{c}$, $\gamma = 1$. A different choice of C and γ can make SVM+ return a constant classifier, which is the solution of the standard SVM, but there is no setting that would make it return $(w^\star, b^\star) = (-2, 3)$.

Note that in this example an even stronger result can be shown: SVM+ cannot reproduce the same *type* of dichotomy, i.e., even if we allowed it to return a line with *any* negative slope going through the same point, the SVM+ would still fail. This shows that there are settings where WSVM performs significantly better than SVM+ due to a fundamental constraint of the latter.

4.5. Is there an SVM−?

We have seen that the SVM+ has a more constrained solution space than the weighted SVM. Lemma 1 gives the exact characterization of that constraint in terms of the relation between the SVM+ dual variables and the incurred loss on the training sample. The WSVM solution space can thus be partitioned into solutions that can be found by SVM+ and the rest. We are now interested if there is a modification to the SVM+ algorithm that would yield solutions from that second part.

Theorem 4 suggests that $\gamma = \rho(\boldsymbol{c}, \boldsymbol{\xi}^\star) \geq 0$, so, intuitively, if we now require $\rho(\boldsymbol{c}, \boldsymbol{\xi}^\star) < 0$, the corresponding γ has to be with a minus:

$$\begin{aligned} \min_{\boldsymbol{w}, b, \tilde{\boldsymbol{w}}, \tilde{b}} \quad & \frac{1}{2}\left(\|\boldsymbol{w}\|^2 - \gamma \left\|\tilde{\boldsymbol{w}}\right\|^2\right) + C \sum_{i=1}^{n} \xi_i(\tilde{\boldsymbol{w}}, \tilde{b}) \\ \text{s.t.} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle + b) \geq 1 - \xi_i(\tilde{\boldsymbol{w}}, \tilde{b}), \qquad \xi_i(\tilde{\boldsymbol{w}}, \tilde{b}) \geq 0. \end{aligned} \tag{11}$$

This problem is clearly non-convex as the objective is now a difference of convex functions. If there was a finite (local) minimizer $(\boldsymbol{w}^\star, b^\star, \tilde{\boldsymbol{w}}^\star, \tilde{b}^\star)$, the KKT conditions would still hold (Borwein & Lewis, 2000, Theorem 2.3.8) for a Lagrange multiplier vector $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$, and one could show a result similar to Lemma 1, but with the reverse inequality.

Unfortunately, however, the problem (11) is unbounded below, which is easy to see: the quadratic term $\left\|\tilde{\boldsymbol{w}}\right\|^2$ grows faster than the linear term $\xi_i(\tilde{\boldsymbol{w}}, \tilde{b})$ and the feasible set is unbounded. This shows that it is not trivial to modify the SVM+ algorithm to obtain solutions from its complement, and it is an open question if such a modification (with non-degenerate solutions) exists at all.

The phenomenon we observe here is that some of the WSVM solutions ($\rho(\boldsymbol{c}, \boldsymbol{\xi}^\star) \geq 0$) can be computed easily within the LUPI framework, while others ($\rho(\boldsymbol{c}, \boldsymbol{\xi}^\star) < 0$) may be completely out of reach. What are the implications of this observation in terms of learning a classifier?

Consider any training sample S of size n for a problem P. Let $f_{\boldsymbol{c},S}$ be a classifier constructed by the WSVM with weights $\boldsymbol{c}$, and let $\boldsymbol{\xi}^\star_{\boldsymbol{c},S}$ be the corresponding loss vector. The set of all admissible weights $\mathbb{R}^n_+$ is partitioned into two subsets, $\mathcal{W}_+$ and $\mathcal{W}_-$, depending on the sign of $\rho(\boldsymbol{c}, \boldsymbol{\xi}^\star_{\boldsymbol{c},S})$. Define the "best" weight vectors in each of the two classes as $\boldsymbol{c}_\pm = \arg\min_{\boldsymbol{c} \in \mathcal{W}_\pm} L(f_{\boldsymbol{c},S})$. If $L(f_{\boldsymbol{c}_-,S}) < L(f_{\boldsymbol{c}_+,S})$, then the best classifier corresponds to the weights that are out of reach for the SVM+, hence, there are no privileged features that will yield an SVM+ classifier as good as $f_{\boldsymbol{c}_-,S}$.

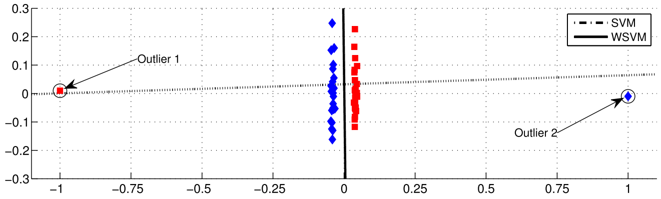

Fig. 3. Illustration of the effect of instance weighting on a toy problem in 2D. Even though the problem is (almost) linearly separable, the two outliers in the training set cause the SVM to have a near chance level performance (horizontal line). Assigning zero weight to the outliers allows the WSVM to recover a near optimal solution (vertical line).

This reasoning motivated us to consider weight generation schemes that are unrelated to SVM+ and which are discussed next.

5. How to choose the weights

Recall that we are interested in ways of incorporating prior knowledge about the *training data*. In the SVM+ approach, the role of additional information is played by the privileged features which are used to estimate the loss on the training sample. The same effect, as we have established, can be achieved by importance weighting. Taking into account the vast amount of work on weighted learning, it seems that re-weighting of misclassification costs is a very powerful method of incorporating prior knowledge. We would like to stress, however, that a critical difference to, e.g., the cost-sensitive learning is that we are ultimately interested in minimizing the *non-weighted* expected loss and *the weights are only used to impose a bias on the learning algorithm*.

We also note that even though the SVM+ solutions are contained within the WSVM solutions, there is *no* implication that any of the two algorithms is "better". If privileged features are available, then SVM+ is a reasonable choice. On the other hand, if there are no privileged features or if one has concerns outlined at the end of Section 4.5, then one may want to consider a more general WSVM with some problem specific scheme for computing weights.

In the following, we investigate two approaches that make different assumptions about what is additionally available to the learning algorithm at training time. The methods operate in a somewhat idealized setting and are mainly aimed at motivating further research on how to choose the weights. They may be thought of as the empirical counterparts of a more theoretical discussion involving the Oracle SVM in Vapnik and Vashist (2009). In particular, the weight learning method of Section 5.3 can be thought of as a way of extracting additional information about the given training sample from a validation sample which is used as a reference.

5.1. Why instance weighting is important?

Let us first motivate why instance weighting can be very important in certain problems.

Consider the toy problem shown in Fig. 3. The data comes from two linearly separable blobs, so it is possible to achieve zero test error on them. However, the training sample has been contaminated with two outliers that lie extremely far from the optimal decision boundary. Since the SVM uses a surrogate loss and not the 0–1 loss, the cost of a point is higher the further the point is from the separating hyperplane. Hence, the SVM "prefers" to keep the two outliers close to the decision boundary, which leads to a near chance level performance on this data set. Instance weighting, on

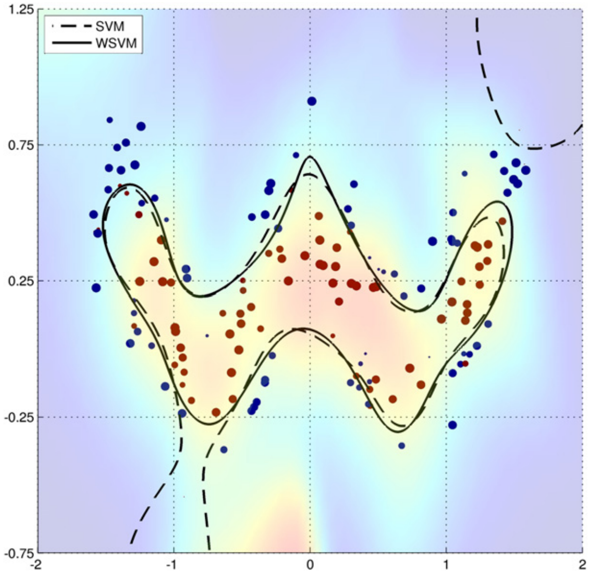

Fig. 4. Importance weighting leads to a more stable estimate of the decision boundary in a non-linear 2D problem. The size of a data point corresponds to its weight, which is computed from an estimate of $P(Y = 1|X)$ shown in background. The WSVM (solid line) is less influenced by outliers than SVM (dashed line) since the outliers are downweighted, which ultimately results in better predictive performance.

the other hand, allows one to alter the cost of each point. In particular, if the two outliers are assigned zero weight, then the WSVM is able to find a near optimal classifier.

The second toy problem shown in Fig. 4 suggests that an estimate of $P(Y = 1|X)$ could be used to compute instance weights and improve predictive performance even in the non-linear case, where the aforementioned problem of extreme outliers is less likely to happen. As before, the issue evolves around the points that lie either too close to or even on the wrong side of the true decision boundary. We used the standard Nadaraya–Watson estimator (22) to obtain an estimate of the conditional probability (shown in background), which was then used to compute instance weights (reflected by the size of points) using the formula (13) introduced below. Note that the outliers are downweighted and have less influence on the WSVM decision boundary (solid line) than on the SVM one (dashed line). That leads to better accuracy, as reported in Section 6.2.

5.2. Access to an estimate of $P(Y = 1|X)$

Clearly, having full access to the conditional probability $P(1|X)$ is a hypothetical scenario since in this case the classification problem is solved. However, it is interesting to see how this type of information could be used in construction of good weights. As the first step, we note that if $P(1|X)$ were available at least for the *training* points one could directly compute the conditional expectation and employ the following estimator

$$L'(f) := \frac{1}{n}\sum_{i=1}^{n}\left[\ell(f(X_i))P(1|X_i) + \ell(-f(X_i))P(-1|X_i)\right],$$

which is an unbiased estimator of $L(f)$:

$$\begin{aligned}\mathbb{E}L'(f) &= \mathbb{E}\left[\ell(f(X))P(1|X) + \ell(-f(X))P(-1|X)\right] \\ &= \mathbb{E}\mathbb{E}\left[\ell(Yf(X))|X\right] = \mathbb{E}\ell(Yf(X)) = L(f).\end{aligned}$$

The property of being biased or not is of asymptotic nature and is arguably of lesser interest in the small sample regime. Following this line of argument, we consider a conservative weighted estimator given by:

$$\hat{L}_w(f) := \frac{1}{n}\sum_{i=1}^{n} w(X_i, Y_i)\ell(Y_i f(X_i)), \tag{12}$$

$$w(X_i, Y_i) := P(Y = Y_i|X = X_i). \tag{13}$$

It is not hard to check that $\hat{L}_w(f)$ is biased:

$$\begin{aligned}\mathbb{E}\hat{L}_w(f) &= \mathbb{E}\mathbb{E}\left[w(X, Y)\ell(Yf(X))|X\right] \\ &= \mathbb{E}\left[\ell(f(X))P(1|X)^2 + \ell(-f(X))P(-1|X)^2\right] \\ &\leq \mathbb{E}\left[\ell(f(X))P(1|X) + \ell(-f(X))P(-1|X)\right] \\ &= \mathbb{E}\mathbb{E}\left[\ell(Yf(X))|X\right] = L(f).\end{aligned}$$

More precisely, $\hat{L}_w(f)$ is *conservative* in the sense that the points far from the decision boundary are upweighted, while the points with $P(1|X) \approx 0.5$ receive relatively low weight. This behavior is due to the $p \mapsto p^2$ transform which is monotonically increasing and is strictly convex on $[0, 1]$. The monotonicity also ensures the following important property of the obtained estimator when ℓ is the 0–1 loss:

$$\arg\min_f \mathbb{E}\hat{L}_w(f) = f^* = \arg\min_f \mathbb{E}L(f),$$

that is, the $\hat{L}_w$ is minimized by the Bayes classifier and *the learning problem is not changed*.

If the bias of $\hat{L}_w$ is a concern, one can let the weights decay to one as the size of the training sample increases. To this end, we consider the following generalization of the weight function in (13):

$$c_\tau(X_i, Y_i) := w^\tau(X_i, Y_i), \tag{14}$$

where $\tau \in [0, \infty)$ is tuned along with the standard regularization parameter. Note that SVM is recovered when the weights are given by $c_0(X_i, Y_i) \equiv 1$.

When $P(Y = 1|X)$ is estimated from a training sample, the WSVM with weights given by (14) will mainly serve as a baseline for the method introduced in the following section. However, it is conceivable that an estimate of $P(Y = 1|X)$ could be available from a different source, e.g., from annotations provided by humans. The latter setting is evaluated in Section 6.4.

5.3. Learning the weights

Given a fixed training sample S, the weights in a weighted SVM parametrize the set of hypotheses that the WSVM can choose from. Hence, they could be learned within the standard framework of risk minimization with the additional twist that the classifier f depends on the weights $\boldsymbol{c}$ *implicitly*:

$$\boldsymbol{c}^\star = \arg\min_{\boldsymbol{c}\in\mathbb{R}^n_+} \mathbb{E}\ell(Yf_{\boldsymbol{c}}(X)), \tag{15}$$

$$f_{\boldsymbol{c}} = \arg\min_{f\in\mathcal{H}} \frac{1}{2}\|f\|^2 + \sum_{i=1}^{n} c_i\ell(y_i f(\boldsymbol{x}_i)). \tag{16}$$

Clearly, the optimization problem (15) cannot be solved in practice since the underlying probability distribution is unknown, hence, we replace $L(f)$ in (15) with an estimator. The latter, however, has to be different from the estimator $\hat{L}_w$ in (16) to avoid overfitting. We follow a simple approach and assume that a second sample S' is available at training time. The problem (15) is thus replaced with

$$\boldsymbol{c}^\star = \arg\min_{\boldsymbol{c}\in\mathbb{R}^n_+} \sum_{i=1}^{N} \ell(y'_i f_{\boldsymbol{c}}(\boldsymbol{x}'_i)).$$

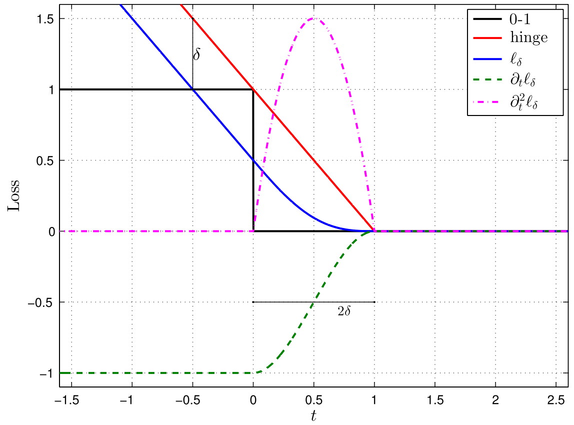

Fig. 5. A 0–1 loss, a hinge loss, an approximate hinge loss ℓ_δ, and its first two derivatives.

This idea follows the method of Chapelle, Vapnik, Bousquet, and Mukherjee (2002) who suggested to tune L2-SVM parameters by minimizing certain estimates of the generalization error using a gradient descent algorithm. The use of the L_2 penalization of the training errors allows one to additionally assume the hard margin case which leads to a very specific derivation of the gradient w.r.t. the parameters. Instead, we proceed with a different approach and use a smooth version of the hinge loss given below in (21). Furthermore, we optimize (16) in the primal as suggested by Chapelle (2007). The weight learning problem can thus be stated as follows.

$$\boldsymbol{c}^\star = \operatorname*{arg\,min}_{\boldsymbol{c}\in\mathbb{R}^n_+} \sum_{i=1}^{N} \ell(y'_i[\bar{K}_i^\top \boldsymbol{\alpha}^\star(\boldsymbol{c}) + b^\star(\boldsymbol{c})]), \tag{17}$$

$$\begin{bmatrix} \boldsymbol{\alpha}^\star \\ b^\star \end{bmatrix} = \operatorname*{arg\,min}_{\boldsymbol{\alpha}, b} \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \sum_{i=1}^{n} c_i \ell(y_i[K_i^\top \boldsymbol{\alpha} + b]), \tag{18}$$

where $\bar{\mathbf{K}}$ is the matrix with $\bar{\mathbf{K}}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}'_j)$, and K_i, $\bar{K}_i$ are the ith columns of $\mathbf{K}$ and $\bar{\mathbf{K}}$.

Note that f depends on the weights implicitly via the second optimization problem and the main challenge in applying the gradient descent is the computation of $\partial\boldsymbol{\alpha}^\star/\partial\boldsymbol{c}$ and $\partial b^\star/\partial\boldsymbol{c}$. These can be computed via implicit differentiation from the optimality conditions as shown below.

Theorem 5. *Let the loss function ℓ be convex and twice continuously differentiable and let the kernel matrix $\mathbf{K}$ be (strictly) positive definite. Define vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ componentwise for $i = 1, \ldots, n$ as*

$$u_i := y_i \ell'(y_i[K_i^\top \boldsymbol{\alpha}^\star + b^\star]),$$
$$v_i := c_i \ell''(y_i[K_i^\top \boldsymbol{\alpha}^\star + b^\star]),$$

where $(\boldsymbol{\alpha}^\star, b^\star)$ is a solution of (18) for a given $\boldsymbol{c}$. If $\boldsymbol{v} \neq \mathbf{0}$, then the solution is unique, $\boldsymbol{\alpha}^\star$ and $b^\star$ are continuously differentiable w.r.t. $\boldsymbol{c}$ and the corresponding gradient can be computed as follows.

$$\begin{bmatrix} \dfrac{\partial\boldsymbol{\alpha}^\star}{\partial\boldsymbol{c}} \\ \dfrac{\partial b^\star}{\partial\boldsymbol{c}} \end{bmatrix} = -\begin{bmatrix} \mathbf{I} + \operatorname{diag}(\boldsymbol{v})\mathbf{K} & \boldsymbol{v} \\ \mathbf{1}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \operatorname{diag}(\boldsymbol{u}) \\ \mathbf{0}^\top \end{bmatrix}. \tag{19}$$

Note that this result can be directly applied to such popular loss functions as the squared hinge loss and the logistic loss, and for the latter it will always hold that $\boldsymbol{v} \neq \mathbf{0}$ unless all weights are zero. If $\boldsymbol{v} = \mathbf{0}$, it can be seen that $\boldsymbol{\alpha}^\star$ is still uniquely defined and is continuously differentiable w.r.t. $\boldsymbol{c}$ if b is considered fixed. The "gradient" in this case is given by

$$\partial\boldsymbol{\alpha}^\star/\partial\boldsymbol{c} = \operatorname{diag}(\boldsymbol{u}), \qquad \partial b^\star/\partial\boldsymbol{c} = \mathbf{0}^\top. \tag{20}$$

Ideally, to be consistent with the discussion about the relation between the SVM+ and the WSVM, we would have to consider the hinge loss in the weight learning problem. However, the hinge loss is not differentiable and Theorem 5 does not apply. Instead, we consider a differentiable approximation of the hinge loss that preserves certain desirable properties of the latter. We have chosen the loss function defined as follows (Fig. 5).

$$\ell_\delta(t) := \begin{cases} 1 - t - \delta & \text{if } t \le 1 - 2\delta \\ \dfrac{(1-t)^3(t - 1 + 4\delta)}{16\delta^3} & \text{if } 1 - 2\delta < t < 1 \\ 0 & \text{if } t \ge 1. \end{cases} \tag{21}$$

Note that, unlike certain other approximations, this function is twice continuously differentiable. Like the hinge loss, it does not penalize points with the margin $t := y_i f(\boldsymbol{x}_i) \ge 1$ and it grows linearly for $t \le 1 - 2\delta$.

With the approximate hinge loss ℓ_δ defined above, the $\boldsymbol{v} \neq \mathbf{0}$ means that at least one of the data points has to fall into the strictly convex region of the loss. Clearly, this presents us with a tradeoff between having a good approximation of the hinge loss (small δ) and a higher chance of being able to compute "correct" gradients and thus make substantial progress in the optimization problem (large δ). We resolve the tradeoff by tuning $\delta \in [0.01, 1]$ on a validation set.

6. Experiments

In this section we present empirical evaluation of the algorithms considered in this paper. In our experiments, we used the implementation of the WSVM by Chang and Lin (2011) and the code for the SVM+ provided by Pechyony and Vapnik (2011). The weight learning problem was solved using our implementation of the BFGS algorithm (Nocedal & Wright, 2006). The general experimental setup is similar to that of Vapnik and Vashist (2009): parameters are tuned on a validation set, which is not used for training, and performance is evaluated on a test set. Training subsets are randomly sampled from a fixed training set, and results over multiple runs are aggregated showing the mean error rate as well as the standard deviation. Depending on the experiment, the validation set is either fixed or subsampled randomly as well. The Gaussian RBF kernel is used in all of the experiments and features are rescaled to be in [0, 1]. The weights in (13) are computed from $\eta(\boldsymbol{x}) = 2\mathrm{P}(1|\boldsymbol{x}) - 1$, which is either given directly by human experts or estimated via:

$$\eta(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i) y_i}{\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i)}, \tag{22}$$

where K_h is the Gaussian kernel with bandwidth h.

Note that in all experiments each algorithm has access to exactly *the same data*, and the only difference between different splits is which data is used to construct a classifier (training) and which is used to tune the hyper-parameters (validation).

6.1. WSVM replicates SVM+

We begin with the experimental verification of our theoretical findings of Section 4. We reproduced the handwritten digit recognition experiment of Vapnik et al. (2009), where the task is to discriminate between 5's and 8's taken from the MNIST database and downsized to 10 × 10 pixels. We used the features provided by the

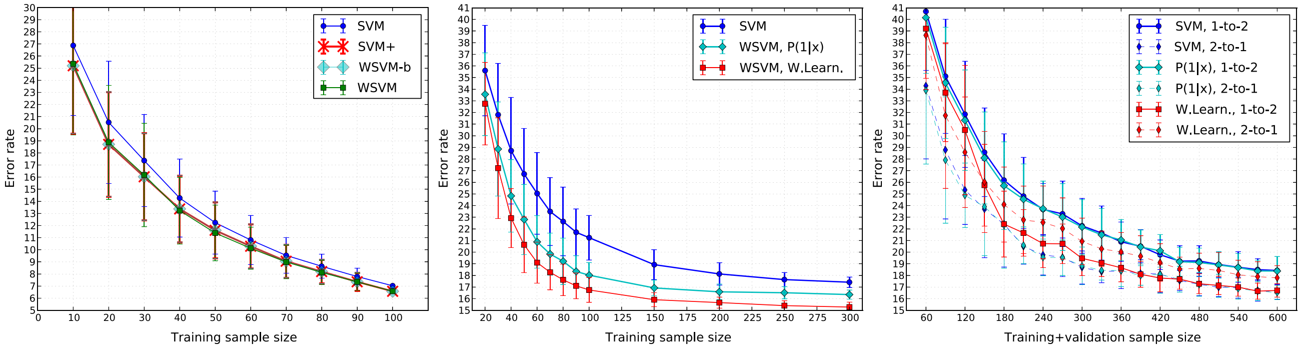

Fig. 6. SVM, SVM+, and WSVM error rates. Left: Reproduction of the experiment of Vapnik et al. (2009). The SVM+ and the WSVM classifiers coincide up to the non-uniqueness of b. Middle: Instance weighting leads to significant performance improvement when a large validation set is available (toy data). Right: Similar setting, but the training-to-validation splits are 1-to-2 and 2-to-1.

authors and obtained similar error rates for both the SVM and the SVM+, see Fig. 6, left. Our results are averaged over 100 runs and include more subsets.

The weights for the WSVM algorithm were computed as $\boldsymbol{c} = \boldsymbol{\alpha}^{\star} + \boldsymbol{\beta}^{\star}$, where $\boldsymbol{\alpha}^{\star}$ and $\boldsymbol{\beta}^{\star}$ come from the SVM+ solution. We observed that $\boldsymbol{\alpha}^{\star}_{\text{WSVM}} \approx \boldsymbol{\alpha}^{\star}_{\text{SVM+}}$. However, we also observed that, in general, $b^{\star}_{\text{WSVM}} \neq b^{\star}_{\text{SVM+}}$, which is explained by the non-uniqueness of b (Theorem 1). If $b^{\star}_{\text{SVM+}}$ from the SVM+ model is used (WSVM-b in the plot), then the two classifiers are identical, but if b is tuned within the constraints imposed by the KKT conditions (WSVM in the plot), then minor differences appear.

6.2. *Toy data*

We now turn to the problem of choosing weights and evaluate the two weight generation schemes introduced in Section 5. In this experiment, data comes from a mixture of 2D Gaussians that form a non-linear shape resembling "W", see Fig. 4. Similar to the previous setting, we sample from a fixed training set of size 400, tune parameters, estimate the $\mathrm{P}(1|\boldsymbol{x})$, and perform weight learning on a validation set of size 4000, and test on a separate set of size 2000. The results are averaged over 50 runs, see Fig. 6, middle. Note that, just like in the experiment of Vapnik et al. (2009), this is an idealistic setting where the validation set is so large that model selection is close to optimal. In practice, one would never split the available sample as 1-to-40, therefore we also evaluate more "reasonable" splits 1-to-2 and 2-to-1 next.

Fig. 6, right, shows results of a similar experiment where the validation sample is not fixed, but rather obtained by splitting the available training data. Since validation samples are now small, the estimation of $\mathrm{P}(1|\boldsymbol{x})$ fails and the corresponding WSVM performs on par with the standard SVM. The weight learning, however, still yields performance improvement on 1-to-2 splits. Moreover, the WSVM with weight learning is able to achieve a similar error rate as the SVM trained on twice as much data. We also observe the effect of overfitting when weight learning is performed on 2-to-1 splits, and we omit it in further experiments. Note that one could have anticipated that for the weight learning to succeed the amount of validation data, in general, has to be at least comparable to or larger than the number of weights that are to be learned.

6.3. *UCI data sets*

In this set of experiments we evaluate weight learning on three data sets from the UCI repository (Frank & Asuncion, 2010). For every data set, we first remove any records with missing values and then split the remaining data randomly into training and test sets of roughly equal size approximately preserving the initial class distribution. Table 1 summarizes characteristics of the obtained data sets. Smaller subsets are then sampled from the training data and split into training and validation sets as 1-to-2 and 2-to-1. The subsets sampling process is repeated 20 to 50 times depending on the amount of data. The rest of the experimental setup is the same as before.

Table 1
Statistics of data sets from the UCI repository.

Data set	Features	Training	Test
BCW	9	351	332
Mammographic	4	420	410
Spambase	57	2430	2171

Breast Cancer Wisconsin (BCW) (Bennett & Mangasarian, 1992): On this data set, the weight learning on the 1-to-2 split performs on par or better than the SVM on both splits, see Fig. 7. Notably, the SVM performed worse on the 2-to-1 split, which we attribute to overfitting. The latter is not too surprising considering the small amount of data and the capacity of the RBF kernel, which makes the weight learning result even more remarkable.

Mammographic Mass (Elter, Schulz-Wendtland, & Wittenberg, 2007): Again, the weight learning performs on par or better than the SVM on all splits for almost all subsets. On the last subset, however, the weight optimization did not yield any improvement, and the resulting performance is the same as that of the corresponding SVM.

Spambase: On this data set, the general outcome is that the weight learning brings roughly the same level of improvement as if twice as much data were used for training the standard SVM. This can be interpreted as a more efficient use of training data given the additional knowledge about importance of each data point.

6.4. *Handwritten digit recognition (5's vs. 8's)*

Finally, we get back to the original handwritten digit recognition experiment of Vapnik et al. (2009) and evaluate our weight generation schemes on that data.

In this experiment, we evaluate the first weight generation scheme (14) under the assumption that digit ranking is available as the additional information, i.e., in addition to the class label ±1, we are also given a confidence score between −1 and 1. This is a reasonable assumption e.g. for data sets where robust annotation is obtained by aggregation of labeling from several human experts and is similar to the setting considered by Wu and Srihari (2004).

We collected additional annotation in the form of ranking from three human experts. The humans were presented with a random sample of the 10×10 pixel digits and were asked to label them using one of 5 possible labels, which we translated to a score in $\{-1, -0.5, 0, 0.5, 1\}$. Each of the 100 digits from the training set was ranked 16 times and the average score was then used as an estimate of $2\mathrm{P}(1|x) - 1$.

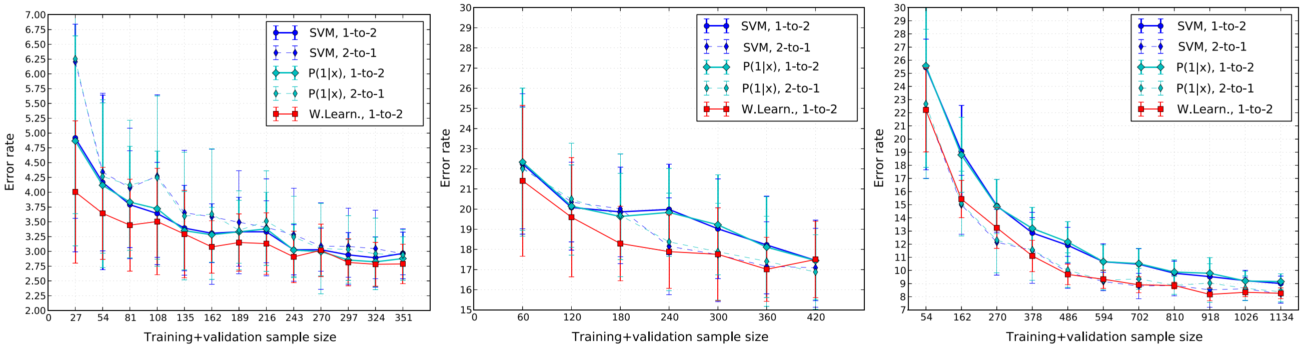

Fig. 7. SVM and WSVM error rates on the UCI repository data sets with training-to-validation splits of 1-to-2 and 2-to-1. Left: Breast Cancer Wisconsin. Middle: Mammographic Mass. Right: Spambase.

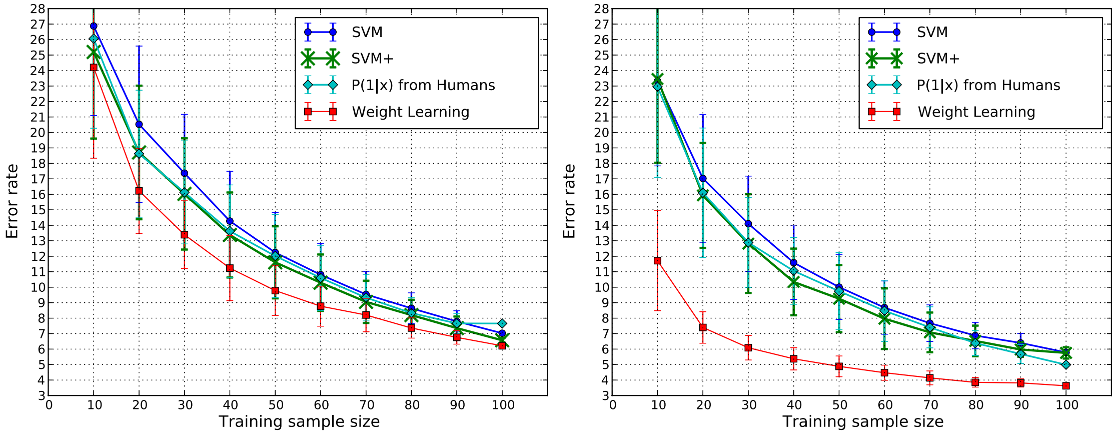

Fig. 8. Error rate comparison in the handwritten digit recognition experiment of Vapnik et al. (2009). P(1|$\boldsymbol{x}$) was estimated from human rankings. Left: The original setting. Right: The extended setting where each digit is translated by 1 pixel in each of the 8 directions.

Fig. 8 shows the corresponding experimental results. We observe that additional information from human experts helps on small subsets, but its influence degrades on larger subsets. This might be in part due to the difference in image representation used by SVMs and humans. In particular, humans' recognition of digits is translation invariant, while the pixel-wise representation is not. This leads us to our final experiment on the extended version of that data set.

We extend the original training sample of 100 digits by shifting each digit by 1 pixel in all 8 directions, thus obtaining 9 times the initial sample size. We assume that both the human rankings and the privileged features from the experiment of Vapnik et al. (2009) are unaffected by such translations and we simply replicate them. The experimental results are presented in Fig. 8, right. Note that the WSVM with human rankings is now consistently on par or better than SVM and is somewhat comparable to SVM+.

Remarkably, weight learning now gives significant performance boost on the extended version of the data set, which shows that it can be successfully combined with other sources of additional information, like the hint about translation invariance in this case. Interestingly enough, Lauer and Bloch (2008) discussed the possibility of combining the virtual sample method, which we used to extend the training set, with weighted learning where each virtual point would be given a confidence score c_i. Our weight learning algorithm does exactly that, but without trying to model the measure of confidence. Instead, it attempts to directly optimize an estimate of the expected loss $L(f)$.

7. Conclusion

We have investigated basic properties of the recently proposed SVM+ algorithm, such as uniqueness of its solution, and have shown that it is closely related to the well-known weighted SVM. We revealed that all SVM+ solutions are constrained to have a certain dependency between the dual variables and the incurred loss on the training sample, and that the prior knowledge from the SVM+ framework can be encoded via instance weights.

That motivated us to consider other sources of additional information about the training data than the one given by privileged features. In particular, we considered the weight learning method in Section 5.3 which allows one to learn weights directly from data (using a validation set). The latter approach is not limited to SVMs and can be extended to other classifiers.

Experimental results confirmed our intuition that importance weighting is a powerful method of incorporating prior knowledge. In the idealized setting, we showed that the weight learning works and yields significant performance improvement. The choice of weights in a more practical setting is left for future work.

Appendix A. The KKT conditions

In convex optimization, the Karush–Kuhn–Tucker (KKT) conditions are necessary and sufficient for a point to be primal and dual optimal with zero duality gap (Boyd & Vandenberghe, 2004).

The KKT conditions corresponding to the SVM+ problem (1) are given below:

$$\sum_{i=1}^{n} \alpha_i^\star y_i \boldsymbol{z}_i = \boldsymbol{w}^\star, \tag{A.1a}$$

$$\sum_{i=1}^{n} \alpha_i^\star y_i = 0, \tag{A.1b}$$

$$\sum_{i=1}^{n}(\alpha_i^\star + \beta_i^\star - C)\tilde{\boldsymbol{z}}_i = \gamma\tilde{\boldsymbol{w}}^\star, \tag{A.1c}$$

$$\sum_{i=1}^{n}(\alpha_i^\star + \beta_i^\star - C) = 0, \tag{A.1d}$$

$$\alpha_i^\star[\langle\tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b}^\star - 1 + y_i(\langle\boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + b^\star)] = 0, \tag{A.1e}$$

$$\beta_i^\star[\langle\tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b}^\star] = 0, \tag{A.1f}$$

$$\langle\tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b}^\star - 1 + y_i(\langle\boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + b^\star) \geq 0, \tag{A.1g}$$

$$\alpha_i^\star \geq 0, \qquad \beta_i^\star \geq 0, \qquad \langle\tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b}^\star \geq 0. \tag{A.1h}$$

And the KKT conditions corresponding to the weighted SVM problem (2) are as follows:

$$\sum_{i=1}^{n}\alpha_i^\star y_i \boldsymbol{z}_i = \boldsymbol{w}^\star, \tag{A.2a}$$

$$\sum_{i=1}^{n}\alpha_i^\star y_i = 0, \tag{A.2b}$$

$$\alpha_i^\star + \beta_i^\star = c_i, \tag{A.2c}$$

$$\alpha_i^\star[\xi_i^\star - 1 + y_i(\langle\boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + b^\star)] = 0, \tag{A.2d}$$

$$\beta_i^\star[\xi_i^\star] = 0, \tag{A.2e}$$

$$\xi_i^\star - 1 + y_i(\langle\boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + b^\star) \geq 0, \tag{A.2f}$$

$$\alpha_i^\star \geq 0, \qquad \beta_i^\star \geq 0, \qquad \xi_i^\star \geq 0. \tag{A.2g}$$

Appendix B. Technical proofs

B.1. Proof of Theorem 2

Theorem. *The solution to the problem (1) is unique in $(\boldsymbol{w}, \tilde{\boldsymbol{w}}, \tilde{b})$ for any $C > 0, \gamma > 0$. If there is a support vector, then b is unique as well, otherwise:*

$$\max_{i\in\mathcal{I}_+}(1 - \langle\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}_i\rangle - \tilde{b}) \leq b \leq \min_{i\in\mathcal{I}_-}(\langle\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b} - 1).$$

Proof. Following Burges and Crisp (1999), let F be the objective function:

$$F = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{2}\left\|\tilde{\boldsymbol{w}}\right\|^2 + C\sum_{i=1}^{n}(\langle\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b}),$$

and define $\boldsymbol{u} := (\boldsymbol{w}, \tilde{\boldsymbol{w}}, \tilde{b})^\top$. Suppose $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are two solutions, then, since the problem is convex, a family of solutions is given by $\boldsymbol{u}_t = (1-t)\boldsymbol{u}_1 + t\boldsymbol{u}_2, t \in [0, 1]$, and $F(\boldsymbol{u}_1) = F(\boldsymbol{u}_2) = F(\boldsymbol{u}_t)$. Expanding $F(\boldsymbol{u}_t) - F(\boldsymbol{u}_1) = 0$ and differentiating w.r.t. t yields:

$$(t-1)\|\boldsymbol{w}_1\|^2 + (1-2t)\langle\boldsymbol{w}_1, \boldsymbol{w}_2\rangle + t\|\boldsymbol{w}_2\|^2 + \gamma[(t-1)\left\|\tilde{\boldsymbol{w}}_1\right\|^2$$
$$+ (1-2t)\langle\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2\rangle + t\left\|\tilde{\boldsymbol{w}}_2\right\|^2]$$
$$+ tC\sum_{i=1}^{n}(\langle\tilde{\boldsymbol{w}}_2 - \tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b}_2 - \tilde{b}_1) = 0,$$
$$\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2 + \gamma\left\|\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}_2\right\|^2 = 0.$$

Since $\gamma > 0$ it follows that $\boldsymbol{w}_1 = \boldsymbol{w}_2$ and $\tilde{\boldsymbol{w}}_1 = \tilde{\boldsymbol{w}}_2$. Plugging that into the first equation yields $\tilde{b}_2 = \tilde{b}_1$. Uniqueness of b now follows from condition (A.1e). If all $\alpha_i = 0$ (i.e., there are no support vectors), then $\boldsymbol{w} = \boldsymbol{0}$ and the result follows from (A.1g). □

B.2. Uniqueness of the dual solution

Proposition 3. *If $(\boldsymbol{\alpha}_1, \tilde{\boldsymbol{\alpha}}_1)$ and $(\boldsymbol{\alpha}_2, \tilde{\boldsymbol{\alpha}}_2)$ are solutions to the optimization problem (3), then*

$$(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) \in \mathcal{N}(\mathbf{YKY}) \cap \mathbf{1}^\perp \cap \boldsymbol{y}^\perp,$$
$$(\tilde{\boldsymbol{\alpha}}_1 - \tilde{\boldsymbol{\alpha}}_2) \in \mathcal{N}(\tilde{\mathbf{K}}) \cap \mathbf{1}^\perp.$$

If $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are solutions to the problem (4), then

$$(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) \in \mathcal{N}(\mathbf{YKY}) \cap \mathbf{1}^\perp \cap \boldsymbol{y}^\perp.$$

Proof. The proof employs the same method as in the proof of Theorem 2 and we only provide the part concerning the WSVM problem.

Let $\mathbf{K}' = \mathbf{YKY}$ and consider a family of solutions $\boldsymbol{\alpha}_t = (1-t)\boldsymbol{\alpha}_1 + t\boldsymbol{\alpha}_2,\ t \in [0, 1]$. Note that $(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) \in \boldsymbol{y}^\perp$ follows directly from the optimization constraints. Expanding $F(\boldsymbol{\alpha}_t) - F(\boldsymbol{\alpha}_1) = 0$ and differentiating w.r.t. t yields:

$$(t-1)\boldsymbol{\alpha}_1^\top\mathbf{K}'\boldsymbol{\alpha}_1 + (1-2t)\boldsymbol{\alpha}_1^\top\mathbf{K}'\boldsymbol{\alpha}_2 + t\boldsymbol{\alpha}_2^\top\mathbf{K}'\boldsymbol{\alpha}_2$$
$$+ \mathbf{1}^\top(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) = 0,$$
$$(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)^\top\mathbf{K}'(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) = 0.$$

It follows that $(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) \in \mathcal{N}(\mathbf{K}')$. Let $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 + \boldsymbol{v}, \boldsymbol{v} \in \mathcal{N}(\mathbf{K}')$, then from the first equation $\mathbf{1}^\top\boldsymbol{v} = 0$, which completes the proof. □

Corollary 3. *If $\mathbf{K}$ has full rank, then solution to the problem (4) is unique. If $\mathbf{K}$ and $\tilde{\mathbf{K}}$ have full rank, then solution to the problem (3) is unique.*

B.3. Proof of Proposition 1

Proposition. *Let $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the WSVM problem (2). The point $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star)$ is primal optimal for any weight vector $\boldsymbol{c} \in \mathcal{W}$, and all such weights are contained in $\mathcal{W}$.*

Proof. The proof consists in a straightforward application of the KKT conditions. The additional constraint that $\sum_i \mu_i = \sum_i \alpha_i^\star$ follows from Proposition 3 since it must hold that $(\boldsymbol{\mu} - \boldsymbol{\alpha}^\star) \in \mathbf{1}^\perp$. □

B.4. Proof of Theorem 3

Theorem. *Let $(\boldsymbol{w}^\star, b^\star, \tilde{\boldsymbol{w}}^\star, \tilde{b}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the SVM+ problem. There exists a choice of weights $\boldsymbol{c}$, namely $\boldsymbol{c} = \boldsymbol{\alpha}^\star + \boldsymbol{\beta}^\star$, and $\boldsymbol{\xi}^\star$ such that $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ is a primal–dual optimal point for the WSVM problem.*

Proof. Given any fixed feasible $\tilde{\boldsymbol{\alpha}}$, the SVM+ problem (3) is equivalent to the WSVM problem (4) with $\boldsymbol{c} = C\mathbf{1} + \tilde{\boldsymbol{\alpha}}$. In particular, if $(\boldsymbol{\alpha}^\star, \tilde{\boldsymbol{\alpha}}^\star)$ is a solution to (3), then $\boldsymbol{\alpha}^\star$ is a solution to (4) with $\boldsymbol{c} = C\mathbf{1} + \tilde{\boldsymbol{\alpha}}^\star = \boldsymbol{\alpha}^\star + \boldsymbol{\beta}^\star$. Let $\xi_i^\star = \langle\tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{z}}_i\rangle + \tilde{b}^\star$, then the point $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ verifies the KKT conditions for the WSVM problem (2). □

B.5. Proof of Lemma 1

Lemma. *Assume any given $C > 0,\ \gamma \geq 0$ and let $(\boldsymbol{w}^\star, b^\star, \tilde{\boldsymbol{w}}^\star, \tilde{b}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the SVM+ problem (1), then the following holds:*

$$\frac{\sum_{i=1}^{n}(\alpha_i^\star + \beta_i^\star)h_i}{\sum_{i=1}^{n}(\alpha_i^\star + \beta_i^\star)} \geq \frac{1}{n}\sum_{i=1}^{n}h_i, \tag{B.1}$$

where $h_i := [1 - y_i(\langle\boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + b^\star)]_+$ is the hinge loss at a point $i = 1, \ldots, n$. If $\gamma = 0$, then (B.1) is satisfied with equality.

Proof. It follows from the KKT conditions that

$$\langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{z}}_i \rangle + \tilde{b}^\star = h_i + \delta_i, \quad \delta_i \geq 0,$$
$$\alpha_i^\star > 0 \vee \beta_i^\star > 0 \Rightarrow \delta_i = 0, \quad (\alpha_i^\star + \beta_i^\star)\delta_i = 0.$$

Multiplying by $(\alpha_i^\star + \beta_i^\star - C)$ and summing up yields

$$\gamma \langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{w}}^\star \rangle = \sum_{i=1}^n (\alpha_i^\star + \beta_i^\star) h_i - C \sum_{i=1}^n (h_i + \delta_i).$$

Note that $C = \frac{1}{n} \sum_{i=1}^n (\alpha_i^\star + \beta_i^\star) > 0$, hence

$$\gamma \langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{w}}^\star \rangle = \sum_{i=1}^n (\alpha_i^\star + \beta_i^\star) h_i - \frac{1}{n} \sum_{i=1}^n (\alpha_i^\star + \beta_i^\star) \sum_{i=1}^n (h_i + \delta_i).$$

Since $\gamma \langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{w}}^\star \rangle \geq 0$, it must hold that

$$\sum_{i=1}^n (\alpha_i^\star + \beta_i^\star) h_i \geq \frac{1}{n} \sum_{i=1}^n (\alpha_i^\star + \beta_i^\star) \sum_{i=1}^n (h_i + \delta_i)$$
$$\geq \frac{1}{n} \sum_{i=1}^n (\alpha_i^\star + \beta_i^\star) \sum_{i=1}^n h_i. \qquad (B.2)$$

Division by $\sum_{i=1}^n (\alpha_i^\star + \beta_i^\star)$ completes the proof. □

B.6. SVM+ reduction to standard SVM

We show that when there is an equality in the previous lemma, then the SVM+ can be reduced to the standard SVM. For simplicity, we only state this result in the linear setting, where $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^d$.

Proposition 4. *Assume the setting of Lemma 1 and let (B.1) be satisfied with equality, then*

$$\langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{x}}_i \rangle + \tilde{b}^\star = h_i, \quad i = 1, \ldots, n.$$

Furthermore, the following holds.

1. *If $\gamma > 0$, then $\tilde{\boldsymbol{w}}^\star = \boldsymbol{0}$ and $\tilde{b}^\star = h_i$ for all i, i.e., the loss on all data points is the same and the hard margin SVM is a special case with $\tilde{b}^\star = 0$.*
2. *If $\gamma = 0$, then $\tilde{\mathbf{X}}\tilde{\boldsymbol{\alpha}}^\star = \boldsymbol{0}$, where $\tilde{\mathbf{X}}$ is the matrix of $\tilde{\boldsymbol{x}}_i$ stacked. If additionally $\operatorname{rank}(\tilde{\mathbf{X}}) = n$, then $\alpha_i^\star + \beta_i^\star = C$ for all i and any vector in $\mathbb{R}^n$ can be represented via $\langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{x}}_i \rangle + \tilde{b}^\star$, hence the soft margin SVM is recovered with $\xi_i^\star = \langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{x}}_i \rangle + \tilde{b}^\star$.*

Proof. It follows from (B.2) that $\delta_i = 0$ and $\langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{x}}_i \rangle + \tilde{b}^\star = h_i$ for $i = 1, \ldots, n$. If $\gamma > 0$, then $\gamma \langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{w}}^\star \rangle = 0$ implies $\tilde{\boldsymbol{w}}^\star = \boldsymbol{0}$ and thus $\tilde{b}^\star = h_i$ for all i.

If $\gamma = 0$, then (A.1c) implies $\tilde{\mathbf{X}}\tilde{\boldsymbol{\alpha}}^\star = \boldsymbol{0}$, where $\tilde{\boldsymbol{\alpha}}^\star = \boldsymbol{\alpha}^\star + \boldsymbol{\beta}^\star - C\mathbf{1}$, as before. If $\operatorname{rank}(\tilde{\mathbf{X}}) = n$, then $\tilde{\mathbf{X}}\tilde{\boldsymbol{\alpha}}^\star = \boldsymbol{0}$ yields $\tilde{\boldsymbol{\alpha}}^\star = \boldsymbol{0}$, and so $\alpha_i^\star + \beta_i^\star = C$ for $i = 1, \ldots, n$. Since $(\tilde{\boldsymbol{x}}_i)_{i=1}^n$ is in this case a basis in $\mathbb{R}^n$ and there is no penalty on $\left\|\tilde{\boldsymbol{w}}\right\|^2$ in the objective function, the SVM+ does not impose any additional constraints compared to the soft margin SVM. The primal–dual optimal point of the SVM+ is thus also optimal for the SVM with $\xi_i^\star = \langle \tilde{\boldsymbol{w}}^\star, \tilde{\boldsymbol{x}}_i \rangle + \tilde{b}^\star$. □

B.7. Proof of Theorem 4

Theorem. *Let $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}_0^\star, \boldsymbol{\beta}_0^\star)$ be a primal–dual optimal point for the WSVM problem with instance weights $\mathbf{c}_0 \in \mathbb{R}_+^n$, not all zero. There exists a choice of C, γ, and correcting features $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^n$ such that $(\boldsymbol{w}^\star, b^\star)$ is optimal for the SVM+problem iff:*

$$\exists \mathbf{c} \in \mathcal{W} : \rho(\mathbf{c}, \boldsymbol{\xi}^\star) := \langle \mathbf{c} - \bar{c}\mathbf{1}, \boldsymbol{\xi}^\star \rangle \geq 0, \qquad (B.3)$$

where $\bar{c} := n^{-1} \sum_{i=1}^n c_i$. If $\rho(\mathbf{c}, \boldsymbol{\xi}^\star) \geq 0$, one such possible choice is as follows:

$$C = \bar{c}, \qquad \gamma = \rho(\mathbf{c}, \boldsymbol{\xi}^\star), \qquad \tilde{x}_i = \xi_i^\star - \tilde{b}^\star, \quad \forall i \qquad (B.4)$$

moreover, the optimal $\tilde{w}^\star$ and $\tilde{b}^\star$ in that case are:

$$\tilde{w}^\star = 1, \qquad \tilde{b}^\star = \langle \mathbf{c}, \boldsymbol{\xi}^\star \rangle / \langle \mathbf{c}, \mathbf{1} \rangle. \qquad (B.5)$$

Proof. (B.3) *is necessary*. Assume there exists an SVM+ setting such that $(\boldsymbol{w}^\star, b^\star, \tilde{\boldsymbol{w}}^\star, \tilde{b}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ is a primal–dual optimal point for the SVM+ problem (1) and let $\mathbf{c} = \boldsymbol{\alpha}^\star + \boldsymbol{\beta}^\star$ (note that $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ and $(\boldsymbol{\alpha}_0^\star, \boldsymbol{\beta}_0^\star)$ may be different). Theorem 3 states that there exists $\boldsymbol{\xi}_0^\star$ such that $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}_0^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ is primal–dual optimal for the WSVM problem with weights $\mathbf{c}$. We need to show that $\boldsymbol{\xi}_0^\star = \boldsymbol{\xi}^\star$. This follows directly from the KKT conditions when all $c_{0,i} > 0$ and $c_i > 0$ since $h_i = [1 - y_i(\langle \boldsymbol{w}^\star, \boldsymbol{x}_i \rangle + b^\star)]_+$ are the same for both problems. If some of the weights are zero, then the corresponding $\xi_i^\star$ is not uniquely defined (it is unbounded from above) and we have to assume that the algorithm returns the value at the lower bound, i.e., $\xi_i^\star = h_i$. Now, given that $\boldsymbol{\xi}_0^\star = \boldsymbol{\xi}^\star$, $\mathbf{c} \in \mathcal{W}$ by Proposition 1 and $\rho(\mathbf{c}, \boldsymbol{\xi}^\star) \geq 0$ by Corollary 2.

(B.3) *is sufficient*. First, consider the case $\rho(\mathbf{c}, \boldsymbol{\xi}^\star) > 0$ and let $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point of the WSVM problem with weights $\mathbf{c}$. We construct $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^n$ and provide $C > 0$, $\gamma > 0$, $\tilde{\boldsymbol{w}}^\star$, and $\tilde{b}^\star$ such that $(\boldsymbol{w}^\star, b^\star, \tilde{\boldsymbol{w}}^\star, \tilde{b}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ is primal–dual optimal for the corresponding SVM+ problem.

It is sufficient to look for one dimensional correcting features that additionally satisfy $\sum_{i=1}^n c_i \tilde{x}_i = 0$. The KKT conditions in this case imply that

$$\tilde{w}^\star = -\frac{C}{\gamma} \sum_{i=1}^n \tilde{x}_i, \qquad C = \frac{1}{n} \sum_{i=1}^n c_i = \bar{c}. \qquad (B.6)$$

We require for all $i = 1, \ldots, n$ that

$$\tilde{w}^\star \tilde{x}_i + \tilde{b}^\star = [1 - y_i(\langle \boldsymbol{w}^\star, \boldsymbol{x}_i \rangle + b^\star)]_+ = \xi_i^\star. \qquad (B.7)$$

Multiplying both sides by c_i and summing up yields

$$\tilde{b}^\star = \langle \mathbf{c}, \boldsymbol{\xi}^\star \rangle / \langle \mathbf{c}, \mathbf{1} \rangle.$$

Plugging (B.6) into (B.7) and solving for $\tilde{x}_i$ one gets:

$$\tilde{x}_i = \pm \sqrt{\frac{\gamma}{\rho(\mathbf{c}, \boldsymbol{\xi}^\star)}} (\tilde{b}^\star - \xi_i^\star). \qquad (B.8)$$

Choosing $\gamma = \rho(\mathbf{c}, \boldsymbol{\xi}^\star)$ and the plus sign in (B.8) for convenience, (B.6) leads to $\tilde{w}^\star = \rho(\mathbf{c}, \boldsymbol{\xi}^\star)/\gamma = 1$.

Now, consider $\rho(\mathbf{c}, \boldsymbol{\xi}^\star) = 0$. Let $\tilde{\mathbf{X}} = [\tilde{\boldsymbol{x}}_1 \cdots \tilde{\boldsymbol{x}}_n]$ and set $\gamma = 0$. Proposition 4 and the KKT conditions imply:

$$C = \bar{c}, \qquad \tilde{\mathbf{X}}(\mathbf{c} - \bar{c}\mathbf{1}) = \boldsymbol{0}, \qquad \tilde{\mathbf{X}}^\top \tilde{\boldsymbol{w}}^\star + \tilde{b}^\star \mathbf{1} = \boldsymbol{\xi}^\star.$$

Hence, the matrix $\tilde{\mathbf{X}}$ must satisfy

$$(\mathbf{c} - \bar{c}\mathbf{1}) \in \mathcal{N}(\tilde{\mathbf{X}}), \qquad (\boldsymbol{\xi}^\star - \tilde{b}^\star \mathbf{1}) \in \mathcal{R}(\tilde{\mathbf{X}}^\top).$$

The above requirements translate to

$$\langle \mathbf{c} - \bar{c}, \boldsymbol{\xi}^\star - \tilde{b}^\star \mathbf{1} \rangle = 0,$$

which holds for (B.4), (B.5), and $\rho(\mathbf{c}, \boldsymbol{\xi}^\star) = 0$. □

B.8. Proof of Proposition 2

Proposition. *Let $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\xi}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ be a primal–dual optimal point for the WSVM problem with instance weights $\mathbf{c} \in \mathbb{R}_+^n$, not all zero. If*

$$\mathcal{N}(\mathbf{YKY}) \cap \mathbf{1}^\perp \cap \boldsymbol{y}^\perp = \{\boldsymbol{0}\},$$

then there exists a choice of C, γ, *and* $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^n$ *such that* $(\boldsymbol{w}^\star, b^\star)$ *is optimal for the SVM+ problem iff:*

$$\rho(\boldsymbol{\alpha}^\star, \boldsymbol{\xi}^\star) = \boldsymbol{\xi}^{\star\top}\big(\mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big)\boldsymbol{\alpha}^\star \geq 0.$$

Proof. Sufficiency follows directly from Theorem 4 since $\boldsymbol{c} = \boldsymbol{\alpha}^\star$ is a valid choice of weights (cf. Definition 1). For necessity, note that $\boldsymbol{\alpha}^\star$ is unique by Proposition 3 and all weights in $\mathcal{W}$ are of the form $\boldsymbol{c} = \boldsymbol{\alpha}^\star + \boldsymbol{\beta}$, $\boldsymbol{\beta} \in \mathcal{V}$. The maximum in (6) corresponds to

$$\begin{aligned} \max_{\boldsymbol{\beta}} \quad & \sum_{i=1}^n \xi_i^\star \beta_i - \frac{1}{n}\sum_{i=1}^n \xi_i^\star \sum_{i=1}^n \beta_i, \\ \text{s.t.} \quad & \beta_i \geq 0, \end{aligned}$$

which is attained at $\boldsymbol{\beta} = \mathbf{0}$ since $\forall i\ \xi_i^\star \beta_i = 0$. □

B.9. Proof of Theorem 5

Theorem. *Let the loss function* ℓ *be convex and twice continuously differentiable and let the kernel matrix* $\mathbf{K}$ *be (strictly) positive definite. Define vectors* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *componentwise for* $i = 1, \dots, n$ *as*

$$\begin{aligned} u_i &:= y_i \ell'(y_i[K_i^\top \boldsymbol{\alpha}^\star + b^\star]), \\ v_i &:= c_i \ell''(y_i[K_i^\top \boldsymbol{\alpha}^\star + b^\star]), \end{aligned}$$

where $(\boldsymbol{\alpha}^\star, b^\star)$ *is a solution of* (18) *for a given* $\boldsymbol{c}$. *If* $\boldsymbol{v} \neq \mathbf{0}$, *then the solution is unique,* $\boldsymbol{\alpha}^\star$ *and* $b^\star$ *are continuously differentiable w.r.t.* $\boldsymbol{c}$ *and the corresponding gradient can be computed as follows.*

$$\begin{bmatrix} \dfrac{\partial \boldsymbol{\alpha}^\star}{\partial \boldsymbol{c}} \\ \dfrac{\partial b^\star}{\partial \boldsymbol{c}} \end{bmatrix} = -\begin{bmatrix} \mathbf{I} + \operatorname{diag}(\boldsymbol{v})\mathbf{K} & \boldsymbol{v} \\ \mathbf{1}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \operatorname{diag}(\boldsymbol{u}) \\ \mathbf{0}^\top \end{bmatrix}. \tag{B.9}$$

Proof. Uniqueness of solution follows from a similar argument as in the proof of Theorem 2 and is obvious for $\boldsymbol{\alpha}$. Let $b_1^\star$ and $b_2^\star$ be two optimal b and define $b_t^\star = (1-t)b_1^\star + tb_2^\star$. Considering the difference of the objective function at $b_t^\star$ and $b_1^\star$ and differentiating twice w.r.t. t, one arrives at

$$(b_2^\star - b_1^\star)\mathbf{1}^\top \boldsymbol{v} = 0 \Rightarrow b_2^\star = b_1^\star.$$

The optimality conditions of (18) yield

$$\mathbf{K}(\boldsymbol{\alpha}^\star + \operatorname{diag}(\boldsymbol{u})\boldsymbol{c}) = \mathbf{0}, \qquad \langle \boldsymbol{u}, \boldsymbol{c}\rangle = 0.$$

Since $\mathbf{K}$ is non-singular it can be dropped from the first equation. Computation of the total derivatives yields the linear system below.

$$\begin{bmatrix} \mathbf{I} + \operatorname{diag}(\boldsymbol{v})\mathbf{K} & \boldsymbol{v} \\ \boldsymbol{v}^\top \mathbf{K} & \mathbf{1}^\top \boldsymbol{v} \end{bmatrix} \begin{bmatrix} \dfrac{\partial \boldsymbol{\alpha}^\star}{\partial \boldsymbol{c}} \\ \dfrac{\partial b^\star}{\partial \boldsymbol{c}} \end{bmatrix} = -\begin{bmatrix} \operatorname{diag}(\boldsymbol{u}) \\ \boldsymbol{u}^\top \end{bmatrix}. \tag{B.10}$$

Note that (B.10) is equivalent to the system in (B.9) since the last equation can be equivalently replaced by the sum of the first n equations minus the last one. To apply the implicit function theorem, it remains to show that the matrix in (B.9) is invertible. Recall that the determinant of a block matrix factors as the determinant of a block and its Schur complement. It is thus sufficient to show that

$$\det(\mathbf{I} + \operatorname{diag}(\boldsymbol{v})\mathbf{K}) \neq 0, \qquad \mathbf{1}^\top(\mathbf{I} + \operatorname{diag}(\boldsymbol{v})\mathbf{K})^{-1}\boldsymbol{v} \neq 0.$$

Assume w.l.o.g. that the first $m \leq n$ components of $\boldsymbol{v}$ are non-zero and define $M := \mathbf{I} + \operatorname{diag}(\boldsymbol{v})\mathbf{K}$. Further,

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} \mathbf{I}_m + \operatorname{diag}(\boldsymbol{v}_m)\mathbf{K}_m & B \\ \mathbf{0}_{n-m,m} & \mathbf{I}_{n-m} \end{bmatrix},$$

where the B block is irrelevant. It now follows that

$$\begin{aligned} \det(M) &= \det(D)\det(A - BD^{-1}C) = \det(A) \\ &= \det(\operatorname{diag}(\boldsymbol{v}_m))\det(\operatorname{diag}(\boldsymbol{v}_m)^{-1} + \mathbf{K}_m) \neq 0, \end{aligned}$$

where we use that $\operatorname{diag}(\boldsymbol{v}_m)^{-1}$ is positive definite since $\boldsymbol{v}_m \succ \mathbf{0}_m$ due to convexity of ℓ. Finally,

$$\begin{aligned} \mathbf{1}^\top M^{-1}\boldsymbol{v} &= \mathbf{1}^\top \begin{bmatrix} A^{-1} & -A^{-1}BD^{-1} \\ \mathbf{0}_{n-m,m} & D^{-1} \end{bmatrix}^{-1} \boldsymbol{v} \\ &= \mathbf{1}_m^\top \left(\operatorname{diag}(\boldsymbol{v}_m)^{-1} + \mathbf{K}_m\right)^{-1} \mathbf{1}_m > 0. \quad \square \end{aligned}$$

References

Bennett, K. P., & Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods & Software*, *1*(1), 23–34.

Beygelzimer, A., Dasgupta, S., & Langford, J. (2009). Importance weighted active learning. In *ICML* (pp. 49–56).

Borwein, J., & Lewis, A. (2000). *Convex analysis and nonlinear optimization: theory and examples*. Springer-Verlag.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Burges, C.J.C., & Crisp, D.J. (1999). Uniqueness of the SVM solution. In *NIPS* (pp. 223–229).

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27. Software available at: http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Computation*, *19*(5), 1155–1178.

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.) (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, *46*(1–3), 131–159.

Chen, J., Liu, X., & Lyu, S. (2012). Boosting with side information. In *ACCV* (pp. 5–9).

Cortes, C., Mansour, Y., & Mohri, M. (2010). Learning bounds for importance weighting. In *NIPS* (pp. 442–450).

Dredze, M., Crammer, K., & Pereira, F. (2008). Confidence-weighted linear classification. In *ICML* (pp. 264–271).

Elkan, C. (2001). The foundations of cost-sensitive learning. In *IJCAI* (pp. 973–978).

Elter, M., Schulz-Wendtland, R., & Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical Physics*, *34*(11), 4164–4172.

Feyereisl, J., & Aickelin, U. (2012). Privileged information for data clustering. *Information Sciences*, *194*, 4–23.

Fouad, S., Tino, P., Raychaudhury, S., & Schneider, P. (2012). Learning using privileged information in prototype based models. In *ICANN* (pp. 322–329).

Frank, A., & Asuncion, A. (2010). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 153–161.

Lauer, F., & Bloch, G. (2008). Incorporating prior knowledge in support vector machines for classification: a review. *Neurocomputing*, *71*(7–9), 1578–1594.

Liang, L., Cai, F., & Cherkassky, V. (2009). Predictive learning with structured (grouped) data. *Neural Networks*, *22*(5–6), 766–773.

Liang, L., & Cherkassky, V. (2008). Connection between SVM+ and multi-task learning. In *IJCNN* (pp. 2048–2054).

Lin, C.-F., & Wang, S.-D. (2002). Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 464–471.

Margineantu, D.D. (2002). Class probability estimation and cost-sensitive classification decisions. In *ECML* (pp. 270–281).

Nocedal, J., & Wright, S. (2006). *Numerical optimization* (2nd ed.). Springer.

Pechyony, D., & Vapnik, V. (2010). On the theory of learnining with privileged information. In *NIPS* (pp. 1894–1902).

Pechyony, D., & Vapnik, V. (2011). Fast optimization algorithms for solving SVM+. In *Statistical learning and data science*. Chapman & Hall.

Schölkopf, B., Herbrich, R., & Smola, A. (2001). A generalized representer theorem. In *COLT* (pp. 416–426).

Schölkopf, B., Simard, P.Y., Smola, A.J., & Vapnik, V.N. (1998). Prior knowledge in support vector kernels. In *NIPS* (pp. 640–646).

Schölkopf, B., & Smola, A. (2002). *Learning with kernels: support vector machines, regularization, optimization and beyond*. University Press Group Limited.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244.

Sugiyama, M., & Müller, K.-R. (2005). Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, *23*(4), 249–279.

Vapnik, V. (2006). *Empirical inference science afterword of 2006*. Springer.

Vapnik, V., & Vashist, A. (2009). A new learning paradigm: learning using privileged information. *Neural Networks*, *22*(5–6), 544–557.

Vapnik, V., Vashist, A., & Pavlovitch, N. (2009). Learning using hidden information (Learning with teacher). In *IJCNN* (pp. 3188–3195).

Wolf, L., & Levy, N. (2013). The SVM-minus similarity score for video face recognition. In *CVPR* (pp. 3523–3530).

Wu, X., & Srihari, R. (2004). Incorporating prior knowledge with weighted margin support vector machines. In *SIGKDD* (pp. 326–333).

Yang, X., Song, Q., & Cao, A. (2005). Weighted support vector machine for data classification. In *IJCNN* (pp. 859–864).

Zadrozny, B., Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *ICDM* (pp. 435–448).