On p-norm linear discrimination

Yana Morenko* Alexander Vinel* Zhaohan Yu* Pavlo Krokhmal*,†

Abstract

We consider a p-norm linear discrimination model that generalizes the model of Bennett and Mangasarian (1992) and reduces to a linear programming problem with p-order cone constraints. The proposed approach for handling linear programming problems with p-order cone constraints is based on reformulation of p-order cone optimization problems as second order cone programming (SOCP) problems when p is rational. Since such reformulations typically lead to SOCP problems with large numbers of second order cones, an "economical" representation that minimizes the number of second order cones is proposed. A case study illustrating the developed model on several popular data sets is conducted.

1 Introduction

Consider two discrete sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^n$ containing k and m points, respectively: $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$, $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$. One of the principal tasks arising in machine learning and data mining is that of *discrimination* of such sets, namely, constructing a surface $f(\mathbf{x}) = 0$ such that $f(\mathbf{x}) < 0$ for any $\mathbf{x} \in \mathcal{A}$ and $f(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{B}$. Of particular interest is the linear separating surface (hyperplane): $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - \gamma = 0$. From the simple fact that any two points $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$ satisfying the inequalities $\mathbf{w}^\top \mathbf{y}_1 - \gamma > 0$, $\mathbf{w}^\top \mathbf{y}_2 - \gamma < 0$ for some $\mathbf{w}$ and γ are located on the opposite sides of the hyperplane $\mathbf{w}^\top \mathbf{x} - \gamma = 0$, it follows that the discrete sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^n$ are considered *linearly separable* if and only if there exist $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{w}^\top \mathbf{a}_i > \gamma > \mathbf{w}^\top \mathbf{b}_j$ for all $i = 1, \ldots, k$, $j = 1, \ldots, m$, with an appropriately chosen γ, or, equivalently,

$$\min_{\mathbf{a}_i \in \mathcal{A}} \mathbf{a}_i^\top \mathbf{w} \;>\; \max_{\mathbf{b}_j \in \mathcal{B}} \mathbf{b}_j^\top \mathbf{w}. \tag{1}$$

Clearly, existence of such a separating hyperplane is not guaranteed (namely, a separating hyperplane exists if the convex hulls of sets $\mathcal{A}$ and $\mathcal{B}$ are disjoint); thus, in general, a separating hyperplane that minimizes some sort of *misclassification error* is desired.

In the next section we introduce a new linear separation model that is based on p-order cone programming, and discuss its key properties. The proposed solution approach, based on a reformulation of p-cone programming problems as second order cone programming (SOCP) problems when p is rational, is presented in Section 3. Section 4 contains a case study on several popular data sets that illustrates the developed discrimination model.

2 p-Norm linear separation: A stochastic optimization analogy

Since definition (1) involves strict inequalities, it is not well suited for mathematical programming models of selecting the "best" linear separator. However, the fact that the separating hyperplane can be scaled by any non-negative factor allows one to formulate the following observation:

Proposition 1 ([4]) *Discrete sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^n$ represented by matrices $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_k)^\top \in \mathbb{R}^{k\times n}$ and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_m)^\top \in \mathbb{R}^{m\times n}$, respectively, are linearly separable if and only if*

$$\mathbf{A}\mathbf{w} \geq \mathbf{e}\gamma + \mathbf{e}, \quad \mathbf{B}\mathbf{w} \leq \mathbf{e}\gamma - \mathbf{e} \quad \textit{for some} \quad \mathbf{w} \in \mathbb{R}^n,\ \gamma \in \mathbb{R}, \tag{2}$$

*Department of Mechanical and Industrial Engineering, University of Iowa, Iowa City, IA 52242, USA
†Corresponding author. E-mail: krokhmal@engineering.uiowa.edu

where $\mathbf{e}$ *is the vector of ones of an appropriate dimension,* $\mathbf{e} = (1, \ldots, 1)^\top$.

Given the linear separability condition (2), the (non-negative) vectors $\mathbf{x}_{\mathcal{A}} = (-\mathbf{Aw} + \mathbf{e}\gamma + \mathbf{e})_+$, $\mathbf{x}_{\mathcal{B}} = (\mathbf{Bw} - \mathbf{e}\gamma + \mathbf{e})_+$, where $t_+ = \max\{0, t\}$, represent misclassification errors: $\mathbf{x}_{\mathcal{A}}$ and/or $\mathbf{x}_{\mathcal{B}} > \mathbf{0}$ if sets $\mathcal{A}$ and $\mathcal{B}$ are not linearly separable. If one considers that points of sets $\mathcal{A}$ and $\mathcal{B}$ represent realizations of (discretely distributed) random vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, respectively, the corresponding elements of vectors $\mathbf{x}_{\mathcal{A}}$, $\mathbf{x}_{\mathcal{B}}$ may be regarded as realizations of random variables $X_{\mathcal{A}}(\mathbf{a}; \mathbf{w}, \gamma) = (-\mathbf{a}^\top\mathbf{w} + \gamma + 1)_+$, $X_{\mathcal{B}}(\mathbf{b}; \mathbf{w}, \gamma) = (\mathbf{b}^\top\mathbf{w} - \gamma + 1)_+$, respectively, that depend parametrically on the decision variables $\mathbf{w}$ and γ. Then, a plausible strategy for selecting $\mathbf{w}$ and γ is one that minimizes, for example, the expected misclassification errors, and which can be formulated as the following stochastic programming problem:

$$\min_{(\mathbf{w},\gamma) \in \mathbb{R}^{n+1}} \Big\{\delta_1 \mathsf{E}\big[(-\mathbf{a}^\top\mathbf{w} + \gamma + 1)_+\big] + \delta_2 \mathsf{E}\big[(\mathbf{b}^\top\mathbf{w} - \gamma + 1)_+\big]\Big\},$$

where $\delta_{1,2}$ serve as "importance" weights of the misclassification errors for points of sets $\mathcal{A}$ and $\mathcal{B}$, respectively. Further, instead of minimizing the expected misclassification error, one may select the parameters $\mathbf{w}$ and γ so as to minimize the risk of misclassification. As it is well known in stochastic optimization and risk analysis, the "risk" associated with random outcome of a decision under uncertainty is often attributed to the "heavy" tails of the corresponding probability distribution. The risk-inducing "heavy" tails of probability distributions, are, in turn, characterized by the distribution's higher moments. Thus, if the misclassifications introduced by a separating hyperplane can be viewed as "random", the misclassification risk may be controlled better if one minimizes not the average, or expected misclassification errors, but their moments of order $p > 1$. This gives rise to the following formulation for linear discrimination of sets $\mathcal{A}$ and $\mathcal{B}$:

$$\min_{(\mathbf{w},\gamma) \in \mathbb{R}^{n+1}} \delta_1 \big\|(-\mathbf{a}^\top\mathbf{w} + \gamma + 1)_+\big\|_p + \delta_2 \big\|(\mathbf{b}^\top\mathbf{w} - \gamma + 1)_+\big\|_p, \quad p \in [1, +\infty], \tag{3}$$

where $\|\cdot\|_p$ is the usual $\mathcal{L}_p$ norm: $\|Y\|_p = \big(\mathsf{E}|Y|^p\big)^{1/p}$ if $p \in [1, \infty)$, and $\|Y\|_\infty = \operatorname{ess\,sup}|Y|$. If $\mathbf{a}$ and $\mathbf{b}$ are uniformly distributed with support sets $\mathcal{A}$ and $\mathcal{B}$, respectively:

$$\mathsf{P}(\mathbf{a} = \mathbf{a}_i) = 1/k, \quad \mathsf{P}(\mathbf{b} = \mathbf{b}_j) = 1/m \quad \text{for all} \quad \mathbf{a}_i \in \mathcal{A}, \ \mathbf{b}_j \in \mathcal{B}, \tag{4}$$

the p-norm linear discrimination problem takes the form

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{n+1}} \ \frac{\delta_1}{k^{1/p}} \big\|(-\mathbf{Aw} + \mathbf{e}\gamma + \mathbf{e})_+\big\|_p + \frac{\delta_2}{m^{1/p}} \big\|(\mathbf{Bw} - \mathbf{e}\gamma + \mathbf{e})_+\big\|_p, \tag{5}$$

where $\|\cdot\|_p$ is a norm in Euclidean space of an appropriate dimension: $\|\mathbf{u}\|_p = (|u_1|^p + \ldots + |u_l|^p)^{1/p}$, $p \in [1, \infty)$ and $\|\mathbf{u}\|_\infty = \max_{i=1,\ldots,l}\{u_i\}$ (in the sequel, it shall be clear from the context whether the $\mathcal{L}_p$ or Euclidean p-norm is used). Further, (5) can be formulated as a p-order cone programming problem (pOCP)

$$\begin{aligned}
\min \quad & \delta_1 k^{-1/p}\, \xi + \delta_2 m^{-1/p}\, \eta && \text{(6a)}\\
\text{s. t.} \quad & \xi \geq \|\mathbf{y}\|_p, && \text{(6b)}\\
& \eta \geq \|\mathbf{z}\|_p, && \text{(6c)}\\
& \mathbf{y} \geq -\mathbf{Aw} + \mathbf{e}\gamma + \mathbf{e}, && \text{(6d)}\\
& \mathbf{z} \geq \mathbf{Bw} - \mathbf{e}\gamma + \mathbf{e}, && \text{(6e)}\\
& \mathbf{z}, \mathbf{y} \geq \mathbf{0}. && \text{(6f)}
\end{aligned}$$

Note that the special case of $p = 1$ and $\delta_1 = \delta_2$ corresponds to the linear discrimination model of Bennett and Mangasarian [4]. The p-cone programming linear separation model (3)–(6) shares many key properties with the LP separation model [4], including the guarantee that an optimal solution of (6) is non-zero in $\mathbf{w}$ for linearly separable sets.

Proposition 2 *When sets* $\mathcal{A}$ *and* $\mathcal{B}$*, represented by matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, are linearly separable, the separating hyperplane* $\mathbf{w}^{*\top}\mathbf{x} = \gamma^*$ *given by an optimal solution of* (5)–(6) *satisfies* $\mathbf{w}^* \neq \mathbf{0}$.

Proof: Zero optimal value of (6a) entails that $-\mathbf{A}\mathbf{w}^* + \mathbf{e}\gamma^* + \mathbf{e} \le \mathbf{0}$, $\mathbf{B}\mathbf{w}^* - \mathbf{e}\gamma^* + \mathbf{e} \le \mathbf{0}$ at optimality, which requires that $\gamma^* \le -1$ and $\gamma^* \ge 1$ simultaneously for $\mathbf{w}^* = \mathbf{0}$ to hold. ∎

Secondly, the p-norm separation model (6) can produce a solution with $\mathbf{w} = \mathbf{0}$ only in a rather special case that is identified by Theorem 1 below.

Theorem 1 *Consider the p-order cone programming problem (6)–(5), where it is assumed without loss of generality that $0 < \delta_1 \le \delta_2$. Then, for any $p \in (1, \infty)$ the p-order cone programming problem (6) has an optimal solution with $\mathbf{w}^* = \mathbf{0}$ if and only if*

$$\frac{\mathbf{e}^\top}{k}\mathbf{A} = \mathbf{v}^\top\mathbf{B}, \quad \text{where} \quad \mathbf{e}^\top\mathbf{v} = 1, \;\; \mathbf{v} \ge \mathbf{0}, \;\; \|\mathbf{v}\|_q \le \frac{\delta_2}{\delta_1 m^{1/p}}, \tag{7a}$$

where q satisfies $p^{-1} + q^{-1} = 1$. In other words, the arithmetic mean of the points in $\mathcal{A}$ must be equal to some convex combination of points in $\mathcal{B}$. In the case of $\delta_1 = \delta_2$ condition (7a) reduces to

$$\frac{\mathbf{e}^\top}{k}\mathbf{A} = \frac{\mathbf{e}^\top}{m}\mathbf{B}, \tag{7b}$$

i.e., the arithmetic means of the points of sets $\mathcal{A}$ and $\mathcal{B}$ must coincide.

Proof: First, let us consider the case when the p-cone discrimination model (6) has an optimal solution with $\mathbf{w}^* = \mathbf{0}$ and demonstrate that (7) must then hold. From the formulation (5) of problem (6) it follows that in the case when $\mathbf{w} = \mathbf{0}$ at optimality, the corresponding optimal value of the objective (6a) is determined as

$$\min_{\gamma \in \mathbb{R}} \left\{ \frac{\delta_1}{k^{1/p}} \left(\sum_{i=1}^{k} (1+\gamma)_+^p \right)^{1/p} + \frac{\delta_2}{m^{1/p}} \left(\sum_{j=1}^{m} (1-\gamma)_+^p \right)^{1/p} \right\} = 2\delta_1,$$

due to the assumption $0 < \delta_1 \le \delta_2$. Next, consider the dual of the p-cone programming problem (6):

$$\begin{aligned} \max \quad & \mathbf{e}^\top\mathbf{u} + \mathbf{e}^\top\mathbf{v} \\ \text{s. t.} \quad & -\mathbf{A}^\top\mathbf{u} + \mathbf{B}^\top\mathbf{v} = \mathbf{0}, \\ & \mathbf{e}^\top\mathbf{u} - \mathbf{e}^\top\mathbf{v} = \mathbf{0}, \\ & \mathbf{0} \le \mathbf{u} \le -\mathbf{s}, \\ & \mathbf{0} \le \mathbf{v} \le -\mathbf{t}, \\ & \|\mathbf{s}\|_q \le \delta_1 k^{-1/p}, \\ & \|\mathbf{t}\|_q \le \delta_2 m^{-1/p}, \end{aligned} \tag{8}$$

where q is such that $1/p + 1/q = 1$. Note that (6) is strictly feasible and bounded from below, since for any $\mathbf{w}_0$, γ_0 and $\varepsilon > 0$ one can select $\mathbf{y}_0 = \varepsilon\mathbf{e} + (-\mathbf{A}\mathbf{w}_0 + \mathbf{e}\gamma_0 + \mathbf{e})_+ > \mathbf{0}$, $\mathbf{z}_0 = \varepsilon\mathbf{e} + (\mathbf{B}\mathbf{w}_0 - \mathbf{e}\gamma_0 + \mathbf{e})_+ > \mathbf{0}$, $\xi_0 = (1+\varepsilon)\|\mathbf{y}_0\|_p > \|\mathbf{y}_0\|_p > 0$, and $\eta_0 = (1+\varepsilon)\|\mathbf{z}_0\|_p > \|\mathbf{z}_0\|_p > 0$ that are feasible to (6). Thus, the duality gap for the primal-dual pair of p-order cone programming problems (6) and (8) is zero [12]. Then, from the first two constraints of (8) we have $\mathbf{A}^\top\mathbf{u}^* = \mathbf{B}^\top\mathbf{v}^*$ as well as $\mathbf{e}^\top\mathbf{u}^* = \mathbf{e}^\top\mathbf{v}^*$, which, given that the optimal objective value of (8) is $2\delta_1$, implies that an optimal $\mathbf{u}^*$ must satisfy

$$\mathbf{e}^\top\mathbf{u}^* = \delta_1. \tag{9a}$$

Also, from (8) it follows that

$$\|\mathbf{u}^*\|_q \le \delta_1 k^{-1/p}. \tag{9b}$$

Then, it is easy to see that the unique solution of system (9) is

$$\mathbf{u}^* = \frac{\delta_1}{k}\mathbf{e} = \left(\frac{\delta_1}{k}, \ldots, \frac{\delta_1}{k} \right)^\top,$$

which corresponds to the point where the surface $(u_1^q + \ldots + u_k^q)^{1/q} = \delta_1 k^{-1/p}$ is tangent to the hyperplane $u_1 + \ldots + u_k = \delta_1$ in the positive of $\mathbb{R}^k$.

Likewise, an optimal $\mathbf{v}^*$ must satisfy $\mathbf{e}^\top \mathbf{v}^* = \delta_1$ and $\|\mathbf{v}^*\|_q \le \delta_2 m^{-1/p}$, but such $\mathbf{v}^*$ is not unique in the case $\delta_2/\delta_1 > 1$. By substituting the obtained characterizations for $\mathbf{u}^*$ and $\mathbf{v}^*$ in the constraint $\mathbf{A}^\top \mathbf{u}^* = \mathbf{B}^\top \mathbf{v}^*$ of the dual, we obtain (7a). When $\delta_1 = \delta_2$, the optimal $\mathbf{v}^*$ is unique: $\mathbf{v}^* = \dfrac{\delta_1}{m}\mathbf{e}$, and yields (7b).

To prove the statement of the Theorem in the opposite direction, assume that, for instance, (7a) holds for certain $\mathbf{u}$ and $\mathbf{v}$. Selecting $\mathbf{u}^* = (\delta_1/k)\,\mathbf{e}$, $\mathbf{v}^* = \delta_1 \mathbf{v}$, and $\mathbf{s}^* = -\mathbf{u}^*$, $\mathbf{t}^* = -\mathbf{v}^*$, it is easy to see that $(\mathbf{u}^*, \mathbf{v}^*, \mathbf{s}^*, \mathbf{t}^*)$ represents a feasible solution of the dual problem (8) with the dual cost of $2\delta_1$. Similarly, the tuple $(\mathbf{w}^*, \gamma^*, \mathbf{y}^*, \mathbf{z}^*, \xi^*, \eta^*)$, where $\mathbf{w}^* = \mathbf{0}$, $\gamma^* = 1$, $\mathbf{y}^* = (\mathbf{e}\gamma^* + \mathbf{e})_+ = 2\mathbf{e}$, $\mathbf{z}^* = (-\mathbf{e}\gamma^* + \mathbf{e})_+ = \mathbf{0}$, $\xi^* = \|\mathbf{y}^*\|_p = 2k^{1/p}$, $\eta^* = \|\mathbf{z}^*\|_p = 0$, represents a feasible solution of the primal problem (6) with the corresponding objective value of $2\delta_1$. Noting the zero duality gap for the constructed pair of feasible solutions of (6) and (8), and recalling that the primal problem is bounded and strictly feasible, we immediately obtain that this pair of primal-dual solutions is optimal [12]. Hence, from (7a) it follows that an optimal solution of (6) exists with $\mathbf{w}^* = \mathbf{0}$. ■

Observe that Theorem 1 implies that in the case of $\delta_1 = \delta_2$, the p-norm discrimination model (6) produces a null separating hyperplane only when the "geometric centers" of the sets $\mathcal{A}$ and $\mathcal{B}$ coincide. In practice, this means that such sets cannot be efficiently separated, at least by a hyperplane, thus an occurrence of a $\mathbf{w}^* = \mathbf{0}$ solution in (6) may be regarded not as a shortfall of formulation (6), but rather as the general unsuitability of such sets $\mathcal{A}$ and $\mathcal{B}$ to linear discrimination. In the case of $\delta_1 < \delta_2$, occurrence of a $\mathbf{w}^* = \mathbf{0}$ solution in (6) does not necessarily signify that sets $\mathcal{A}$ and $\mathcal{B}$ are hardly amenable to linear separation. In this case Theorem 1 only claims that the "geometric center" of $\mathcal{A}$ must lie within the convex hull of set $\mathcal{B}$, so that linear discrimination can still be a feasible approach, albeit at a cost of significant misclassification errors.

In order for a $\mathbf{w}^* = \mathbf{0}$ solution to occur only under the stricter condition (7b) when misclassification preferences for sets $\mathcal{A}$ and $\mathcal{B}$ are different, the p-norm linear discrimination model can be extended by applying norms of different orders to misclassifications of points in $\mathcal{A}$ and $\mathcal{B}$:

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{n+1}} k^{-1/p_1} \left\|(-\mathbf{A}\mathbf{w} + \mathbf{e}\gamma + \mathbf{e})_+\right\|_{p_1} + m^{-1/p_2} \left\|(\mathbf{B}\mathbf{w} - \mathbf{e}\gamma + \mathbf{e})_+\right\|_{p_2}, \quad p_{1,2} \in (1,\infty). \tag{10}$$

Intuitively, a norm of higher order places more "weight" on the outliers. For example, use of $p = 1$ norm entails minimization of the average of misclassifications; in contrast, application of the $p = \infty$ norm implies minimization of the largest misclassification for a set. Thus, by selecting appropriately the orders p_1 and p_2 in (10) one may introduce tolerance preferences on misclassifications of points of sets $\mathcal{A}$ and $\mathcal{B}$. At the same time, it can be shown that the occurrence of $\mathbf{w}^* = \mathbf{0}$ solution in (10) would signal the presence of the aforementioned singularity about the sets $\mathcal{A}$ and $\mathcal{B}$. Namely, we have

Theorem 2 *The p-order cone programming problem* (10), *where $p_1, p_2 \in (1,\infty)$, has an optimal solution with $\mathbf{w}^* = \mathbf{0}$ if and only if* (7b) *holds.*

We conclude this section by pointing out a connection between the p-norm separation model and the classical Support Vector Machine (SVM) model. SVM models are widely used in classification problems (see some recent works in, e.g., [5, 9, 14]). The linear SVM for non-separable sets can be written as a quadratic programming problem of the form

$$\begin{aligned}
\min \quad & \tfrac{1}{2}\|\mathbf{w}\|_2 + C_1 \mathbf{e}^\top \boldsymbol{\varepsilon}_1 + C_2 \mathbf{e}^\top \boldsymbol{\varepsilon}_2 && \text{(11a)}\\
\text{s. t.} \quad & \mathbf{A}\mathbf{w} - \mathbf{e}\gamma \ge \mathbf{e} - \boldsymbol{\varepsilon}_1 && \text{(11b)}\\
& -\mathbf{B}\mathbf{w} + \mathbf{e}\gamma \ge \mathbf{e} - \boldsymbol{\varepsilon}_2 && \text{(11c)}\\
& \boldsymbol{\varepsilon}_{1,2} \ge \mathbf{0} && \text{(11d)}
\end{aligned}$$

where $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are misclassification vectors for sets $\mathcal{A}$ and $\mathcal{B}$, respectively, and $C_1, C_2 > 0$.

Proposition 3 *If the misclassification weight coefficients in the p-norm separation model* (6) *and the SVM model* (11) *coincide, $C_1 = \delta_1/k_1$ and $C_2 = \delta_2/k_2$, the optimal value V^*_{SVM} of SVM problem* (11) *can be* *bounded as*

$$V_p^* \le V_{\text{SVM}}^* \le V_p^* + \tfrac{1}{2}\|\mathbf{w}^*\|_2,$$

where V_p^* *is the optimal value of* p*-norm problem* (6) *and* $\mathbf{w}^*$ *is an optimal solution of* (6).

Proof: By renaming variables $\boldsymbol{\varepsilon}_1 = \mathbf{y}$, $\boldsymbol{\varepsilon}_2 = \mathbf{z}$, problem (11) can be rewritten as

$$\min\left\{\tfrac{1}{2}\|\mathbf{w}\|_2 + C_1\xi + C_2\eta \;\middle|\; \xi \ge \|\mathbf{y}\|_1,\;\; \eta \ge \|\mathbf{z}\|_1,\; \text{(6d)},\; \text{(6e)},\; \text{(6f)}\right\}. \tag{12}$$

Setting $C_1 = \delta_1/k$, $C_2 = \delta_2/m$ and taking into account that $\|\mathbf{x}\|_p \ge \|\mathbf{x}\|_q$ for $1 \le p < q$, it is easy to see that

$$\begin{aligned}&\tfrac{1}{2}\|\mathbf{w}^*\|_2 + C_1\xi^* + C_2\eta^* \ge \tfrac{1}{2}\|\mathbf{w}^{**}\|_2 + C_1\xi^{**} + C_2\eta^{**}\\ &\ge C_1\xi^{**} + C_2\eta^{**} \ge C_1\,\xi^*_{(1)} + C_2\,\eta^*_{(1)} \ge C_1\xi^* + C_2\eta^*,\end{aligned}$$

where $\mathbf{w}^{**}, \xi^{**}, \eta^{**}$ are the optimal values of the variables in the SVM problem (12), $\mathbf{w}^*, \xi^*, \eta^*$ are optimal solutions of the p-norm separation model (6), and $\xi^*_{(1)}, \eta^*_{(1)}$ are optimal solutions of (6) with $p = 1$. ■

In the next section we discuss the details of practical implementation of the p-norm linear discrimination model (6).

3 A second order cone programming approach to p-order cone programming problems

The p-order cone constraints (6b)–(6c) are central to practical implementation of the p-norm separation method (6). In the special cases of $p = 1$ or $p = \infty$, p-order cone constraints reduce to linear inequalities; specifically, the $p = 1$ version of model (6) has been studied in [4]. In general, the amenability of 1-norm to implementation via linear constraints has been exploited in a variety approaches and applications, too numerous to cite here. Another prominent special case of is that of $p = 2$, when (6b)–(6c) represent second order, or quadratic cones. The second order cone programming (SOCP) constitutes a well-developed subject of convex optimization, and a number of efficient self-dual “long-step” interior point (IP) SOCP algorithms have been developed in the literature and implemented in software [1, 2, 13]. The “general” case of $p \in (1, 2) \cup (2, \infty)$, when the p-cone is not self-dual, has received relatively limited attention in the literature. IP approaches to p-order cone programming have been considered in, e.g., [6, 11, 15]; a polyhedral approximation approach was proposed in [10].

In this work, we pursue an approach to solving p-cone programming problems that is based on the possibility to represent a p-order cone via a sequence of second order cones when p is rational [1, 12]. Reformulation of a rational-order p-cone programming problem as a SOCP problem allows for employing the efficient self-dual SOCP methods, albeit at a cost of a large number of second order cones required for such a reformulation. Moreover, since such a reformulation is not unique, in Section 3.2 we introduce a constructive “economical” representation of rational-order p-cones via second order cones.

3.1 Representation of rational-order p-cones with second order cones

Without loss of generality, consider a p-cone in the positive orthant of $\mathbb{R}^{n+1}$

$$t \ge (w_1^p + \ldots + w_n^p)^{1/p}, \quad (t, w_1, \ldots, w_n)^\top \ge \mathbf{0}. \tag{13}$$

In the case when the parameter p is a positive rational number, $p = r/s$, where $r, s \in \mathbb{N}$, then, for instance, the following “lifted” representation of the p-cone set (13) can be constructed in $\mathbb{R}_+^{2n+1}$ [1, 10]:

$$t \ge u_1 + \ldots + u_n, \qquad u_j \ge 0, \quad j = 1, \ldots, n, \tag{14a}$$

$$w_j^R \le u_j^s t^{r-s} w^{R-r}, \qquad j = 1, \ldots, n, \tag{14b}$$

where $R = 2^\rho$, $\rho = \lceil \log_2 r \rceil$. Then, each nonlinear inequality (14b) can equivalently be replaced by a sequence of three-dimensional (3D) rotated quadratic cones $z^2 \le xy$; such a representation, however, is not unique. Observe that each side of inequalities (14b) contains 2^ρ factors; this allows one to construct a lifted representation for (14b) via $2^\rho - 1$ 3D rotated quadratic cones using the "tower of variables" technique [3]:

$$w^2 \le v_{\rho-1,1} v_{\rho-1,2} \tag{15a}$$
$$v_{l,i}^2 \le v_{l-1,2i-1}\, v_{l-1,2i}, \quad i = 1, \dots, 2^{\rho-l}, \quad l = 2, \dots, \rho-1, \tag{15b}$$
$$v_{1,i}^2 \le u^2, \quad i = 1, \dots, \lfloor s/2 \rfloor, \tag{15c}$$
$$v_{1,i}^2 \le ut, \quad i = \lfloor s/2 \rfloor + 1, \dots, \lceil s/2 \rceil, \tag{15d}$$
$$v_{1,i}^2 \le t^2, \quad i = \lceil s/2 \rceil + 1, \dots, \lfloor r/2 \rfloor, \tag{15e}$$
$$v_{1,i}^2 \le tw, \quad i = \lfloor r/2 \rfloor + 1, \dots, \lceil r/2 \rceil, \tag{15f}$$
$$v_{1,i}^2 \le w^2, \quad i = \lceil r/2 \rceil + 1, \dots, \lfloor R/2 \rfloor, \tag{15g}$$
$$w, v_{l,i}, u, t \ge 0,$$

where subscripts j are suppressed for brevity. The set of inequalities (15) can be visualized as a binary tree whose nodes represent the variables in (15). Each inequality in (15) can then be viewed as a subgraph with two arcs that connect the "parent" node (the variable at the left-hand side of the inequality) to the two "child" nodes (the variables at the right-hand side of the same inequality). Given this binary structure, the set of second order cones in (15) can be regarded as partitioned into ρ levels indexed by l, where the variable w in (15a) constitutes the root node of the tree, and belongs to ρ-level, while variables u, t, w in (15d)–(15g) represent the leaf nodes, or 0-level nodes of the tree.

In [10] it has been shown that among the $2^\rho - 1$ inequalities (15) there are only $O(\rho) = O(\log_2 r)$ non-degenerate second order cones, while the rest reduce to linear inequalities that can be omitted. The following bounds on the number of non-degenerate quadratic cones in (15) follow directly from the arguments in [10]:

Proposition 4 ([10]) *When p is a positive rational number, $p = r/s$, such that $r > s$ and the greatest common divisor of r and s is 1, a p-order cone in the positive orthant of $\mathbb{R}^{n+1}$ can equivalently be represented by C_p three-dimensional quadratic cones, where C_p satisfies*

$$n\rho \le C_p \le n(2\rho - 1), \quad \rho = \lceil \log_2 r \rceil. \tag{16}$$

It it easy to see that the order in which the variables u, t, and w are assigned to the leaf nodes in the binary tree (15) can significantly affect the number of non-degenerate quadratic cones needed to represent a rational-order p-cone in $\mathbb{R}^{n+1}$. As an illustration, consider the case $p = 3$; direct application of (15) yields $\rho = 2$, $R = 4$, and a representation of $p = 3$ cone (13) that involves $3n$ 3D rotated quadratic cones:

$$t \ge u_1 + \ldots + u_n; \quad w_j^2 \le v_{1j} v_{2j}, \quad v_{1j}^2 \le u_j t, \quad v_{2j}^2 \le t w_j, \quad j = 1, \dots, n. \tag{17}$$

On the other hand, it is easy to verify that reordering the leaf nodes inequalities (15c)–(15g) allows for reducing the number of 3D quadratic cones necessary to represent a $p = 3$ cone in $\mathbb{R}_+^{n+1}$ to $2n$:

$$t \ge u_1 + \ldots + u_n; \quad w_j^2 \le t v_j, \quad v_j^2 \le u_j w_j, \quad j = 1, \dots, n. \tag{18}$$

Observe that the number of second order cones in representations (17) and (18) correspond to the upper and lower bounds in (16), respectively.

Since a reduction in the number of second order cone inequalities in (15) leads to a reduction in the number of quadratic cones representing a rational-order p-cone (13) by the order of dimensionality n of the p-cone, it is of interest to devise an "economical" second order cone representation of rational-order cones.

3.2 An "economical" representation of rational-order p-cone via second order cones

Below we demonstrate that the lower bound on C_p in (16) is achievable for any rational $p \ge 1$. To this end, consider the following convex pointed cone in $\mathbb{R}_+^4$:

$$\mathcal{P} = \left\{ \mathbf{y} \in \mathbb{R}_+^4 \;\middle|\; y_0^{k_0} - y_1^{k_1} y_2^{k_2} y_3^{k_3} \le 0 \right\}, \tag{19}$$

that satisfies the next four properties:
(P1) $k_0, k_1, k_2, k_3 \in \mathbb{Z}_+$;
(P2) $k_0 = k_1 + k_2 + k_3$;
(P3) $k_1 + k_2 + k_3 = 2^q$ for some integer $q \geq 1$;
(P4) exactly two numbers among k_1, k_2, and k_3 are odd.

Proposition 5 *Cone $\mathcal{P}$ (19) that satisfies (P1)–(P4) can be represented as an intersection of at most q three-dimensional cones of the form $\{ \mathbf{x} \in \mathbb{R}^3_+ \mid x_3^2 \leq x_1 x_2 \}$.*

Proof: The process of building such a representation of $\mathcal{P}$ is based on successive lifting of $\mathcal{P}$ into spaces of dimensions greater than previous by 1, in such a way that the degree of the polynomial in (19) is reduced in half each time. First, assume that $k_1, k_3, k_3 > 0$ are all different, and $q \geq 2$. Without loss of generality, let k_1, k_2 be odd and such that $k_2 > k_1$, and consider the following set in $\mathbb{R}^5_+$:

$$\begin{gathered} \mathcal{P}^* = \{ \mathbf{y} \in \mathbb{R}^5_+ \mid y_0^{\nu_0} - y_4^{\nu_4} y_2^{\nu_2} y_3^{\nu_3} \leq 0, \;\; y_4^2 \leq y_1 y_2 \}, \\ \text{where} \quad \nu_0 = k_0/2, \quad \nu_2 = (k_2 - k_1)/2, \quad \nu_4 = k_1, \quad \nu_3 = k_3/2. \end{gathered} \tag{20}$$

It is easy to see that any $(y_0, \ldots, y_3) \in \mathcal{P}$ can be extended to $(y_0, \ldots, y_4) \in \mathcal{P}^*$, and any $(y_0, \ldots, y_4) \in \mathcal{P}^*$ is such that $(y_0, \ldots, y_3) \in \mathcal{P}$. As k_1 and k_2 are odd and positive integers by assumption, due to (P4) k_3 is even, whence ν_3 is a positive integer. The above assumption also implies that $k_2 - k_1$ is even, meaning that ν_2 is a positive integer. Similarly, ν_0 is integer and $\nu_0 = 2^{q-1}$. Also, observe that $\nu_1 + \nu_2 + \nu_3 = (k_1 + k_2 + k_3)/2 = k_0/2 = \nu_0$. So, the first cone in (20) satisfies properties (P1)–(P3). Next, observe that $\nu_4 = k_1$ is odd, thus out of two integers ν_2, ν_3 exactly one should be odd for $\nu_2 + \nu_3 + \nu_4 = 2^{q-1}$ to hold. Thus, condition (P4) holds as well.

Note that if in our assumption $k_1 = k_2$, then $\nu_2 = 0$ in (20), but all conditions still hold. Consider the case when $q \geq 2$ and one of k_1, k_2, k_3 is zero, assume it is k_3. Then k_1, k_2 should be odd by (P4). Performing the same transformation, we obtain

$$\mathcal{P}^{**} = \{ \mathbf{y} \in \mathbb{R}^5_+ \mid y_0^{\nu_0} - y_4^{\nu_4} y_2^{\nu_2} \leq 0, \; y_4^2 \leq y_1 y_2 \}, \; \nu_0 = k_0/2, \; \nu_2 = (k_2 - k_1)/2, \; \nu_4 = k_1. \tag{21}$$

The first cone of $\mathcal{P}^{**}$ still has properties (P1)–(P4), and $(y_0, \ldots, y_3) \in \mathcal{P}$ can be extended to $(y_0, \ldots, y_4) \in \mathcal{P}^{**}$, and any $(y_0, \ldots, y_4) \in \mathcal{P}^{**}$ is such that $(y_0, \ldots, y_3) \in \mathcal{P}$.

If $q = 1$, then one of k_1, k_2, k_3 is zero, and two others are necessarily equal to 1. In this case $\mathcal{P}$ is already a quadratic cone. Thus, the above lifting transformation can be carried out no more than $q - 1$ times, and the conic set $\mathcal{P}$ (19) can be represented by at most q quadratic cones using at most $q - 1$ new variables. ■

With the help of Proposition 5 we can now establish the following result on second order cone representation of rational-order p-cones:

Theorem 3 *Let $p > 1$ be a positive rational number, $p = r/s$, where the greatest common divisor of r and s is 1. Then a p-order cone in the positive orthant of $\mathbb{R}^{n+1}$ can equivalently be represented by $n\lceil \log_2 r \rceil$ three-dimensional rotated quadratic cones.*

Proof: In accordance to (13)–(14b), the problem of representing a (r/s)-cone in $\mathbb{R}^{n+1}_+$ via second order cones can be reduced to finding a second order cone representation of n sets of the form

$$\mathcal{Q} = \left\{ \mathbf{y} \in \mathbb{R}^3_+ \;\middle|\; y_3^R - y_1^s y_2^{r-s} y_3^{R-r} \leq 0 \right\}, \tag{22}$$

where $R = 2^\rho$, $\rho = \lceil \log_2 r \rceil$. Observe that cone $\mathcal{Q}$ is equivalent to intersection of cone $\mathcal{P}$ (19), where $k_1 = s$, $k_2 = r - s$, $k_3 = R - r$, with a hyperplane $y_0 = y_3$. Indeed, properties (P1)–(P3) are obvious, and (P4) holds since if r and s do not have common divisor greater than 1, neither do $r - s$ and s, whereby $r - s$ and s cannot be both even.

Note that an iteration of the lifting procedure described in Proposition 5 corresponds to a specific order in which the variables at some level of the binary tree are arranged. For example, the first iteration of lifting corresponds to arranging the 0-level variables $\{w, t, u\} = \{y_1, y_2, y_3\}$ in pairs corresponding to second order cone constraints, such that y_1 and y_2 make k_1 pairs, or $y_4^2 \leq y_1 y_2$ non-degenerate cones; the remaining $k_2 - k_1$ variables y_2 form $(k_2 - k_1)/2$ pairs, or degenerate cones ${y_4'}^2 \leq y_2^2$, and k_3 variables y_3 form $k_3/2$ pairs, or degenerate cones ${y_4''}^2 \leq y_3^3$, assuming that $k_1 < k_2$ are odd. Obviously, the degenerate cones can simply be disregarded.

Hence, by Proposition 5, $\mathcal{Q}$ admits representation by at most $\rho = \lceil \log_2 r \rceil$ second order cones; combining this with Proposition 4, one obtains that each of n sets of the form $\mathcal{Q}$ admits representation using exactly $\rho = \lceil \log_2 r \rceil$ second order cones. ■

It is well known that second order cone sets admit an equivalent semidefinite representation in the form of linear matrix inequalities (LMIs). In general, p-order cones are not LMI-representable in the space of original variables (see an example for $p = 4$ cone in [7, 8]), but admit lifted LMI representations.

Corollary 1 *Conic set* $\mathcal{Q}$ (22) *admits a lifted representation in the form of LMI*

$$\mathcal{Q}^* = \left\{ \mathbf{y} \in \mathbb{R}_+^{\rho+2} \;\middle|\; \sum_{i=1}^{\rho+2} \mathbf{A}_i y_i \succeq \mathbf{0} \right\},$$

where $\mathbf{A}_i \in \mathbb{R}^{2\rho \times 2\rho}$ *are symmetric matrices, in the sense that the projection of* $\mathcal{Q}^*$ *onto the space of variables* (y_1, y_2, y_3) *coincides with* $\mathcal{Q}$.

4 Computational study

In this section we report computational results on using the p-norm discrimination model (5)–(6) for linear separation of sets. In particular, we employ the presented above "economical" SOCP reformulation approach to solving pOCP problem (6) in the case when p is rational, and compare it with the polyhedral approximation technique of [10].

In our computational experiments we used three data sets from UCI Machine Learning Repository. The first data set is Wisconsin Breast Cancer data set with a total of 683 instances and 9 attributes. It contains 444 instances with benign diagnosis and 239 instances with malignant diagnosis. The second data set, Cleveland Heart Disease data set, contains 281 instances with 13 attributes, of them 125 instances correspond to positive diagnosis and 156 instances correspond to negative diagnosis. Finally, the Pima Indians Diabetes data set reports 768 instances with 8 attributes, including 266 instances of positive diagnosis and 502 instances of negative diagnosis. Both the Wisconsin Breast Cancer and Cleveland Heart Disease data sets (in their then-up-to-date versions) were used in [4].

For each data set, training and testing was performed by randomly selecting 100 training sets with equal number of points of both types, and testing the obtained separator on the data not included in the training set. For computational purposes, the data in training data sets was normalized and scaled by a factor of 10^4; the same transformation was then applied to testing data. After the training and testing procedures were performed, the average misclassification error on testing set was computed. It is important to comment on selection of parameter p in (6): as a general rule that follows from our numerical experiments and is consistent with the motivation presented in Section 2, smaller values of p (around $p = 2$) are beneficial for well-separable data sets with smaller misclassification errors, whereas larger values of $p \geq 3$ allow for reducing large misclassification errors in linear separation. With this in mind, a particular value of p can be selected during the training procedure.

Table 1 reports the average out-of-sample misclassification error for each data set, together with the respective "best" value of p at which this error was obtained. It also includes results for the cases of $p = 1$, which corresponds to minimization of the average of misclassifications due to [4], $p = \infty$, corresponding to minimization of the largest misclassification errors, and SVM model (11). Figures 1, 2, and 3 illustrate the behavior of the misclassification error in the described data sets with respect to the value of parameter p in (5)–(6), which was varied in the range of 1.0 to 4.0 with a 0.1 step. As it follows from Table 1 and Figures 1–3, the p-norm separation model (5)–(6) with $p > 1$ allows for an improved classification accuracy as compared to the cases of $p = 1$ proposed in [4], the SVM model (11), and the worst-error approach of $p = \infty$.

In addition to classification capabilities of the p-norm linear separation model (5)–(6), its computational properties were investigated. In particular, for all the data sets described above we compared the running times of the cutting plane procedure for polyhedral approximations of problem (6) due to [10], denoted as LP/CP, and the "economical" SOCP reformulation of (6), along with the corresponding results for SVM model (11) and $p = \infty$ case. All models were coded in C++ and CPLEX 12.2 solver was used to solve the resulting LP, SOCP, and QP problems. A dual-core 3GHz CPU computer with 2GB of RAM was used to run the computations. Figure 4 illustrates corresponding running times on the example of the Wisconsin Breast Cancer data set, along with the values of the parameter $\rho = \lceil \log_2 r \rceil$, where $p = r/s$, which is proportional to the number of second order cones in the SOCP reformulation of rational-order p-cone programming problem (6). From Figure 4 it follows that the solution times for SOCP reformulation of a rational-order p-cone programming model (6) are highly correlated to the number of second order cones in the reformulated problem. On the other hand, solution times of a polyhedral approximation of (6) solved with a cutting plane method (LP/CP) exhibit relatively little dependence on the value of the parameter p, and are competitive with the running times of the SVM model. Computational performance of the considered models on other data sets is very similar to that presented in Figure 4.

Table 1: Classification results for different data sets: the lowest average misclassification error, the corresponding value of p, and misclassification error for the cases of $p = 1$, $p = \infty$, and SVM model (11).

Dataset	Error	Best p	$p = 1$	SVM	$p = \infty$
Wisconsin Breast Cancer Dataset	3.95%	1.8	4.11%	4.03%	4.21%
Cleveland Heart Disease Dataset	18.7%	3.8	19.5%	18.98%	19.11%
Pima Indians Diabetes Dataset	31.82%	3.4	35.29%	34.02%	33.51%

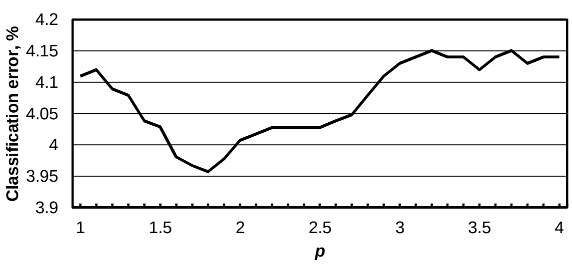

Figure 1: Misclassification error as a function of p for Wisconsin Breast Cancer data set.

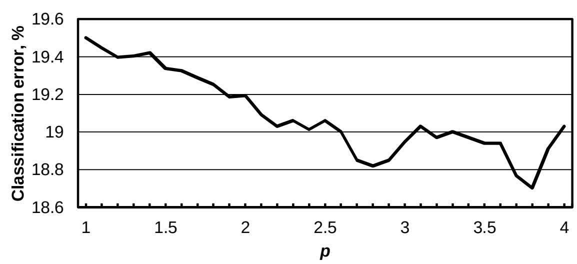

Figure 2: Misclassification error as a function of p for Cleveland Heart Disease data set.

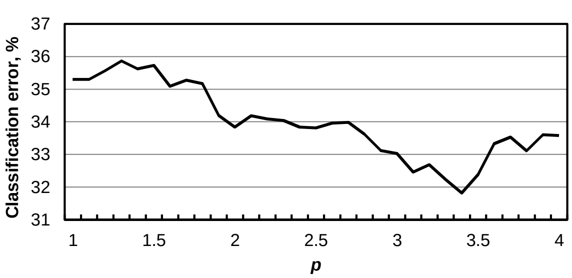

Figure 3: Misclassification error as a function of p for Pima Indians Diabetes data set.

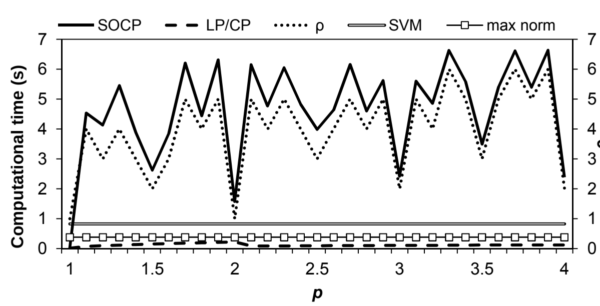

Figure 4: Average running time for instances of Wisconsin Breast Cancer data set.

Acknowledgments

The authors would like to acknowledge partial support of the U. S. Air Force Office of Scientific Research and National Science Foundation.

References

[1] Alizadeh, F. and Goldfarb, D. (2003) "Second-order cone programming," *Mathematical Programming*, **95** (1), 3–51.

[2] Andersen, E. D., Roos, C., and Terlaky, T. (2003) "On implementing a primal-dual interior-point method for conic quadratic optimization," *Mathematical Programming*, **95** (2), 249–277.

[3] Ben-Tal, A. and Nemirovski, A. (2001) "On polyhedral approximations of the second-order cone," *Mathematics of Operations Research*, **26** (2), 193–205.

[4] Bennett, K. P. and Mangasarian, O. L. (1992) "Robust linear programming separation of two linearly inseparable sets," *Optimization Methods and Software*, **1** (1), 23–34.

[5] Carrizosa, E. and Morales, D. R. (2013) "Supervised classification and mathematical optimization," *Computers & Operations Research*, **40**, 150–165.

[6] Glineur, F. and Terlaky, T. (2004) "Conic Formulation for l_p-Norm Optimization," *Journal of Optimization Theory and Applications*, **122** (2), 285–307.

[7] Helton, W. and Nie, J. (2010) "Semidefinite representation of convex sets," *Mathematical Programming*, **122** (1), 21–64.

[8] Helton, W. and Vinnikov, V. (2007) "Linear matrix inequality representation of sets," *Communications in Pure and Applied Mathematics*, **60** (5), 654–674.

[9] Khemchandani, R., Jayadeva, and Chandra, S. (2009) "Knowledge based proximal support vector machines," *European Journal of Operational Research*, **195**, 914–923.

[10] Krokhmal, P. and Soberanis, P. (2010) "Risk optimization with p-order conic constraints: A linear programming approach," *European Journal of Operational Research*, **301** (3), 653–671.

[11] Nesterov, Y. E. (2012) "Towards non-symmetric conic optimization," *Optimization Methods & Software*, **27** (4–5), 893–917.

[12] Nesterov, Y. E. and Nemirovski, A. (1994) *Interior Point Polynomial Algorithms in Convex Programming*, volume 13 of *Studies in Applied Mathematics*, SIAM, Philadelphia, PA.

[13] Sturm, J. F. (1998) "Using SeDuMi 1.0x, a MATLAB toolbox for optimization over symmetric cones," *Manuscript*.

[14] Trafalis, T. and Gilbert, R. (2006) "Robust classification and regression using support vector machines," *European Journal of Operational Research*, **173**, 893–909.

[15] Xue, G. and Ye, Y. (2000) "An efficient algorithm for minimizing a sum of p-norms," *SIAM Journal on Optimization*, **10** (2), 551–579.