MAXIMIZATION OF MINIMUM MARGIN FOR STATISTICAL PARAMETER ESTIMATION

YOICHI IZUNAGA AND YOSHITSUGU YAMAMOTO

Department of Social Systems and Management Discussion Paper Series No.1317

Abstract. This short note was prompted by a presentation at a regular seminar of *team PhilOpt))))* graduate students in December 2013. The problem originates in a linear discriminant problem, for which the maximum likelihood estimation ends up with a complicated likelihood function. In this note we propose to apply the support vector machine to the problem and provide some basic results.

1. Introduction

This paper is concerned with a multi-class classification problem of n objects, each of which is endowed with an m-dimensional attribute vector $\boldsymbol{x}^i = (x^i_1, x^i_2, \ldots, x^i_m)^\top \in \mathbb{R}^m$ and a label ℓ_i. The underlying statistical model assumes that object i receives label k, i.e., $\ell_i = k$, when the latent variable y_i determined by

$$y_i = \boldsymbol{w}^\top \boldsymbol{x}^i + \epsilon_i = \sum_{j=1}^{m} w_j x^i_j + \epsilon_i$$

falls between two thresholds p_k and p_{k+1}, where ϵ_i represents a random noise whose probabilistic property is not known. Namely, attribute vectors of objects are loosely separated by hyperplanes $H(\boldsymbol{w}, p_k) = \{\, \boldsymbol{x} \in \mathbb{R}^m \mid (\boldsymbol{w})^\top \boldsymbol{x} = p_k \,\}$ for $k = 1, 2, \ldots, l$ which share a common normal vector $\boldsymbol{w}$, then each object is given a label according to the layer it is located in. Note that neither y_i's, w_j's nor p_k's are observable. Our problem is to find the vector $\boldsymbol{w} \in \mathbb{R}^m$ as well as the thresholds $p_1, p_2, \ldots, p_l$ that best fit the input data $\{\, (\boldsymbol{x}^i, \ell_i) \mid i \in N \,\}$.

This problem is a variation of the multi-class classification problem, for which several learning algorithms of the support vector machine (*SVM* for short) have been proposed such as one-versus-the-rest approach, one-versus-one approach, decision tree approach, and the all-together approach. We refer the reader to Chapters 4.1.2 and 7.1.3 of Bishop [1], Chapter 10.10 of Vapnik [6] and Tatsumi *et al.* [5] and references therein. What distinguishes the problem from other multi-class classification problems is that the identical normal vector should be shared by all the separating hyperplanes.

2. Definitions and notation

Throughout the paper $N = \{1, 2, \ldots, i, \ldots, n\}$ denotes the set of n objects and $\boldsymbol{x}^i = (x^i_1, x^i_2, \ldots, x^i_m)^\top \in \mathbb{R}^m$ denotes the attribute vector of object i. The predetermined set of labels is $L = \{0, 1, \ldots, k, \ldots, l\}$ and the label assigned to object i is denoted by ℓ_i. Let $N(k) = \{\, i \in N \mid \ell_i = k \,\}$ be the set of objects with label $k \in L$, and for notational convenience we write $n(k) = |N(k)|$ for $k \in L$, and $N(k..k') = N(k) \cup N(k+1) \cup \cdots \cup N(k')$

Date: March 5, 2014.

Key words and phrases. Support vector machine, Multi-class classification, Hard margin, Soft margin, Quadratic program, Kernel technique.

This research is partly supported by The Okawa Foundation for Information and Telecommunications.

for $k, k' \in L$ such that $k < k'$. We can assume that $i < j$ holds when $\ell_i < \ell_j$ for $i, j \in N$, by rearranging the objects if necessary, hence $N(0) = \{1, 2, \ldots, n(0)\}$, $N(1) = \{n(0)+1, \ldots, n(0)+n(1)\}$, and so forth. For succinct notation we define

$$X = \begin{bmatrix} \cdots & \boldsymbol{x}^i & \cdots \end{bmatrix}_{i \in N} \in \mathbb{R}^{m \times n} \tag{2.1}$$

the matrix of columns $\boldsymbol{x}^i$, and

$$K = X^\top X = \begin{bmatrix} & (\boldsymbol{x}^i)^\top \boldsymbol{x}^j & \\ & & \end{bmatrix}_{i,j \in N} \in \mathbb{R}^{n \times n}, \tag{2.2}$$

and denote the k-dimensional zero vector by $\mathbf{0}_k$ and the k-dimensional vector of 1's by $\mathbf{1}_k$.

3. Maximization of Minimum Margin for Separable Case

Henceforth we assume that $N(k) \neq \emptyset$ for all $k \in L$ for the sake of simplicity, and adopt the notational convention that $p_0 = -\infty$ and $p_{l+1} = +\infty$. We say that an instance $\{ (\boldsymbol{x}^i, \ell_i) \mid i \in N \}$ is *separable* if there exist $\boldsymbol{w} \in \mathbb{R}^m$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_l)^\top \in \mathbb{R}^l$ such that

$$p_k < \boldsymbol{w}^\top \boldsymbol{x}^i < p_{k+1} \quad \text{for } i \in N(k) \text{ and } k \in L.$$

Clearly an instance is separable if and only if there are $\boldsymbol{w}$ and $\boldsymbol{p}$ such that

$$p_k + 1 \leq \boldsymbol{w}^\top \boldsymbol{x}^i \leq p_{k+1} - 1 \quad \text{for } i \in N(k) \text{ and } k \in L.$$

For each $k \in L \setminus \{0\}$ we see that

$$\max_{i \in N(k-1)} (\boldsymbol{w})^\top \boldsymbol{x}^i \leq p_k - 1 < p_k < p_k + 1 \leq \min_{j \in N(k)} (\boldsymbol{w})^\top \boldsymbol{x}^j,$$

hence

$$\min_{j \in N(k)} \frac{\boldsymbol{w}^\top}{\|\boldsymbol{w}\|} \boldsymbol{x}^j - \max_{i \in N(k-1)} \frac{\boldsymbol{w}^\top}{\|\boldsymbol{w}\|} \boldsymbol{x}^i \geq \frac{2}{\|\boldsymbol{w}\|}.$$

Then the margin between $\{ \boldsymbol{x}^i \mid i \in N(k-1) \}$ and $\{ \boldsymbol{x}^j \mid j \in N(k) \}$ is at least $2/\|\boldsymbol{w}\|$. Hence the maximization of the minimum margin is formulated as the quadratic programming

$$\left| \begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} & p_k + 1 \leq (\boldsymbol{x}^i)^\top \boldsymbol{w} \leq p_{k+1} - 1 \qquad \text{for } i \in N(k) \text{ and for } k \in L \end{array} \right.$$

more explicitly with the notation introduced in Section 2

$$(H) \quad \left| \begin{array}{lll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 & \\ \text{subject to} & 1 - (\boldsymbol{x}^i)^\top \boldsymbol{w} + p_{\ell_i} \leq 0 & \text{for } i \in N(1..l) \\ & 1 + (\boldsymbol{x}^i)^\top \boldsymbol{w} - p_{\ell_i + 1} \leq 0 & \text{for } i \in N(0..l-1). \end{array} \right.$$

The constraints therein are called *hard margin* constraints, and we name this problem (H). The leading coefficient $1/2$ of the objective function is for the sake of notational simplicity in further discussion.

Among the learning algorithms for multi-class SVM are one-versus-one approach, one-versus-the-rest approach and all-together approach by Weston and Watkins [7]. See, for example, Bishop [1] and Vapnik [6]. When the one-versus-the-rest approach is applied to our problem, one would solve

$$\left|\begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}^k\|^2 \\ \text{subject to} & (\boldsymbol{x}^i)^\top \boldsymbol{w}^k - p_k - 1 \geq (\boldsymbol{x}^j)^\top \boldsymbol{w}^k - p_k + 1 \quad \begin{array}{l}\text{for } i \in N(k) \\ \text{and } j \in N \setminus N(k)\end{array} \end{array}\right.$$

for finding the normal vector $\boldsymbol{w}^k$ of a hyperplane that hopefully separates the objects of class k from the rest. Besides the widely recognized imbalance between the size of $N(k)$ and the rest $N \setminus N(k)$ (see, for example, Fung and Mangasarian [3]), this problem would almost certainly suffer infeasibility when a middle class k is considered, i.e., $1 \leq k \leq l-1$. In the all-together approach one would solve

$$\left|\begin{array}{ll} \text{minimize} & \dfrac{1}{2}\displaystyle\sum_{k \in L} \|\boldsymbol{w}^k\|^2 \\ \text{subject to} & (\boldsymbol{x}^i)^\top \boldsymbol{w}^k - p_k - 1 \geq (\boldsymbol{x}^i)^\top \boldsymbol{w}^{k'} - p_{k'} + 1 \quad \begin{array}{l}\text{for } k' \in L \setminus \{k\},\ i \in N(k) \\ \text{and } k \in L,\end{array} \end{array}\right.$$

and assign the label

$$\operatorname{argmax}\{\, (\boldsymbol{w}^k)^\top \bar{\boldsymbol{x}} - p_k \mid k \in L \,\}$$

to an object with an attribute vector $\bar{\boldsymbol{x}}$. If one requires all $\boldsymbol{w}^k$ be identical, this formulation would become meaningless because of the cancellation of $(\boldsymbol{x}^i)^\top \boldsymbol{w}^k$ and $(\boldsymbol{x}^i)^\top \boldsymbol{w}^{k'}$.

The one-versus-one approach would solve

$$\left|\begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}^{(k,k')}\|^2 \\ \text{subject to} & (\boldsymbol{x}^i)^\top \boldsymbol{w}^{(k,k')} - p_{(k,k')} - 1 \geq (\boldsymbol{x}^j)^\top \boldsymbol{w}^{(k,k')} - p_{(k,k')} + 1 \quad \begin{array}{l}\text{for } i \in N(k) \\ \text{and } j \in N(k')\end{array} \end{array}\right.$$

for every possible pair (k, k') of classes. This approach would require significantly heavy computation burden when there are many classes, and many of the above problems would be rather trivial and provide no useful information. Our formulation (H) could be obtained by adding the constraint that $\boldsymbol{w}^{(k,k')}$ be identical and deleting constraints for a pair of non-adjacent classes.

4. Dual of Hard Margin Problem

The Lagrangian function for the hard margin problem (H) introduced in the previous section is

$$L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i \in N(1..l)} \alpha_i (1 - (\boldsymbol{x}^i)^\top \boldsymbol{w} + p_{\ell_i}) + \sum_{i \in N(0..l-1)} \beta_i (1 + (\boldsymbol{x}^i)^\top \boldsymbol{w} - p_{\ell_i+1}),$$

where α_i and β_i are nonnegative Lagrangian multipliers. Denoting

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_i, \ldots \alpha_n)^\top \in \mathbb{R}^n \quad \text{and} \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_i, \ldots, \beta_n)^\top \in \mathbb{R}^n$$

with the convention that

$$\alpha_i = 0 \quad \text{for all } i \in N(0) \quad \text{and} \quad \beta_i = 0 \quad \text{for all } i \in N(l),$$

the Lagrangian function is compactly written as

$$L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - (X(\boldsymbol{\alpha} - \boldsymbol{\beta}))^\top \boldsymbol{w} + (A\boldsymbol{\alpha} - B\boldsymbol{\beta})^\top \boldsymbol{p} + \boldsymbol{\alpha}^\top \mathbf{1}_n + \boldsymbol{\beta}^\top \mathbf{1}_n, \tag{4.1}$$

where the matrix X is defined in (2.1), and

$$A = \begin{bmatrix} \mathbf{0}_{n(0)}^\top & \mathbf{1}_{n(1)}^\top & & & \\ & & \mathbf{1}_{n(2)}^\top & & \\ & & & \ddots & \\ & & & & \mathbf{1}_{n(l)}^\top \end{bmatrix} \in \mathbb{R}^{l \times n}, \tag{4.2}$$

$$B = \begin{bmatrix} \mathbf{1}_{n(0)}^\top & & & & \\ & \mathbf{1}_{n(1)}^\top & & & \\ & & \ddots & & \\ & & & \mathbf{1}_{n(l-1)}^\top & \mathbf{0}_{n(l)}^\top \end{bmatrix} \in \mathbb{R}^{l \times n}. \tag{4.3}$$

Let

$$\omega(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min\{\, L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \mid (\boldsymbol{w}, \boldsymbol{p}) \in \mathbb{R}^{m+l} \,\},$$

then the *Lagrangian dual of the hard margin problem* is

$$(dH) \quad \left| \begin{array}{ll} \text{maximize} & \omega(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{subject to} & (\boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \mathbf{0}_{2n}. \end{array} \right.$$

Since L is a convex function with respect to $(\boldsymbol{w}, \boldsymbol{p})$, a point $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ attains the minimum $\omega(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for a given $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ if and only if the partial derivatives of L with respect to $(\boldsymbol{w}, \boldsymbol{p})$ vanish at $(\boldsymbol{w}^*, \boldsymbol{p}^*)$, i.e.,

$$\frac{\partial L}{\partial (\boldsymbol{w}, \boldsymbol{p})}(\boldsymbol{w}^*, \boldsymbol{p}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{0}_{n+l}.$$

This condition reduces to

$$\boldsymbol{w}^* - X(\boldsymbol{\alpha} - \boldsymbol{\beta}) = \mathbf{0}_n \tag{4.4}$$

and

$$A\boldsymbol{\alpha} - B\boldsymbol{\beta} = \mathbf{0}_l. \tag{4.5}$$

Plugging these equalities into L, we obtain

$$\omega(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\left\{ \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top K(\boldsymbol{\alpha} - \boldsymbol{\beta}) - \mathbf{1}_n^\top (\boldsymbol{\alpha} + \boldsymbol{\beta}) \right\},$$

where $K = X^\top X$ as defined in (2.2). Note that the variable $\boldsymbol{p}$ disappears due to the equality condition (4.5). Deleting the leading negative coefficient -1, the Lagrangian dual of the hard margin problem is given as

$$(dH) \quad \left| \begin{array}{ll} \text{minimize} & \dfrac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top K(\boldsymbol{\alpha} - \boldsymbol{\beta}) - \mathbf{1}_n^\top (\boldsymbol{\alpha} + \boldsymbol{\beta}) \\ \text{subject to} & A\boldsymbol{\alpha} - B\boldsymbol{\beta} = \mathbf{0}_l \\ & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)} \\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)} \\ & \boldsymbol{\alpha} \geq \mathbf{0}_n \\ & \boldsymbol{\beta} \geq \mathbf{0}_n, \end{array} \right.$$

where $\boldsymbol{\alpha}_{N(0)}$ is a vector of α_i's for $i \in N(0)$ and $\boldsymbol{\beta}_{N(l)}$ is a vector of β_i's for $i \in N(l)$. This is a convex quadratic minimization problem since K is nonnegative definite. Let $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ denote an optimum solution of (dH). Then an optimum normal vector $\boldsymbol{w}^*$ and an optimum threshold vector $\boldsymbol{p}^*$ of the primal problem are given by

$$\boldsymbol{w}^* = X(\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*) \tag{4.6}$$

$$p_k^* = \frac{1}{2}\left(\max_{i \in N(k-1)} (\boldsymbol{w}^*)^\top \boldsymbol{x}^i + \min_{i \in N(k)} (\boldsymbol{w}^*)^\top \boldsymbol{x}^i\right) \quad \text{for } k \in L \setminus \{0\}. \tag{4.7}$$

5. Kernel Technique for Hard Margin Problem

The matrix K of the Lagrangian dual of the hard margin problem (dH) is composed of the inner products $(\boldsymbol{x}^i)^\top \boldsymbol{x}^j$ for $i, j \in N$, which enables us to apply the *kernel technique* simply by replacing them by $\kappa(\boldsymbol{x}^i, \boldsymbol{x}^j)$ for some appropriate kernel function κ.

Let $\phi\colon \mathbb{R}^m \to \mathbb{R}^\nu$ be a function, possibly unknown, such that

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{y})$$

holds for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$, where $\mathbb{R}^\nu$ is some higher dimensional space, so-called *feature space*. In the sequel we denote $\tilde{\boldsymbol{x}} = \phi(\boldsymbol{x})$. The kernel technique considers the vectors $\tilde{\boldsymbol{x}}^i \in \mathbb{R}^\nu$ instead of $\boldsymbol{x}^i \in \mathbb{R}^m$, and finds the ν-dimensional normal vector $\tilde{\boldsymbol{w}} \in \mathbb{R}^\nu$ and thresholds $p_1, \ldots, p_l$. Therefore the matrices X and K should be replaced by $\tilde{X}$ consisting of vectors $\tilde{\boldsymbol{x}}^i$ and $\tilde{K} = \tilde{X}^\top \tilde{X}$, respectively. Note that the latter matrix is given as

$$\tilde{K} = \left[\kappa(\boldsymbol{x}^i, \boldsymbol{x}^j)\right]_{i,j \in N} \in \mathbb{R}^{n \times n}$$

by the kernel function κ. Solving the dual hard margin problem (dH) with K replaced by $\tilde{K}$ to find $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, the optimal normal vector $\tilde{\boldsymbol{w}}^* \in \mathbb{R}^\nu$ of the primal problem would be given as

$$\tilde{\boldsymbol{w}}^* = \tilde{X}(\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*),$$

which is in general not available due to the absence of an explicit representation of $\tilde{X}$. However the value of $(\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}}^j$ necessary to compute the thresholds p_k^* according to (4.7) is obtained as the inner product of $\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*$ and the jth column of $\tilde{K}$. In fact

$$\begin{aligned}(\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}}^j &= (\tilde{X}(\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*))^\top \tilde{\boldsymbol{x}}^j = (\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)^\top \tilde{X}^\top \tilde{\boldsymbol{x}}^j \\ &= (\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)^\top \begin{pmatrix} \vdots \\ (\tilde{\boldsymbol{x}}^i)^\top \tilde{\boldsymbol{x}}^j \\ \vdots \end{pmatrix} = (\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)^\top \begin{pmatrix} \vdots \\ \kappa(\boldsymbol{x}^i, \boldsymbol{x}^j) \\ \vdots \end{pmatrix}. \end{aligned} \tag{5.1}$$

Suppose we are given a new object with attribute vector $\boldsymbol{x} \in \mathbb{R}^n$ to assign a label. In the same way as above we have

$$(\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}} = (\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)^\top \begin{pmatrix} \vdots \\ \kappa(\boldsymbol{x}^i, \boldsymbol{x}) \\ \vdots \end{pmatrix}. \tag{5.2}$$

Then by locating the threshold interval into which this value falls, we can assign a label to the new object.

6. Soft Margin Problem for Non-Separable Case

Similarly to the binary SVM, introducing nonnegative slack variables ξ_{-i} and ξ_{+i} for $i \in N$ relaxes the hard margin constraints to *soft margin* constraints:

$$p_k + 1 - \xi_{-i} \le \boldsymbol{w}^\top \boldsymbol{x}^i \le p_{k+1} - 1 + \xi_{+i} \quad \text{for } i \in N(k) \text{ and for } k \in L.$$

Positive values of variables ξ_{-i} and ξ_{+i} mean misclassification, hence they should be as small as possible. If we penalize positive ξ_{-i} and ξ_{+i} by adding $\sum_{i \in N}(\xi_{-i} + \xi_{+i})$ to the objective function, we have the following *primal soft margin problem* with *1-norm penalty*.

$$(S_1) \quad \left| \begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + c\,\mathbf{1}_n^\top(\boldsymbol{\xi}_- + \boldsymbol{\xi}_+) \\ \text{subject to} & p_k + 1 - \xi_{-i} \le \boldsymbol{w}^\top \boldsymbol{x}^i \le p_{k+1} - 1 + \xi_{+i} \quad \text{for } i \in N(k) \text{ and for } k \in L \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \ge \mathbf{0}_n, \end{array} \right.$$

where $\boldsymbol{\xi}_- = (\xi_{-1}, \ldots, \xi_{-n}), \boldsymbol{\xi}_+ = (\xi_{+1}, \ldots, \xi_{+n})$ and c is a *penalty parameter*. When *2-norm penalty* is employed, we have

$$(S_2) \quad \left| \begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + \dfrac{1}{2}\,c\,(\|\boldsymbol{\xi}_-\|^2 + \|\boldsymbol{\xi}_+\|^2) \\ \text{subject to} & p_k + 1 - \xi_{-i} \le \boldsymbol{w}^\top \boldsymbol{x}^i \le p_{k+1} - 1 + \xi_{+i} \quad \text{for } i \in N(k) \text{ and for } k \in L \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \ge \mathbf{0}_n. \end{array} \right.$$

Lemma 6.1. *The nonnegativity constraints on variables ξ_{-i} and ξ_{+i} of problem (S_2) are redundant.*

Proof. Let $(\boldsymbol{w}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+)$ be a feasible solution of (S_2) with the nonnegativity constraints removed. If $\xi_{-i} < 0$ for some $i \in N$, replacing it with zero will reduce the objective function value. Therefore $\boldsymbol{\xi}_-$ and $\boldsymbol{\xi}_+$ are nonnegative at any optimum solution of (S_2). □

Thus our problem with 2-norm penalty reduces to

$$(S_2) \quad \left| \begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + \dfrac{1}{2}\,c\,(\|\boldsymbol{\xi}_-\|^2 + \|\boldsymbol{\xi}_+\|^2) \\ \text{subject to} & p_k + 1 - \xi_{-i} \le \boldsymbol{w}^\top \boldsymbol{x}^i \le p_{k+1} - 1 + \xi_{+i} \quad \text{for } i \in N(k) \text{ and for } k \in L. \end{array} \right.$$

As proposed in Mangasarian and Musicant [4], the addition of a term $\|\boldsymbol{p}\|^2$ to the objective function yields the following two formulations (S_{12}) and (S_{22}).

$$(S_{12}) \quad \left| \begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + c\,\mathbf{1}_n^\top(\boldsymbol{\xi}_- + \boldsymbol{\xi}_+) + \dfrac{1}{2}\,d\,\|\boldsymbol{p}\|^2 \\ \text{subject to} & p_k + 1 - \xi_{-i} \le \boldsymbol{w}^\top \boldsymbol{x}^i \le p_{k+1} - 1 + \xi_{+i} \quad \text{for } i \in N(k) \text{ and for } k \in L \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \ge \mathbf{0}_n, \end{array} \right.$$

and

$$(S_{22}) \quad \left| \begin{array}{ll} \text{minimize} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + \dfrac{1}{2}\,c\,(\|\boldsymbol{\xi}_-\|^2 + \|\boldsymbol{\xi}_+\|^2) + \dfrac{1}{2}\,d\,\|\boldsymbol{p}\|^2 \\ \text{subject to} & p_k + 1 - \xi_{-i} \le \boldsymbol{w}^\top \boldsymbol{x}^i \le p_{k+1} - 1 + \xi_{+i} \qquad \text{for } i \in N(k) \text{ and for } k \in L, \end{array} \right.$$

where d is a penalty parameter.

Naturally, we could add to each of the above formulations the constraints

$$p_{k'} + 1 - \xi_{-i} \le \boldsymbol{w}^\top \boldsymbol{x}^i \le p_{k''} - 1 + \xi_{+i} \quad \text{for } k', k'' \in L \text{ such that } k' \le k < k''$$

for $i \in N(k)$. It would, however, inflate the problem size and most of those constraints would be likely redundant. Therefore we will not discuss this formulation.

7. 1-Norm Penalty and Kernel Technique

In this section we will make the Lagrangian dual of the soft margin problem (S_1) and show how the kernel technique applies to the problem.

7.1. Dual of Soft Margin Problem (S_1).

The Lagrangian function for (S_1) is

$$\begin{aligned} L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - (X(\boldsymbol{\alpha} - \boldsymbol{\beta}))^\top \boldsymbol{w} + (A\boldsymbol{\alpha} - B\boldsymbol{\beta})^\top \boldsymbol{p} \\ + \boldsymbol{\alpha}^\top \mathbf{1}_n + \boldsymbol{\beta}^\top \mathbf{1}_n + (c\mathbf{1}_n - \boldsymbol{\alpha})^\top \boldsymbol{\xi}_- + (c\mathbf{1}_n - \boldsymbol{\beta})^\top \boldsymbol{\xi}_+ \\ - \boldsymbol{\lambda}^\top \boldsymbol{\xi}_- - \boldsymbol{\mu}^\top \boldsymbol{\xi}_+, \end{aligned} \tag{7.1}$$

and the Lagrangian relaxation problem for a given nonnegative multiplier vector $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^{4n}_+$ is

$$\left| \begin{array}{ll} \text{minimize} & L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} & (\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+) \in \mathbb{R}^{m+l+2n}. \end{array} \right.$$

Thank to the convexity of the problem, the optimality condition of $(\boldsymbol{w}^*, \boldsymbol{p}^*, \boldsymbol{\xi}^*_-, \boldsymbol{\xi}^*_+)$ is simply given as

$$\begin{aligned} \frac{\partial L}{\partial(\boldsymbol{w}, \boldsymbol{p})}(\boldsymbol{w}^*, \boldsymbol{p}^*, \boldsymbol{\xi}^*_-, \boldsymbol{\xi}^*_+, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}_{n+l} \\ \frac{\partial L}{\partial(\boldsymbol{\xi}_-, \boldsymbol{\xi}_+)}(\boldsymbol{w}^*, \boldsymbol{p}^*, \boldsymbol{\xi}^*_-, \boldsymbol{\xi}^*_+, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}_{2n}, \end{aligned}$$

each of which reduces to

$$\boldsymbol{w}^* = X(\boldsymbol{\alpha} - \boldsymbol{\beta}) \tag{7.2}$$

$$A\boldsymbol{\alpha} - B\boldsymbol{\beta} = \mathbf{0}_l \tag{7.3}$$

$$\boldsymbol{\alpha} \le c\mathbf{1}_n \tag{7.4}$$

$$\boldsymbol{\beta} \le c\mathbf{1}_n \tag{7.5}$$

by virtue of the nonnegativity of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. The *complementarity condition*

$$(c\mathbf{1}_n - \boldsymbol{\alpha}^*)^\top \boldsymbol{\xi}^*_- = (c\mathbf{1}_n - \boldsymbol{\beta}^*)^\top \boldsymbol{\xi}^*_+ = 0 \tag{7.6}$$

holds for a primal optimal solution $(\boldsymbol{\xi}^*_-, \boldsymbol{\xi}^*_+)$ and a dual optimal solution $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. Denoting the optimum objective function value of the above relaxation problem by $\omega(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, the *Lagrangian dual of the soft margin problem* is

$$\left| \begin{array}{ll} \text{maximize} & \omega(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} & (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \ge \mathbf{0}_{4n} \end{array} \right.$$

Plugging (7.2), (7.3) and (7.6) into the Lagrangian function, the Lagrangian dual problem is rewritten as

$$(dS_1)\quad \left|\begin{array}{ll} \text{minimize} & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top K(\boldsymbol{\alpha}-\boldsymbol{\beta}) - \mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta}) \\ \text{subject to} & A\boldsymbol{\alpha} - B\boldsymbol{\beta} = \mathbf{0}_l \\ & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)} \\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)} \\ & \mathbf{0}_n \le \boldsymbol{\alpha} \le c\mathbf{1}_n \\ & \mathbf{0}_n \le \boldsymbol{\beta} \le c\mathbf{1}_n. \end{array}\right.$$

This differs from the dual of the hard margin problem (dH) only in the additional upper bound constraints (7.4) and (7.5) of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

7.2. **Kernel Technique.** Kernel technique can apply to the soft margin problem in the same way as discussed in Section 5. The problem to solve is the dual problem (dS_1) in the previous subsection with K replaced by $\tilde{K}$.

An optimal solution $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ of the kernel version of (dS_1) and an optimal solution $(\boldsymbol{\xi}_-^*, \boldsymbol{\xi}_+^*)$ of the primal problem meet the *complementarity condition*

$$(c\mathbf{1}_n - \boldsymbol{\alpha}^*)^\top \boldsymbol{\xi}_-^* = (c\mathbf{1}_n - \boldsymbol{\beta}^*)^\top \boldsymbol{\xi}_+^* = 0.$$

Then we have

$$\begin{aligned} \alpha_i^* < 1 \ \text{ implies } \ \xi_{-i}^* = 0, \ \text{i.e., } p_k^* + 1 \le (\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}}^i \\ \beta_i^* < 1 \ \text{ implies } \ \xi_{+i}^* = 0, \ \text{i.e., } (\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}}^i \le p_{k+1}^* - 1. \end{aligned}$$

Therefore the thresholds should be determined by

$$p_k^* = \frac{1}{2}\left(\max_{i \in N(k-1); \beta_i^* < 1} (\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}}^i + \min_{i \in N(k); \alpha_i^* < 1} (\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}}^i \right). \tag{7.7}$$

In the same way as in Section 5 we have

$$(\tilde{\boldsymbol{w}}^*)^\top \tilde{\boldsymbol{x}} = (\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)^\top \begin{pmatrix} \vdots \\ \kappa(\boldsymbol{x}^i, \boldsymbol{x}) \\ \vdots \end{pmatrix}. \tag{7.8}$$

for $\boldsymbol{x} \in \mathbb{R}^n$. Thus p_k^*'s can be obtained without knowing $\tilde{\boldsymbol{w}}^*$.

8. Dual of Other Formulations

Since the derivation of the optimality condition and the dual problem is technical, we defer the complete derivation to Appendix. We also omit the application of kernel technique.

8.1. **Dual of Soft Margin Problem (S_2).** The dual of (S_2) is

$$(dS_2)\quad \left|\begin{array}{ll} \text{minimize} & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top K(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{1}{2c}(\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + \boldsymbol{\beta}^\top\boldsymbol{\beta}) - \mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta}) \\ \text{subject to} & A\boldsymbol{\alpha} - B\boldsymbol{\beta} = \mathbf{0}_l \\ & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)} \\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)} \\ & \boldsymbol{\alpha} \ge \mathbf{0}_n \\ & \boldsymbol{\beta} \ge \mathbf{0}_n. \end{array}\right.$$

With the quadratic term $(\boldsymbol{\alpha}^\top\boldsymbol{\alpha}+\boldsymbol{\beta}^\top\boldsymbol{\beta})$ added, the objective function becomes strictly convex, which may lighten the computational burden. Furthermore,

$$c\,\boldsymbol{\xi}_-^* = \boldsymbol{\alpha}^* \quad \text{and} \quad c\,\boldsymbol{\xi}_+^* = \boldsymbol{\beta}^*$$

hold between a primal and a dual optimum solutions. From (8.1), we have

$$\alpha_i^* = 0 \ \text{ implies } \ p_{l_i}^* + 1 \le (\boldsymbol{w}^*)^\top \boldsymbol{x}^i$$
$$\beta_i^* = 0 \ \text{ implies } \ (\boldsymbol{w}^*)^\top \boldsymbol{x}^i \le p_{l_i+1}^* - 1.$$

Therefore the optimal threshold $\boldsymbol{p}^*$ of the primal problem (S_2) is determined by

$$p_k^* = \frac{1}{2}\left(\max_{i\in N(k-1);\beta_i^*=0}(\boldsymbol{w}^*)^\top\boldsymbol{x}^i + \min_{i\in N(k);\alpha_i^*=0}(\boldsymbol{w}^*)^\top\boldsymbol{x}^i\right).$$

8.2. **Dual of Soft Margin Problem** (S_{12})**.** The dual of (S_{12}) is

$$(dS_{12})\left|\begin{array}{ll} \text{minimize} & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top K(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{1}{2d}(\boldsymbol{\alpha}^\top\ \boldsymbol{\beta}^\top)M\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} - \mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta}) \\ \text{subject to} & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)} \\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)} \\ & \mathbf{0}_n \le \boldsymbol{\alpha} \le c\mathbf{1}_n \\ & \mathbf{0}_n \le \boldsymbol{\beta} \le c\mathbf{1}_n, \end{array}\right.$$

and the optimal solution $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ of (S_{12}) is determined by

$$\boldsymbol{w}^* = X(\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)$$
$$\boldsymbol{p}^* = -\frac{1}{d}(A\boldsymbol{\alpha}^* - B\boldsymbol{\beta}^*).$$

8.3. **Dual of Soft Margin Problem** (S_{22})**.** The dual of (S_{22}) is

$$(dS_{22})\left|\begin{array}{ll} \text{minimize} & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top K(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{1}{2c}(\boldsymbol{\alpha}^\top\boldsymbol{\alpha}+\boldsymbol{\beta}^\top\boldsymbol{\beta}) + \frac{1}{2d}(\boldsymbol{\alpha}^\top\boldsymbol{\beta}^\top)M\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} \\ & \quad -\mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta}) \\ \text{subject to} & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)} \\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)} \\ & \boldsymbol{\alpha} \ge \mathbf{0}_n \\ & \boldsymbol{\beta} \ge \mathbf{0}_n, \end{array}\right.$$

and the optimal solution $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ of (S_{22}) is determined by

$$\boldsymbol{w}^* = X(\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)$$
$$\boldsymbol{p}^* = -\frac{1}{d}(A\boldsymbol{\alpha}^* - B\boldsymbol{\beta}^*).$$

References

[1] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[2] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge, 2000.

[3] G.M. Fung and O.L. Mangasarian, "Multicategory proximal support vector machine classifiers", *Machine Learning* **59** (2005) 77–97.

[4] O.L. Mangasarian and D.R. Musicant, "Successive overrelaxation for support vector machines", *IEEE Transactions on Neural Networks* **10** (1999) 5, 1032–1037.

[5] K. Tatsumi, K. Hayashida, R. Kawachi and T. Tanino, "Multiobjective muticlass support vector machines maximizing geometric margins", *Pacific Journal of Optimization* **6** (2010) 115–140.

[6] V.N. Vapnik, *Statistical Learning Theory*, John-Wiley & Sons, New York, 1998.

[7] J. Weston and C. Watkins, "Multi-class support vector machines", Technical Report CSD-TR-98-04, Department of Computer Science, Royal Hloolway, University of London, 1998.

Appendix A. Dual of Soft Margin Problem (S_2)

The Lagrangian function for the problem (S_2) is

$$\begin{aligned} L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+, \boldsymbol{\alpha}, \boldsymbol{\beta}) =& \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - (X(\boldsymbol{\alpha} - \boldsymbol{\beta}))^\top \boldsymbol{w} + (A\boldsymbol{\alpha} - B\boldsymbol{\beta})^\top \boldsymbol{p} \\ & + \boldsymbol{\alpha}^\top \mathbf{1}_n + \boldsymbol{\beta}^\top \mathbf{1}_n + (\frac{1}{2}c\,\boldsymbol{\xi}_- - \boldsymbol{\alpha})^\top \boldsymbol{\xi}_- + (\frac{1}{2}c\,\boldsymbol{\xi}_+ - \boldsymbol{\beta})^\top \boldsymbol{\xi}_+, \end{aligned}$$

and the Lagrangian relaxation problem for a given $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{2n}_+$ is

$$\left| \begin{array}{ll} \text{minimize} & L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{subject to} & (\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+) \in \mathbb{R}^{m+l+2n}. \end{array} \right.$$

Since L is a convex function with respect to $(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+)$, the optimality condition of $(\boldsymbol{w}^*, \boldsymbol{p}^*, \boldsymbol{\xi}^*_-, \boldsymbol{\xi}^*_+)$ for a given Lagrangian multiplier vector $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is

$$\frac{\partial L}{\partial(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+)}(\boldsymbol{w}^*, \boldsymbol{p}^*, \boldsymbol{\xi}^*_-, \boldsymbol{\xi}^*_+) = \mathbf{0}_{m+l+2n},$$

which reduces to

$$\boldsymbol{w}^* = X(\boldsymbol{\alpha} - \boldsymbol{\beta}) \tag{A.1}$$

$$A\boldsymbol{\alpha} - B\boldsymbol{\beta} = \mathbf{0}_l \tag{A.2}$$

$$c\,\boldsymbol{\xi}^*_- - \boldsymbol{\alpha} = c\,\boldsymbol{\xi}^*_+ - \boldsymbol{\beta} = \mathbf{0}_n. \tag{A.3}$$

Substituting (A.1), (A.2) and (A.3) for L, we obtain the optimal value $\omega(\boldsymbol{\alpha}, \boldsymbol{\beta})$ of the Lagrangian relaxation problem

$$\omega(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\left\{ \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top K(\boldsymbol{\alpha} - \boldsymbol{\beta}) - \mathbf{1}^\top(\boldsymbol{\alpha} + \boldsymbol{\beta}) + \frac{1}{2c}(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\beta}) \right\}.$$

Then the Lagrangian dual problem of (S_2) forms the following (dS_2):

$$(dS_2) \quad \left| \begin{array}{ll} \text{minimize} & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top K(\boldsymbol{\alpha} - \boldsymbol{\beta}) + \frac{1}{2c}(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\beta}) - \mathbf{1}_n^\top(\boldsymbol{\alpha} + \boldsymbol{\beta}) \\ \text{subject to} & A\boldsymbol{\alpha} - B\boldsymbol{\beta} = \mathbf{0}_l \\ & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)} \\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)} \\ & \boldsymbol{\alpha} \geq \mathbf{0}_n \\ & \boldsymbol{\beta} \geq \mathbf{0}_n. \end{array} \right.$$

From (A.3) we have

$$\begin{aligned} \alpha^*_i = 0 \;\text{ implies }\; & p^*_{\ell_i} + 1 \leq (\boldsymbol{w}^*)^\top \boldsymbol{x}^i \\ \beta^*_i = 0 \;\text{ implies }\; & (\boldsymbol{w}^*)^\top \boldsymbol{x}^i \leq p^*_{\ell_i+1} - 1. \end{aligned}$$

Therefore the optimal thresholds $\boldsymbol{p}^*$ of (S_2) are determined by

$$p^*_k = \frac{1}{2}\left(\max_{i \in N(k-1); \beta^*_i = 0} (\boldsymbol{w}^*)^\top \boldsymbol{x}^i + \min_{i \in N(k); \alpha^*_i = 0} (\boldsymbol{w}^*)^\top \boldsymbol{x}^i \right).$$

Appendix B. Dual of Soft Margin Problem (S_{12})

The Lagrangian function of (S_{12}) is

$$\begin{aligned}L(\boldsymbol{w},\boldsymbol{p},\boldsymbol{\xi}_-,\boldsymbol{\xi}_+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\mu}) =& \frac{1}{2}\|\boldsymbol{w}\|^2 + c\mathbf{1}_n^\top(\boldsymbol{\xi}_- + \boldsymbol{\xi}_+) + \frac{1}{2}d\,\|\boldsymbol{p}\|^2 \\ &+ \sum_{i\in N(1..l)} \alpha_i(1 - (\boldsymbol{x}^i)^\top\boldsymbol{w} + p_{l_i} - \xi_{-i}) \\ &+ \sum_{i\in N(0..l-1)} \beta_i(1 + (\boldsymbol{x}^i)^\top\boldsymbol{w} - p_{l_i+1} - \xi_{+i}) - \boldsymbol{\lambda}^\top\boldsymbol{\xi}_- - \boldsymbol{\mu}^\top\boldsymbol{\xi}_+ \\ =& \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \frac{1}{2}d\,\boldsymbol{p}^\top\boldsymbol{p} - (X(\boldsymbol{\alpha}-\boldsymbol{\beta}))^\top\boldsymbol{w} + (A\boldsymbol{\alpha} - B\boldsymbol{\beta})^\top\boldsymbol{p} \\ &+ \mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta}) + (c\mathbf{1}_n - \boldsymbol{\alpha} - \boldsymbol{\lambda})^\top\boldsymbol{\xi}_- + (c\mathbf{1}_n - \boldsymbol{\beta} - \boldsymbol{\mu})^\top\boldsymbol{\xi}_+.\end{aligned}$$

where $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are nonnegative Lagrange multiplier vectors corresponding to the inequality constraints and the nonnegativity of $\boldsymbol{\xi}_-$ and $\boldsymbol{\xi}_+$, respectively. For a given multiplier vector $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\mu}) \in \mathbb{R}^{4n}_+$, the Lagrangian relaxation problem is given as

$$\left|\begin{array}{ll}\text{minimize} & L(\boldsymbol{w},\boldsymbol{p},\boldsymbol{\xi}_-,\boldsymbol{\xi}_+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\mu}) \\ \text{subject to} & (\boldsymbol{w},\boldsymbol{p},\boldsymbol{\xi}_-,\boldsymbol{\xi}_+) \in \mathbb{R}^{m+l+2n}.\end{array}\right.$$

In the same way as the previous discussion, the optimality condition of $(\boldsymbol{w}^*,\boldsymbol{p}^*,\boldsymbol{\xi}_-^*,\boldsymbol{\xi}_+^*)$ is

$$\frac{\partial L}{\partial(\boldsymbol{w},\boldsymbol{p},\boldsymbol{\xi}_-,\boldsymbol{\xi}_+)}(\boldsymbol{w}^*,\boldsymbol{p}^*,\boldsymbol{\xi}_-^*,\boldsymbol{\xi}_+^*) = \mathbf{0}_{m+l+2n},$$

which reduces to

$$\boldsymbol{w}^* = X(\boldsymbol{\alpha}-\boldsymbol{\beta}) \tag{B.1}$$

$$d\,\boldsymbol{p}^* = -(A\boldsymbol{\alpha} - B\boldsymbol{\beta}) \tag{B.2}$$

$$c\mathbf{1}_n - \boldsymbol{\alpha} - \boldsymbol{\lambda} = c\mathbf{1}_n - \boldsymbol{\beta} - \boldsymbol{\mu} = \mathbf{0}_n. \tag{B.3}$$

Then the optimal value $\omega(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\mu})$ of the Lagrangian relaxation problem is given as

$$\begin{aligned}\omega(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\mu}) &= -\left\{\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top K(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{1}{2d}(A\boldsymbol{\alpha} - B\boldsymbol{\beta})^\top(A\boldsymbol{\alpha} - B\boldsymbol{\beta}) - \mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta})\right\} \\ &= -\left\{\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top K(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{1}{2d}(\boldsymbol{\alpha}^\top\ \boldsymbol{\beta}^\top)M\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} - \mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta})\right\},\end{aligned}$$

where

$$M = \begin{bmatrix} A^\top A & -A^\top B \\ -B^\top A & B^\top B \end{bmatrix} \in \mathbb{R}^{2n\times 2n}.$$

Then the Lagrangian dual of (S_{12}) is

$$(dS_{12})\quad \left|\begin{array}{ll}\text{minimize} & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top K(\boldsymbol{\alpha}-\boldsymbol{\beta}) + \frac{1}{2d}(\boldsymbol{\alpha}^\top\ \boldsymbol{\beta}^\top)M\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} - \mathbf{1}_n^\top(\boldsymbol{\alpha}+\boldsymbol{\beta}) \\ \text{subject to} & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)} \\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)} \\ & \mathbf{0}_n \le \boldsymbol{\alpha} \le c\mathbf{1}_n \\ & \mathbf{0}_n \le \boldsymbol{\beta} \le c\mathbf{1}_n.\end{array}\right.$$

The optimal solution $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ of (S_{12}) is determined by

$$\begin{aligned}\boldsymbol{w}^* &= X(\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)\\ \boldsymbol{p}^* &= -\frac{1}{d}(A\boldsymbol{\alpha}^* - B\boldsymbol{\beta}^*)\end{aligned}$$

from the optimality condition.

Appendix C. Dual of Soft Margin Problem (S_{22})

The Lagrangian function of (S_{22}) is

$$\begin{aligned}L(\boldsymbol{w}, \boldsymbol{p}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+, \boldsymbol{\alpha}, \boldsymbol{\beta}) =& \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}c(\|\boldsymbol{\xi}_-\|^2 + \|\boldsymbol{\xi}_+\|^2) + \frac{1}{2}d\|\boldsymbol{p}\|^2\\ &+ \sum_{i\in N(1..l)} \alpha_i(1 - (\boldsymbol{x}^i)^\top \boldsymbol{w} + p_{l_i} - \xi_{-i})\\ &+ \sum_{i\in N(0..l-1)} \beta_i(1 + (\boldsymbol{x}^i)^\top \boldsymbol{w} - p_{l_i+1} - \xi_{+i}),\\ =& \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \frac{1}{2}d(\boldsymbol{p}^\top \boldsymbol{p}) - (X(\boldsymbol{\alpha} - \boldsymbol{\beta}))^\top \boldsymbol{w} + (A\boldsymbol{\alpha} - B\boldsymbol{\beta})^\top \boldsymbol{p}\\ &+ \mathbf{1}_n^\top(\boldsymbol{\alpha} + \boldsymbol{\beta}) + (\frac{1}{2}c\,\boldsymbol{\xi}_- - \boldsymbol{\alpha})^\top \boldsymbol{\xi}_- + (\frac{1}{2}c\,\boldsymbol{\xi}_+ - \boldsymbol{\beta})^\top \boldsymbol{\xi}_+.\end{aligned}$$

The optimality condition of $(\boldsymbol{w}^*, \boldsymbol{p}^*, \boldsymbol{\xi}_-^*, \boldsymbol{\xi}_+^*)$ is

$$\begin{aligned}\boldsymbol{w}^* &= X(\boldsymbol{\alpha} - \boldsymbol{\beta})\\ d\,\boldsymbol{p}^* &= -(A\boldsymbol{\alpha} - B\boldsymbol{\beta})\\ c\,\boldsymbol{\xi}_-^* &= \boldsymbol{\alpha}\\ c\,\boldsymbol{\xi}_+^* &= \boldsymbol{\beta}.\end{aligned}$$

Plugging these equalities into L, $\omega(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is given as

$$\begin{aligned}\omega(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\Big\{&\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top K(\boldsymbol{\alpha} - \boldsymbol{\beta}) + \frac{1}{2c}(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\beta})\\ &+ \frac{1}{2d}(A\boldsymbol{\alpha} - B\boldsymbol{\beta})^\top (A\boldsymbol{\alpha} - B\boldsymbol{\beta}) - \mathbf{1}_n^\top(\boldsymbol{\alpha} + \boldsymbol{\beta})\Big\}.\end{aligned}$$

Therefore we obtain the Lagrangian dual of (S_{22}):

$$(dS_{22})\quad \left|\begin{array}{ll} \text{minimize} & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top K(\boldsymbol{\alpha} - \boldsymbol{\beta}) + \frac{1}{2c}(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\beta}) + \frac{1}{2d}(\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top) M \begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}\\ & \quad -\mathbf{1}_n^\top(\boldsymbol{\alpha} + \boldsymbol{\beta})\\ \text{subject to} & \boldsymbol{\alpha}_{N(0)} = \mathbf{0}_{n(0)}\\ & \boldsymbol{\beta}_{N(l)} = \mathbf{0}_{n(l)}\\ & \boldsymbol{\alpha} \geq \mathbf{0}_n\\ & \boldsymbol{\beta} \geq \mathbf{0}_n.\end{array}\right.$$

The optimal solution $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ of (S_{22}) is determined by

$$\begin{aligned}\boldsymbol{w}^* &= X(\boldsymbol{\alpha}^* - \boldsymbol{\beta}^*)\\ \boldsymbol{p}^* &= -\frac{1}{d}(A\boldsymbol{\alpha}^* - B\boldsymbol{\beta}^*).\end{aligned}$$

(Y. Izunaga) Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
E-mail address: s1130131@sk.tsukuba.ac.jp

(Y. Yamamoto) Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
E-mail address: yamamoto@sk.tsukuba.ac.jp