CCIT Report #533

Journal of Machine Learning Research X (2005) X-X Submitted 1/2005; Published X/XX

May 2005

Feature Selection by Global Minimization of a Generalization Bound

Dori Peleg DORIP@TX.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Haifa 32000, Israel

Ron Meir RMEIR@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Haifa 32000, Israel

Editor: Unknown

Abstract

A feature selection algorithm is presented based on the global minimization of a data-dependent generalization error bound. Feature selection and scaling algorithms often lead to non-convex optimization problems, which in many previous approaches were addressed through gradient descent procedures, which can only guarantee convergence to a local minimum. We propose an alternative approach, whereby the global solution of the non-convex optimization problem is derived by an equivalent convex conic optimization problem. Highly competitive numerical results on both artificial and real-world data sets are reported. The relation of the algorithm to the support vector machine algorithm is also discussed.

Keywords: Feature Selection, Dimensionality Reduction, Classification, Generalization Error Bounds, Statistical Learning Theory.

1. Introduction

This paper presents a new approach to feature selection for classification where the goal is to learn a decision rule from a training set of pairs $S_n = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{n}$, where $x^{(i)} \in \mathbb{R}^d$ are input patterns and $y^{(i)} \in \{-1, 1\}$ are the corresponding labels. The goal of a classification algorithm is to find a separating function $f(\cdot)$, based on the training set, which will generalize well, i.e. classify new patterns with as few errors as possible. Feature selection schemes often utilize, either explicitly or implicitly, scaling variables, $\{\sigma_j\}_{j=1}^{d}$, which multiply each feature. The aim of such schemes is to optimize an objective function over $\sigma \in \mathbb{R}^d$.

Feature selection can be viewed as the special case $\sigma_j \in \{0, 1\}$, $j = 1, \ldots, d$, where a feature j is removed if $\sigma_j = 0$. The more general case of feature *scaling* is considered here, namely $\sigma_j \geq 0$, $j = 1, \ldots, d$. Clearly feature selection is a special case of feature scaling.

The overwhelming majority of feature selection algorithms in the literature, separate the feature selection and classification tasks, while solving either a combinatorial or a non-convex optimization problem (Grandvalet and Canu, 2003), (Weston et al., 2000), (Weston et al., 2003), (Perkins et al., 2003), (Guyon et al., 2002), (Bradley and Mangasarian, 1998). Additionally, an overview of feature selection algorithms is available in (Guyon and Elisseeff, 2003). In either case there is no guarantee of efficiently locating the global optimum. This is particularly problematic in large scale classification tasks which may initially contain several thousand features. Moreover, the objective of many feature selection algorithms is not related to the Generalization Error (GE). Even for global solutions of such algorithms there is no theoretical guarantee of proximity to the minimum of the GE.

To overcome the above shortcomings we propose a feature selection algorithm based on the Global Minimization of an Error Bound (GMEB). This approach consists of simultaneously finding the optimal *linear* classifier and the optimal scaling factors of each feature by minimizing a GE bound. As in previous feature selection algorithms, a non-convex optimization problem must be solved. A novelty of this paper is the use of the *equivalent* optimization problems concept (Boyd and Vandenberghe, 2004, pp. 130-136), whereby a global optimum is guaranteed in polynomial time.

The development of the GMEB algorithm begins with the design of a GE bound for feature selection. This is followed by formulating an optimization problem which minimizes this bound. Invariably, the resulting problem is non-convex. To avoid the drawbacks of solving non-convex optimization problems, an equivalent convex optimization problem is formulated whereby the *exact* global optimum of the non-convex problem can be computed. Next the convex optimization task is reduced to a rotated conic quadratic programming problem for which efficient solvers are available.

Additionally, the GMEB algorithm can function as a linear classifier for classification problems in which feature selection is not required. A link to the standard Support Vector Machine (SVM) is established in the absence of feature selection. Comparative numerical results on both artificial and real-world datsets are reported.

2. Optimization background and notation

The notation and definitions were taken from (Boyd and Vandenberghe, 2004). All vectors are column vectors unless transposed. The nonnegative real numbers are termed $\mathbb{R}_+$. The domain of function f is denoted by $\mathbf{dom}\, f$. Vectors with all components zero or one are denoted $\mathbf{0}, \mathbf{1}$ respectively. Componentwise inequality between vectors x and y are denoted by $x \preceq y$. The significance of the matrix inequality, $A \succeq B$, between two symmetric matrices A and B is that the eigenvalues of the matrix $A - B$ are nonnegative. The relative interior of set C is denoted by $\mathbf{relint}\, C$. Finally, $\mathbf{diag}\,(x)$ is a diagonal matrix with diagonal entries $x_1, \ldots, x_n$.

2.1 Definitions

Definition 1 (LMI) *The condition*

$$A(x) = x_1 A_1 + \ldots + x_n A_n \preceq B$$

where $B, A_i \in S^m$, i.e. symmetric matrices of dimensions $[m \times m]$, is called a linear matrix inequality (LMI) in $x \in \mathbb{R}^n$.

Definition 2 (Epigraph) *The epigraph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as*

$$\mathbf{epi}\, f = \{(x,t)|x \in \mathbf{dom}\, f, f(x) \le t\}$$

which is a subset of $\mathbb{R}^{n+1}$.

Definition 3 (Quasiconvex function) *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called quasiconvex if its domain and all its sublevel sets*

$$S_\alpha = \{x \in \mathbf{dom}\, f | f(x) \le \alpha\},$$

for $\alpha \in \mathbb{R}$, are convex.

For example the function $f(x) = \sqrt{|x|}$ is quasiconvex but is not convex. A function is *quasiconcave* if $-f$ is quasiconvex.

Definition 4 (Perspective of a function) *If $f : \mathbb{R}^n \to \mathbb{R}$, then the perspective of f is the function $g : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by*

$$g(x,t) = tf(x/t),$$

with domain

$$\mathbf{dom}\, g = \{(x,t)|x/t \in \mathbf{dom}\, f, t > 0\}.$$

The perspective operation preserves convexity: If f is a convex function, then so is its perspective function g. For proof see (Boyd and Vandenberghe, 2004, p.89).

2.2 Generic optimization problems

The *standard* optimization problem is one of the form

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \le 0 & i = 1, \ldots, m \\ & h_j(x) = 0 & j = 1, \ldots, p, \end{array} \tag{1}$$

with variable $x \in \mathbb{R}^n$. The set of points for which the objective and all the constraint functions are defined,

$$\mathcal{D} = (\cap_{i=0}^m \mathbf{dom}\, f_i) \cap (\cap_{j=1}^p \mathbf{dom}\, h_j)$$

is called the *domain* of the optimization problem (1). A point $x \in \mathcal{D}$ is *feasible* if it satisfies the constraints $f_i(x) \le 0$, $i = 1, \ldots, m$ and $h_j(x) = 0$, $j = 1, \ldots, p$.

A *convex optimization problem* is a problem of the form (1) in which the functions $f_i(x) : \mathbb{R}^n \to \mathbb{R}$, $i = 0, \ldots, m$, are convex and the functions $h_j(x) : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \ldots, p$, are affine (Boyd and Vandenberghe, 2004, pp. 136-137).

A *quasiconvex optimization problem* is a problem of form (1) in which $f_0(x)$ is quasiconvex, the functions $f_i(x) : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, are convex and the functions $h_j(x) : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \ldots, p$, are affine.

A *semidefinite program* (SDP) has the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & x_1 F_1 + \ldots + x_n F_n + G \preceq 0 \\ & Ax = b, \end{array} \tag{2}$$

where $G, F_1, \ldots, F_n \in S^k$, and $x \in \mathbb{R}^n, b \in \mathbb{R}^b, A \in \mathbb{R}^{m \times n}$. The inequality is a LMI.

The *epigraph form* of the standard problem (1) is the problem

$$\begin{array}{lll} \text{minimize} & t & \\ \text{subject to} & f_0(x) - t \leq 0 & \\ & f_i(x) \leq 0 & i = 1, \ldots, m \\ & h_j(x) = 0 & j = 1, \ldots, p, \end{array} \tag{3}$$

with variables $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$. The pair (x, t) is optimal for (3) if and only if x is optimal for (1) and $t = f_0(x)$.

2.3 Duality

Consider the optimization problem (1), under the assumption that its domain $\mathcal{D}$ is nonempty. The *Langrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with the problem (1) is defined as

$$L(x; \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x),$$

with $\mathbf{dom}\, L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$. The variable λ_i is termed the *Lagrange multiplier* associated with the ith inequality constraint $f_i(x) \leq 0$; similarly the variable ν_j is termed the Lagrange multiplier associated with the jth equality constraint $h_j(x) = 0$. The vectors λ and ν are called the *dual variables* associated with problem (1).

The *dual function* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as the minimum value of the Lagrangian over $x \in \mathbb{R}^n$: for $\lambda \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^p$,

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu).$$

When the Lagrangian is unbounded below in x, the dual function takes on the value $-\infty$.

The *dual problem* associated with problem (1) is

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq \mathbf{0}. \end{array} \tag{4}$$

In the context of duality, the original problem (1) is termed the *primal problem.* The optimal values of the primal problem (1) and the dual problem (4) are $p^\star, d^\star$ respectively. If $d^\star = p^\star$ then *strong duality* holds.

Theorem 5 (Slater (Boyd and Vandenberghe, 2004)) *Consider primal problem* (1)*. We assume without loss of generality that the first k inequality constraints are affine. If*

(1) *is a convex optimization problem, then strong duality holds if there exists* $x \in \mathbf{relint}\,\mathcal{D}$ *with:*

1. $f_i(x) \leq 0, \quad i = 1, \ldots, k.$
2. $f_i(x) < 0, \quad i = k+1, \ldots, m.$
3. $h_j(x) = 0, \quad j = 1, \ldots, p.$

An application of Theorem 5 is given in Lemma 18.

Lemma 6 (Schur's complement) *Consider a matrix* $X \in S^n$ *partitioned as*

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where $A \in S^k$. *If* $\det A \neq 0$, *the matrix*

$$S = C - B^T A^{-1} B$$

is called the Schur complement of A *in* X. *If* $A \succ 0$, *then* $X \succeq 0$ *if and only if* $S \succeq 0$.

3. The Generalization Error Bounds

We establish GE bounds which are used to motivate an effective algorithm for feature scaling. Consider a sample $S_n = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$, $x^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, $y^{(i)} \in \mathcal{Y}$, where $(x^{(i)}, y^{(i)})$ are generated independently from some distribution P. A set of nonnegative variables $\sigma = (\sigma_1, \ldots, \sigma_d)^T$ is introduced to allow the additional freedom of feature scaling.

For a soft classifier f, the $0-1$ loss is the probability of error given by $\mathbf{P}\left(yf(x) \leq 0\right) = \mathbf{E}I\left(yf(x) \leq 0\right)$, where $I(\cdot)$ is the indicator function.

Definition 7 *The margin cost function* $\phi_\gamma : \mathbb{R} \to \mathbb{R}_+$ *is defined as* $\phi_\gamma(z) = 1 - z/\gamma$ *if* $z \leq \gamma$, *and zero otherwise. Note that* $I\left(yf(x) \leq 0\right) \leq \phi_\gamma(yf(x))$.

Consider a classifier f for which the input features have been rescaled, namely $f(\Sigma x)$ is used instead of $f(x)$. Let $\mathcal{F}$ be some class of functions and let $\hat{\mathbf{E}}_n$ denote the empirical mean. Using standard GE bounds, one can establish that for any choice of σ, with probability at least $1 - \delta$, for any $f \in \mathcal{F}$

$$\mathbf{P}\left(yf(\Sigma x) \leq 0\right) \leq \hat{\mathbf{E}}_n \phi_\gamma\left(yf(\Sigma x)\right) + \Omega(f, \delta, \sigma), \tag{5}$$

for some appropriate complexity measure Ω depending on the bounding technique.

The scaling variables σ transform the linear classifiers from $f(x) = w^T x + b$ to $f(x) = w^T \Sigma x + b$, where $\Sigma = \mathbf{diag}(\sigma)$. It may seem at first glance that these classifiers are essentially the same since w can be redefined as Σw. However, the role of σ is to offer an extra degree of freedom to scale the features *independently* of w, in a way which can be exploited by an optimization algorithm as we will show below.

Unfortunately, (5) cannot be used directly when attempting to select optimal values of the variables σ because the bound is not *uniform* in σ. In particular, we need a result which holds with probability $1 - \delta$ for *every* choice of σ.

Definition 8 *The indices of training patterns with labels $\{-1, 1\}$ are denoted by I_-, I_+ respectively. The cardinalities of the sets I_-, I_+ are n_-, n_+ respectively. The empirical mean of the second order moment of the* j*th feature over the training patterns belonging to indices I_-, I_+ are*

$$v_j^- = \frac{1}{n_-} \sum_{i \in I_-} \left(x_j^{(i)} \right)^2 , \quad v_j^+ = \frac{1}{n_+} \sum_{i \in I_+} \left(x_j^{(i)} \right)^2$$

respectively.

Theorem 9 *Fix $B, r, \gamma > 0$, and suppose that $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ are chosen independently at random according to some probability distribution P on $\mathcal{X} \times \{\pm 1\}$, where $\|x\| \leq r$ for $x \in \mathcal{X}$. Define the class of functions $\mathcal{F}$*

$$\mathcal{F} = \left\{ f : f(x) = w^T \Sigma x + b, \ \|w\| \leq B, \ |b| \leq r, \ \sigma \succeq \mathbf{0} \right\} .$$

Let σ_0 be an arbitrary positive number, and set $\grave{\sigma}_j = 2 \max(\sigma_j, \sigma_0)$. Then with probability at least $1 - \delta$, for every function $f \in \mathcal{F}$

$$P\left(yf(x) \leq 0 \right) \leq \hat{\mathbf{E}}_n \phi_\gamma \left(yf(x) \right) + \frac{2B}{\gamma} \left(\frac{\sqrt{n_+}}{n} \sqrt{\sum_{j=1}^d v_j^+ \grave{\sigma}_j^2} + \frac{\sqrt{n_-}}{n} \sqrt{\sum_{j=1}^d v_j^- \grave{\sigma}_j^2} \right) + \frac{\Lambda(\sigma, \gamma, \delta)}{\sqrt{n}} ,$$

where $K(\sigma) = (B\|\grave{\sigma}\| + 1)r$,

$$\Lambda(\sigma, \gamma, \delta) = \frac{2r}{\gamma} + K(\sigma) \sqrt{2 \sum_{j=1}^d \ln \log_2 \frac{\grave{\sigma}_j}{\sigma_0}} + K(\sigma) \left(\frac{2}{\gamma} + 1 \right) \sqrt{2 \ln \frac{2}{\delta}} .$$

The proof appears in appendix A.

Note that v_j^- and v_j^+ are related to the variance of the jth feature of the respective classes. Thus the bound of Theorem 9 is related to the between class variance. In previous work, for example (Meir and Zhang, 2003), the training points were treated without regard to their labels. In this approach the bound would have been related to the overall variance of each feature.

Remark 10 *We note that Theorem 9 assumed that the ℓ_2 norm is used to define the constraint on w. In fact, the techniques developed in (Meir and Zhang, 2003) allow us to derive bounds which hold for* any *convex constraint. In particular, one can use ℓ_p norms of the form $\|w\|_p$, and derive appropriate generalization bounds.*

In principle, we would like to minimize the r.h.s. of (6) with respect to the variables w, σ, b. However, in this work the focus is only on the *data-dependent* terms in (6), which include the empirical error term and the weighted norms of σ. Note that all other terms of (6) are of the same order of magnitude (as a function of n), but do not depend explicitly on the data. It should be commented that the extra terms appearing in the bound arise because of the assumed unboundedness of σ. Assuming σ to be bounded, e.g. $\sigma \preceq s$, as is the case in most other bounds in the literature, one may replace σ by s in all terms except the first two, thus removing the explicit dependence on σ.

The data-dependent terms of the GE bound (6) are the basis of the objective function

$$\frac{1}{n\gamma}\sum_{i=1}^{n}\phi_{\gamma}\left(y^{(i)}f(x^{(i)})\right)+\frac{4\sqrt{n_+}}{n\gamma}\sqrt{\sum_{j=1}^{d}v_j^+\sigma_j^2}+\frac{4\sqrt{n_-}}{n\gamma}\sqrt{\sum_{j=1}^{d}v_j^-\sigma_j^2}\ , \tag{6}$$

where the variables are subject to $w^Tw \le 1,\ \sigma \succeq 0$. The transition was performed by setting $B = 1$ and replacing $\grave{\sigma}$ by 2σ (assuming that $\sigma > \sigma_0$).

Due to the fact that only the sign of f determines the estimated labels, it can be multiplied by any positive factor and produce identical results. The constraint on the norm of w induces a normalization on the classifier $f(x) = w^Tx + b$, without which the classifier is not unique. However, by introducing the scale variables σ, the classifier was transformed to $f(x) = w^T\Sigma x + b$. Hence, despite the constraint on w, the classifier is again not unique. If the variable γ in (6) is set to an arbitrary positive constant then the solution is unique again. This is true because γ appears in (6) only in the expressions $\frac{b}{\gamma}, \frac{\sigma_1}{\gamma}, \ldots, \frac{\sigma_d}{\gamma}$. By setting $\gamma = 1$ the new objective function is

$$\frac{1}{n}\sum_{i=1}^{n}\phi_{1}\left(y^{(i)}f(x^{(i)})\right)+\frac{C_+\sqrt{n_+}}{n}\sqrt{\sum_{j=1}^{d}v_j^+\sigma_j^2}+\frac{C_-\sqrt{n_-}}{n}\sqrt{\sum_{j=1}^{d}v_j^-\sigma_j^2}\ , \tag{7}$$

where $C_+ = C_- = 4$. In many classification algorithms the final classifier is based on a choice of a hyperparameter. In (7) there is a tradeoff between the penalty on the training errors and the number of features and a tradeoff between the second moment of each class. However, since current bounding techniques are not sufficiently tight, the values of C_+, C_- from the bound are not appropriate for all classification problems. Therefore we propose that C_+, C_- are chosen via a Cross Validation (CV) scheme. These hyperparameters enable fine-tuning a general classifier to a specific classification task.

Next, a generalization error bound for linear classification without feature scaling is presented. The following bound is presented to show that the addition of the scaling variables is indispensable. The bound of Theorem 11 is different from the bound of Theorem 9.

As opposed to the case where feature scaling is allowed, here the parameter γ is not superfluous. Therefore the bound is uniform with respect to the margin parameter γ.

Theorem 11 *Let the conditions of Theorem 9 hold, except that $\sigma_j = 1$ for all j (i.e. no feature scaling is allowed). Let γ_0 be an arbitrary positive number and set $\grave{\gamma} = 2\max(\gamma, \gamma_0)$. Then with probability at least $1-\delta$ for all $f \in \mathcal{F}$, where $\mathcal{F} = \{f:\ f(x) = w^Tx + b,\ \|w\| \le B,\ |b| \le r\}$, and for all $\gamma \ge \gamma_0$,*

$$P\left(yf(x) \le 0\right) \le \hat{\mathbf{E}}_n\phi_\gamma\left(yf(x)\right) + \frac{1}{nC\grave{\gamma}} + \frac{\Lambda(\gamma,\delta)}{\sqrt{n}} \tag{8}$$

where $K = (B+1)r$, $v_j = \frac{1}{n}\sum_{i=1}^{n}\left(x_j^{(i)}\right)^2$,

$$\begin{aligned}\Lambda(\gamma,\delta) &= K\left(\sqrt{2\ln\log_2\tfrac{\grave{\gamma}}{\gamma_0}} + \sqrt{2\ln\tfrac{2}{\delta}}\right)\ ,\\ C &= \left(4B\sqrt{\textstyle\sum_{j=1}^{d}v_j} + 2r\sqrt{n} + 2K\sqrt{2n\ln\tfrac{2}{\delta}}\right)^{-1}\ .\end{aligned}$$

Note that the separation of the training examples according to their labels can be easily performed in Theorem 11 as in Theorem 9. However it would only change the value of C, which is a data-dependant constant.

Similarly to the feature scaling bound, the objective function for classification without feature scaling is,

$$\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma}\left(y^{(i)}f(x^{(i)})\right)+\frac{1}{2nC\gamma}, \tag{9}$$

subject to $\|w\| \leq 1$. In Section 5 we will derive the algorithm based on Theorem 11 and uncover its relation to the standard SVM algorithm.

4. Derivation of the GMEB algorithm

The problem of minimizing (7) can be expressed as

$$\begin{array}{ll} \text{minimize} & \frac{1}{n}\mathbf{1}^T\xi + \frac{C_+\sqrt{n_+}}{n}\sqrt{\sum_{j=1}^{d} v_j^+ \sigma_j^2} + \frac{C_-\sqrt{n_-}}{n}\sqrt{\sum_{j=1}^{d} v_j^- \sigma_j^2} \\ \text{subject to} & w^T w \leq 1 \\ & y^{(i)}(\sum_{j=1}^{d} x_j^{(i)} w_j \sigma_j + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, n \\ & \xi, \sigma \succeq 0, \end{array} \tag{10}$$

with variables $w, \sigma \in \mathbb{R}^d$, $\xi \in \mathbb{R}^n$, $b \in \mathbb{R}$.

Remark 12 *Consider a solution of problem* (10) *in which $\sigma_j^\star = 0$ for some feature j. Only the constraint $w^T w \leq 1$ affects the value of $w_j^\star$. A unique solution is established by setting $\sigma_j^\star = 0 \Rightarrow w_j^\star = 0$. If the original solution $w^\star$ satisfies the constraint $w^T w \leq 1$ then the amended solution will also satisfy the constraint and will not affect the value of the objective function.*

We begin by converting the second and third terms of the objective function (7) into constraints. The purpose of this step is to allow convexification techniques which can't be applied directly to problem (10).

Lemma 13 *Denote optimization problem I as*

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq a \\ & x \in \mathcal{X}, \end{array}$$

and optimization problem II as

$$\begin{array}{ll} \text{minimize} & f(x) + bg(x) \\ \text{subject to} & x \in \mathcal{X}, \end{array}$$

where $b > 0$. Let x_1^a and x_2^b be the global solutions of problems I and II with parameters a and b, respectively. Then for any value of parameter b in problem II, for the parameter $a = g(x_2^b)$ holds $x_1^a \equiv x_2^b$.

Proof.

First we will show that given $a = g(x_2^b)$ the solution x_1^a of problem I is active. Then we will prove that in this case $x_1^a \equiv x_2^b$.

We will falsely assume that for $a = g(x_2^b)$ the solution x_1^a of problem I is inactive. Hence $g(x_1^a) < a = g(x_2^b)$. Additionally, since x_1^a is a global solution of problem I and x_2^b is a feasible point of problem I then $f(x_1^a) \leq f(x_2^b)$. Combining the inequalities with $b > 0$ we get $f(x_1^a) + bg(x_1^a) < f(x_2^b) + bg(x_2^b)$. On the other hand since x_2^b is a global solution of problem 2 then $f(x_1^a) + bg(x_1^a) \geq f(x_2^b) + bg(x_2^b)$. This is a contradiction.

Consequently the solution x_1^a of problem I is active and $g(x_1^a) = a = g(x_2^b)$. Denote optimization problem III as

$$\begin{array}{ll} \text{minimize} & f(x) + ba \\ \text{subject to} & g(x) = a \\ & x \in \mathcal{X}. \end{array}$$

The minimizer and minimal value of problem III are the same as those of problem II. Since the term ba is a constant, the minimizer of problem III is the minimizer of problem I. □

In accordance with the equivalence relationship of Lemma 13 we propose to solve the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T \xi \\ \text{subject to} & w^T w \leq 1 \\ & y^{(i)}(\sum_{j=1}^d x_j^{(i)} w_j \sigma_j + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, n \\ & R_+ \geq \sum_{j=1}^d v_j^+ \sigma_j^2 \\ & R_- \geq \sum_{j=1}^d v_j^- \sigma_j^2 \\ & \xi, \sigma \succeq \mathbf{0} \end{array} \tag{11}$$

with variables $w, \sigma \in \mathbb{R}^d$, $\xi \in \mathbb{R}^n$, $b \in \mathbb{R}$.

The functions $w_j \sigma_j$ in the second inequality constraints are neither convex nor concave (in fact they are quasiconcave). To make matters worse, the functions $w_j \sigma_j$ are multiplied by constants $-y^{(i)} x_j^{(i)}$ which can be either positive or negative. Consequently problem (11) is *not* a convex optimization problem. The objective of Section 4.1 is to find the global minimum of (11) in polynomial time despite its non-convexity.

4.1 Convexification

In this paper the informal definition of equivalent optimization problems is adopted from (Boyd and Vandenberghe, 2004, pp. 130–135): two optimization problems are called *equivalent* if from a solution of one, a solution of the other is found, and vice versa. Instead of detailing a complicated formal definition of general equivalence, the specific equivalence relationships utilized in this paper are either formally introduced or cited from (Boyd and Vandenberghe, 2004).

The functions $w_j \sigma_j$ in problem (11) are not convex and the signs of the multiplying constants $-y^{(i)} x_j^{(i)}$ are data dependant. The only functions that remain convex irrespective of the sign of the constants which multiply them are linear functions. Therefore the functions $w_j \sigma_j$ must be transformed into linear functions.

However, such a transformation must also maintain the convexity of the objective function and the other constraints. For this purpose the *change of variables* equivalence relationship, described in Theorem 14, was utilized.

Theorem 14 (Change of variables) *Consider optimization problem*

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \ldots, m. \end{array} \tag{12}$$

Suppose $\phi : \mathbb{R}^n \to \mathbb{R}^n$ *is one-to-one, with image covering the problem domain* $\mathcal{D}$*, i.e.,* $\phi(\mathbf{dom}\,\phi) \supseteq \mathcal{D}$ *. We define functions* $\tilde{f}_i$ *as*

$$\tilde{f}_i(z) = f_i(\phi(z)), \quad i = 0, \ldots, m$$

Now consider the problem

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(z) \\ \text{subject to} & \tilde{f}_i(z) \le 0, \quad i = 1, \ldots, m \end{array} \tag{13}$$

with variable z*. Problems* (12) *and* (13) *are said to be related by the change of variable* $x = \phi(z)$ *and are equivalent: If* x *solves the problem* (12)*, then* $z = \phi^{-1}(x)$ *solves problem*(13)*; if* z *solves problem* (13)*, then* $x = \phi(z)$ *solves problem* (12)*.*

The proof is detailed in (Boyd and Vandenberghe, 2004, p.130).

The transformation $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ was used on the variables w, σ:

$$\sigma_j = +\sqrt{\tilde{\sigma}_j}, \; w_j = \frac{\tilde{w}_j}{\sqrt{\tilde{\sigma}_j}}, \quad j = 1, \ldots, d \tag{14}$$

where $\mathbf{dom}\,\phi = \{(\tilde{\sigma}, \tilde{w}) | \tilde{\sigma} \succeq \mathbf{0}\}$. If $\tilde{\sigma}_j = 0$ then $\sigma_j = w_j = 0$ without regard to the value of $\tilde{w}_j$, in accordance with remark 12. Transformation (14) is clearly one-to-one and $\phi(\mathbf{dom}\,\phi) \supseteq \mathcal{D}$.

Theorem 15 *The problem*

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T \xi \\ \text{subject to} & y^{(i)}(\tilde{w}^T x^{(i)} + b) \ge 1 - \xi_i, \;\; i = 1, \ldots, n \\ & \sum_{j=1}^d \frac{\tilde{w}_j^2}{\tilde{\sigma}_j} \le 1 \\ & R_+ \ge (v^+)^T \tilde{\sigma} \\ & R_- \ge (v^-)^T \tilde{\sigma} \\ & \xi, \tilde{\sigma} \succeq 0 \end{array} \tag{15}$$

with the variables $\tilde{w}, \tilde{\sigma} \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n$ *is convex and equivalent to the primal non-convex problem* (11) *with transformation* (14)*.*

Note that since $\tilde{w}_j = w_j \sigma_j$, the new classifier is $f(x) = \tilde{w}^T x + b$. Therefore there is no need to use transformation (14) to obtain the desired classifier. Also one can use Schur's complement (Lemma 6) to transform the non-linear constraint into a sparse LMI constraint

$$\begin{bmatrix} \Sigma & w \\ w^T & 1 \end{bmatrix} \succeq 0.$$

Thus problem (15) can be reformulated as a SDP problem. The primal problem therefore, consists of $n + 2d + 1$ variables, $2n + d + 2$ linear inequality constraints and a LMI of $[(d+1) \times (d+1)]$ dimensions. Although the primal problem (15) is convex, it heavily relies on the number of features d which is typically the bottleneck for feature selection problems. To alleviate this dependency the dual problem is formulated.

Remark 16 *If one of the constraints $R_+ \geq (v^+)^T\tilde{\sigma}$ or $R_- \geq (v^-)^T\tilde{\sigma}$ is active and the other one inactive, then Problem 15 reduces to the l_1 SVM problem where each feature j was first divided by v_j^+ or v_j^- respectively.*

4.2 The dual optimization problem

In this section the dual optimization problem associated with problem (15) is formulated, leading to a simpler optimization problem and further insights regarding the primal problem (15).

Theorem 17 (Dual problem) *The dual optimization problem associated with problem* (15) *is*

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T\mu - \mu_1 - R_+\mu_+ - R_-\mu_- \\ \text{subject to} & \left(\sum_{i=1}^n \mu_i y^{(i)} x_j^{(i)}, 2\mu_1, (\mu_+ v_j^+ + \mu_- v_j^-)\right) \in K^r \quad , j = 1, \ldots, d \\ & \mu^T y = 0 \\ & \mathbf{0} \preceq \mu \preceq \mathbf{1} \\ & \mu_+, \mu_- \geq 0, \end{array} \tag{16}$$

where K^r is the Rotated Quadratic Cone (RQC) $K^r = \{(x, y, z) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} | x^T x \leq 2yz, y \geq 0, z \geq 0\}$. The optimization variables are $\mu \in \mathbb{R}^n, \mu_1, \mu_+, \mu_- \in \mathbb{R}$.

The proof is given in appendix B. The dual problem (16) is a RQC problem which can be solved efficiently with conic programming solvers. The number of variables is $n + 3$, there are $2n + 2$ linear inequality constraints, a single linear equality constraint and d RQC inequality constraints. In order to use the dual problem to derive the solution of the primal problem, strong duality must hold.

Lemma 18 (Strong duality) *Strong duality holds between problem* (15) *and problem* (16)*.*

Proof. The primal problem (15) is a convex optimization problem. The point

$$\tilde{w} = \mathbf{0}, \quad \tilde{\sigma} = \frac{1}{2} \min\left\{\frac{R_+}{\mathbf{1}^T v^+}, \frac{R_-}{\mathbf{1}^T v^-}\right\}\mathbf{1}, \quad b = 0, \quad \xi = \mathbf{1},$$

is in the relative interior of the problem domain which satisfies the Slater conditions in theorem 5. □

In order to extract the primal solution from the dual solution we propose the following theorem.

Theorem 19 (Extracting the primal solution from the dual solution) *Consider an equivalent problem to the dual problem* (16)*:*

$$\begin{array}{ll} \text{maximize} & f_0(\mu, \mu_1, \mu_+, \mu_-) \\ \text{subject to} & s = XY\mu \\ & \frac{s_j^2}{4\mu_1} - (\mu_+ v_j^+ + \mu_- v_j^-) \leq 0, \quad j = 1, \dots, d \\ & \mu^T y = 0 \\ & \mathbf{0} \preceq \mu \preceq \mathbf{1} \\ & \mu_+, \mu_- \geq 0, \end{array} \tag{17}$$

where the objective function is defined as

$$f_0(\mu, \mu_1, \mu_+, \mu_-) = \begin{cases} \mathbf{1}^T\mu - \mu_1 - R_+\mu_+ - R_-\mu_- & \textit{if} \quad \mu_1 \geq 0 \\ -\infty & \textit{otherwise.} \end{cases}$$

The primal solution $\{\tilde{w}, b\}$ is equal to the lagrange multipliers of the equality constraints $s = XY\mu$ and $\mu^T y = 0$ respectively.

The proof is given in appendix C.

4.3 The GMEB algorithm

The GMEB algorithm consists of the following steps:

1. In order to avoid a dependency on the mean of the features, as a preprocessing step, the features of the training patterns should be set to zero mean and the features of the test set shifted accordingly.
2. Derive the solution $\{\tilde{w}, b\}$ by doing one of the following:
   (a) Solve the primal optimization problem (15).
   (b) Solve the dual optimization problem (16) and use Theorem 19 to calculate $\{\tilde{w}, b\}$.

The final classifier is

$$f(x) = \text{sign}\left(\tilde{w}^T x + b\right) .$$

5. Relation to the standard SVM algorithm

In this section we explore the relation of the proposed GMEB algorithm to the standard SVM algorithm. In the absence of feature scaling, minimizing (9) produces the SVM classifier with a specific choice of the SVM hyperparameter C. The problem of minimizing (9) can be expressed as

$$\begin{array}{ll} \text{minimize} & \frac{1}{\gamma}\left(\mathbf{1}^T\xi + \frac{1}{2C}\right) \\ \text{subject to} & y^{(i)}(w^T x^{(i)} + b) \geq \gamma - \xi_i, \;\; i = 1, \dots, n \\ & w^T w \leq 1 \\ & \xi \succeq \mathbf{0}, \end{array} \tag{18}$$

where the classifier is sgn $\big(w^T x + b\big)$.

Problem (18) is a quasiconvex optimization problem with a convex over a linear objective function and convex constraints.

Theorem 20 *Consider problem*

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\hat{w}\|_2 + C\mathbf{1}^T\hat{\xi} \\ \text{subject to} & y^{(i)}(\hat{w}^T x^{(i)} + \hat{b}) \geq 1 - \hat{\xi}_i,\ i = 1, \ldots, n \\ & \hat{\xi} \succeq \mathbf{0}. \end{array} \tag{19}$$

Problem (19) *is equivalent to problem* (18) *with the transformation* $w = \hat{w}/\|\hat{w}\|_2$, $b = \hat{b}/\|\hat{w}\|_2$, $\xi = \hat{\xi}/\|\hat{w}\|_2$, $\gamma = 1/\|\hat{w}\|_2$.

The proof is an application of Lemma 23 of equivalence which is proved in appendix D.

Note that the classifier sgn $\left(\frac{\hat{w}^T x + \hat{b}}{\|\hat{w}\|_2}\right)$ of problem (18) is equivalent to the classifier sgn $\big(w^T x + b\big)$ of the problem (19) because it is multiplied by a positive constant[1]. Thus the only difference between the classifier of the GMEB algorithm without feature scaling to a standard SVM algorithm with the linear kernel and C from theorem 11 is that in the SVM algorithm the norm of w is squared.

6. Experiments

Several algorithms were comparatively evaluated on a number of artificial and real world two class problem datasets. The GMEB algorithm is a linear classifier and therefore it was compared only to linear classifiers.

6.1 Algorithms

The GMEB algorithm was compared to the linear SVM (standard SVM with linear kernel), the l_1 SVM classifier (Fung and Mangasarian, 2000) and the RFE algorithm (Guyon et al., 2002). The SVM algorithms were chosen for comparison because of their relationship to the GMEB algorithm. Furthermore, all the aforementioned algorithms consist of solving convex optimization problems and are among the best classification algorithms.

6.2 Experimental Methodology

The algorithms are compared by two criteria: the number of selected features and the error rates. In feature scaling algorithms, i.e. when using continuous rather than binary parameters, the weight assigned by a linear classifier to a feature j, determines whether it should be 'selected' or 'rejected'. This weight must fulfil at least one of the following two requirements:

1. Absolute measure - $|w_j| \geq \epsilon$.
2. Relative measure - $\frac{|w_j|}{\max_j\{|w_j|\}} \geq \epsilon$.

[1] The case $w = \mathbf{0}$ is a degenerate case in which the classifier uniformly produces 1 or -1.

In this paper $\epsilon = 0.01$ was used. Note that the weights of the GMEB algorithm are $\tilde{w}$ instead of w.

The definition of the error rate is intrinsically entwined with the protocol for determining the hyperparameter. Given an *a-priori* partitioning of the dataset into training and test sets, the following protocol for determining the value of R_+, R_- and defining the error rate is suggested:

1. Define a set $\mathcal{R}$ of values of the hyperparameters R_+, R_- for all datasets. The set $\mathcal{R}$ consists of a predetermined number of values. For each algorithm the cardinality $|\mathcal{R}| = 49$ was used.

2. Calculate the N-fold CV error for each value of R_+, R_- from set $\mathcal{R}$ on the *training* set. Five fold CV was used throughout all the datasets.

3. Use the classifier with the value of R_+, R_- which produced the lowest CV error to classify the test set. This is the reported error rate.

If the dataset is not partitioned *a-priori* into a training and test set, it is randomly divided into n_p contiguous training and 'test' sets. Each training set contains $n\frac{n_p-1}{n_p}$ patterns and the corresponding test set consists of $\frac{n}{n_p}$ patterns. Once the dataset is thus partitioned, the above steps $1-3$ can be implemented. The error rate and the number of selected features are then defined as the average on the n_p problems. The value $n_p = 10$ was used for all datasets, where an *a-priori* partitioning was not available.

Two types of real-world classification tasks were selected for the experiments. One required major feature selection, while in the other, feature selection was of lesser importance. The hyperparameter sets $\mathcal{R}$ used for the GMEB algorithm consisted of 7×7 linearly spaced values between 1 and 10. These ranges were determined by trial and error. However for almost all the datasets some feature selection was performed by the algorithm.

The linear and the l_1 SVM algorithms require one hyperparameter C. For the SVM algorithms the set $\mathcal{R}$ consisted of the values $\frac{\Lambda}{1-\Lambda}$ where $\Lambda = \{0.02, 0.04, \ldots, 0.98\}$, i.e. 49 linearly spaced values between 0.02 and 0.98. On the other hand, the RFE algorithm requires two hyperparameters. The first is the number of selected features and the second is the SVM hyperparameter C. The number of selected features was assigned 7 linearly spaced values from 1 to d and C was assigned $\frac{\hat{\Lambda}}{1-\hat{\Lambda}}$ where the set $\hat{\Lambda}$ is 7 linearly spaced values between 0.02 and 0.98.

An exhaustive search for the optimal value of the hyperparameters would have produced better results for all algorithms. This is especially true for the GMEB and RFE algorithms which require the selection of two hyperparameters. However in such an approach a fair comparison would have been much more complicated to quantify.

6.3 Data sets

The algorithms in Section 6.1 with the methodology of Section 6.2 were tested on two types of synthetic datasets, and eight real-world problems. The first four real-world datasets are known to require feature selection.

6.3.1 Synthetic datasets

1. The synthetic dataset is described in (Weston et al., 2000). Six features out of 202 were relevant. The probability of $y = 1$ or -1 was equal. With a probability of 0.7, the first six features were drawn from the Gaussian distribution described for case I in table 1. Otherwise they were drawn as case II. The remaining features were noise drawn as $x_i \sim \mathcal{N}(0, 20)$, for $i = 7, \ldots, 202$.

Table 1: Gaussian distributions for first six features.

case	$i \in \{1, 2, 3\}$	$i \in \{4, 5, 6\}$
I	$x_i \vert y \sim \mathcal{N}(yi, 1)$	$x_i \vert y \sim \mathcal{N}(0, 1)$
II	$x_i \vert y \sim \mathcal{N}(0, 1)$	$x_i \vert y \sim \mathcal{N}(y(i-3), 1)$

2. n_{sel} features out of d were relevant for classification. The probability of $y = 1$ or -1 was equal. Features $x_j, j = 1, \ldots, n_{sel}$ were drawn as $x|y \sim \mathcal{N}(y\mathbf{1}, \Lambda)$, with the covariance matrix $\Lambda = VDV^T$. The matrix D was diagonal with the values $1, 2, \ldots, n_{sel}$ and V a randomly generated unitary matrix. The remaining features were normally distributed white noise $\mathcal{N}(0, 1)$.

6.3.2 Real-World datasets

The number of features, the number of patterns and the partitioning into train and test sets of the real-world datasets are detailed in Table 2. The datasets were taken form the UCI repository unless stated otherwise. Dataset (1) is termed Wisconsin Diagnostic Breast Cancer 'WDBC', (2) 'Multiple Features' dataset, which was first introduced by (Perkins et al., 2003), (3) the 'Internet Advertisements' dataset, was separated into a training and test set randomly, (4) the 'Colon' dataset, taken from (Weston et al., 2000), (5) the 'BUPA' dataset, (6) the 'Pima Indians Diabetes' dataset, (7) the 'Cleveland heart disease' dataset, and (8), the 'Ionosphere' dataset.

6.4 Experimental results

This subsection includes a comparison of the performance of the GMEB, RFE, l_1 SVM and linear SVM algorithms on the synthetic and real-world datasets.

6.4.1 Synthetic data set

Table 3 provides a comparison of the GMEB algorithm with the RFE and SVM algorithms on synthetic datasets of type 1. The Bayes error is 0.4%. For further comparison see (Rakotomamonjy, 2003). Note that the number of features selected by the l_1 SVM and GMEB algorithms increases with the sample size. A possible explanation for this observation is that with only a few training patterns a small training error can be achieved by many subsets containing a small number of features, i.e. a sparse solution. The particular subset selected is essentially random, leading to a large test error.

An illustration of the dependance of the GMEB algorithm on its hyperparmeters R_+ and R_- is portrayed in Figure 1. The GMEB algorithm was tested on three synthetic datasets of type 2 with 100 features, 200 training patterns and 2,10,20 informative features respectively.

Table 2: The real-world datasets. The set of values of hyperparameter C for the linear SVM and RFE algorithm for datasets 1,5,6 had to be set to Λ and $\hat{\Lambda}$ respectively to allow convergence.

Dataset No.	Name	Features	Patterns
1	WDBC	30	569
2	Digits	649	200/1800
3	Internet Ads	1558	200/3080
4	Colon	2000	62
5	BUPA	6	345
6	Pima	8	768
7	Cleveland heart	13	297
8	Ionosphere	34	351

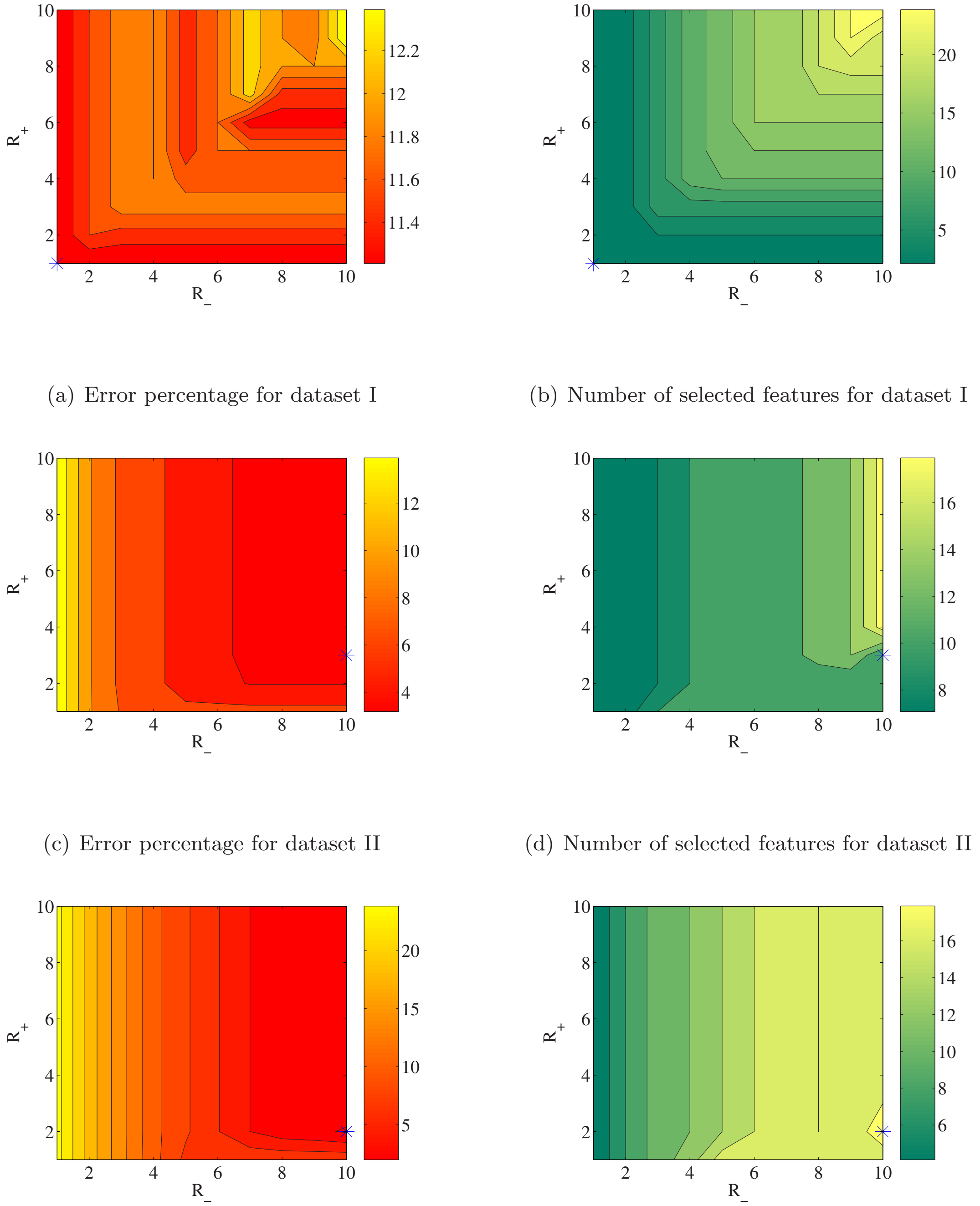
Table 3: Mean and standard deviation of the mean of test error percentage on synthetic datasets 1 given n training patterns. The number of selected features is in brackets.

n	SVM	l_1 SVM	RFE	GMEB
10	46.2 ± 1.9 (197.1±2.1)	49.6 ± 1.9 (77.7±83.8)	46.1 ± 8.6 (38.0±40.7)	**33.8 ± 14.2 (3.7±2.1)**
20	44.9 ± 2.1 (196.8±1.9)	38.5 ± 12.7 (10.7±6.1)	30.3 ± 16.4 (13.3±34.7)	**13.9 ± 7.2 (4.8±2.7)**
30	43.6 ± 1.7 (196.7±2.8)	27.4 ± 12.4 (14.5±8.7)	22.9 ± 12.3 (9.9±29.1)	**7.1 ± 5.6 (5.1±2.3)**
40	41.8 ± 1.9 (197.2±1.8)	19.2 ± 6.9 (16.2±11.1)	19.4 ± 8.5 (6.6±30.7)	**5.0 ± 3.5 (5.5±2.1)**
50	41.9 ± 1.8 (196.6±2.6)	16.0 ± 5.3 (18.4±11.3)	16.9 ± 5.6 (**1.0±0.0**)	**3.1 ± 2.7** (5.1±1.8)

The hyperparameter sets $\mathcal{R}$ used for the GMEB algorithm consisted of 10×10 linearly spaced values between 1 and 10.

The full potential of the GMEB algorithm is demonstrated on synthetic datasets of type 2. Ten datasets were generated with 1000 features, 100 training patterns, 1000 testing patterns and $n_{sel} = 2$. These datasets were far more challenging than synthetic datasets of type 1 because the number of features from which a feature selection algorithm had to find the relevant features was an order of magnitude larger. Moreover, the irrelevant features of synthetic datasets type 1 were given large variance (standard deviation 20) and the relevant features were independent, which is not indicative of real-world problems. In synthetic datasets of type 2 the standard deviation of the irrelevant features was 1 and the two relevant features were correlated.

The linear SVM algorithm achieved 30.1% ± 1.7% test error by using 922.9 ± 11.6 features. The l_1 SVM algorithm achieved 10.8% ± 2.3% test error by using 6.6 ± 9.8 features. The RFE algorithm achieved 17.8% ± 1.8% test error by using 1.0 ± 0 features. The GMEB algorithm achieved 11.3% ± 2.3% test error by using 2.8 ± 1.3 features. The Bayes error is 9.6%. From the results on the synthetic datasets it is evident that the GMEB algorithm has the potential to deal with difficult feature selection problems which contain a vast majority of irrelevant features.



(a) Error percentage for dataset I

(b) Number of selected features for dataset I

(c) Error percentage for dataset II

(d) Number of selected features for dataset II

(e) Error percentage for dataset III

(f) Number of selected features for dataset III

Figure 1: The error percentage and the number of selected features for synthetic dataset with 100 features, 200 training patterns and 2,10,20 informative features termed as dataset I, II, III respectively. The optimal value is marked by a star.

6.4.2 Real World datasets

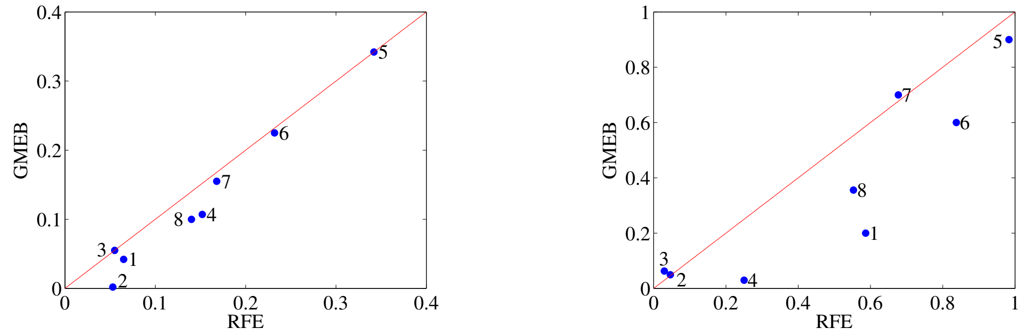
In this section the error rates and number of selected features of the GMEB and the SVM algorithms on the datasets described in Section 6.3.2 are presented.

Table 4: The performance of the algorithms on the real-world datasets (mean and standard deviation of the mean).

No.	Linear SVM	l_1 SVM	RFE	GMEB
1	5.3±0.8 (27.3±0.3)	4.9±1.1 (16.4±1.3)	6.5±0.9 (17.6±2.3)	**4.2±0.9** (**6.0±0.3**)
2	0.3 (616)	3.5 (**15**)	5.3 (30)	**0.2** (32)
3	5.3 (322)	**4.7** (**12**)	5.5 (46)	5.5 (98)
4	13.6±5.9 (1941.8±1.9)	**10.7±4.4** (**23.3±1.5**)	15.2±5.7 (500.7±89.6)	**10.7±4.4** (59.1±25.0)
5	**33.1±3.5** (6.0±0.0)	33.6±3.6 (5.9±0.1)	34.2±3.4 (5.9±0.1)	34.2±4.4 (**5.4±0.5**)
6	22.8±1.5 (5.8±0.2)	22.9±1.4 (5.8±0.2)	23.2±1.6 (6.7±0.3)	**22.5±1.8** (**4.8±0.2**)
7	17.5±1.9 (11.6±0.2)	16.8±1.6 (10.7±0.3)	16.8±2.1 (**8.8±0.6**)	**15.5±2.0** (9.1±0.3)
8	11.7±2.6 (32.8±0.2)	12.0±2.3 (27.9±1.6)	14.0±2.9 (18.8±2.4)	**10.0±2.3** (**12.1±1.7**)

The number of features, the number of patterns and the partitioning into train and test sets of the real-world datasets are detailed in Table 2. The datasets can be separated into datasets which require considerable feature selection (1-4) and those which do not (5-8).

The GMEB algorithm attained the lowest error rates for the majority of both types of datasets investigated while using a comparatively small number of features. A graphical comparison to the RFE algorithm is available in Figure 2.



(a) Error rate

(b) Fraction of selected features

Figure 2: A comparison of the error percentage and the fraction of selected features between the GMEB and RFE algorithms for the eight real-world datasets.

6.5 Discussion

The GMEB algorithm performs comparatively well against the other algorithms, in regard to both the test error and the number of selected features. A possible explanation is that the other algorithms perform both classification and feature selection with the same variable w. In contrast, the GMEB algorithm performs the feature selection and classification simultaneously, while using variables σ and w respectively. The use of two variables also allows the GMEB algorithm to reduce the weight of a feature j with both w_j and σ_j, while the l_1 SVM uses only w_j. Perhaps this property of GMEB could explain why it produces comparable (and at times better) results than the competing algorithms both in classification problems where feature selection is and is not required.

7. Summary and future work

This paper presented a feature selection algorithm motivated by minimizing a GE bound. The *global* optimum of the objective function is found by solving a non-convex optimization problem. The equivalent optimization problems technique reduces this task to a conic RCQ problem. This enabled an extension of the GMEB algorithm to large scale classification problems.

The GMEB classifier is a linear classifier. Linear classifiers are the most important type of classifiers in a feature selection framework because feature selection is highly susceptible to overfitting.

We believe that the GMEB algorithm is just the first of a series of algorithms which may globally minimize increasingly tighter bounds on the generalization error.

Acknowledgment R.M. is partially supported by the fund for promotion of research at the Technion and by the Ollendorff foundation of the Electrical Engineering department at the Technion.

References

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004. http://www.stanford.edu/~boyd/cvxbook.html.

Paul S. Bradley and Olvi L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proc. 15th International Conf. on Machine Learning*, pages 82–90. Morgan Kaufmann, San Francisco, CA, 1998.

Glenn Fung and Olvi L. Mangasarian. Data selection for support vector machines classifiers. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 64–70, 2000. URL `ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-02.pdf`.

Yves Grandvalet and Stéphane Canu. Adaptive scaling for feature selection in svms. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 553–560. MIT Press, 2003.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, March 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

Ron Meir and Tong Zhang. Generalization bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Simon Perkins, Kevin Lacker, and James Theiler. Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, 3:1333–1356, March 2003.

Alain Rakotomamonjy. Variable selection using svm based criteria. *The Journal of Machine Learning Research*, 3:1357–1370, 2003. ISSN 1533-7928.

Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *The Journal of Machine Learning Research*, 3: 1439–1461, March 2003. ISSN 1533-7928.

Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. Feature selection for SVMs. In *Advances in Neural Information Processing Systems 13*, pages 668–674, 2000.

Appendix A. Proof of theorem 9

Before presenting our GE bound, we begin with a simple lemma.

Lemma 21 *Let $\{i_1, i_2, \ldots, i_s\}$, $i \in \mathbb{N}$ be an s-tuple of positive integers. Then, for any s and $i_j \in \mathbb{N}$,*

$$\sum_{j=1}^{s}(i_j - 1) + 1 \leq \prod_{j=1}^{s} i_j.$$

Proof: We proceed by induction on s. For $s = 1$ the claim holds trivially. Assume it holds for s, and proceed to $s + 1$. We have

$$\begin{aligned}\sum_{j=1}^{s+1}(i_j - 1) + 1 &= \sum_{j=1}^{s}(i_j - 1) + (i_{s+1} - 1) + 1 \\ &\leq \prod_{j=1}^{s} i_j + i_{s+1} - 1 \\ &\leq \prod_{j=1}^{s+1} i_j,\end{aligned}$$

where the last inequality follows from the observation that $in+1-n-i = (n-1)(i-1) \geq 0$ for $i \geq 1, n \geq 1$. □

Consider the d-dimensional grid defined by the coordinates $(i_1, \ldots, i_d)$, $i_j \in \mathbb{N}$, and introduce a mapping ζ from $(\mathbb{N})^d$ to $\mathbb{N}$ by requiring

$$i \leq \prod_{j=1}^{d} i_j \quad ; \quad i = \zeta(i_1, \ldots, i_d). \tag{20}$$

This can be done, for example, by setting $\zeta(i_1, \ldots, i_d)$ to be the Manhattan distance of the grid point $(i_1, \ldots, i_d)$ from the origin, given by $1 + \sum_{j=1}^{d}(i_j - 1)$. Equally (Manhattan) distant points are arbitrarily assigned an index i consistent with (20) so that a unique mapping is achieved. Lemma 21 establishes (20) in this case.

Assume initially that $0 \leq \sigma_j \leq s_j$, $j = 1, 2, \ldots, d$, and set

$$\mathcal{F}_s = \{f : \ f \in \mathcal{F}, \ 0 \preceq \sigma \preceq s\}$$

where $s = (s_1, \ldots, s_d)$.

Since ϕ is Lipschitz with constant $1/\gamma$, according to Theorem 8 in [MeiZha03][2], with probability at least $1-\delta/2$, for every $f \in \mathcal{F}_s$,

$$\mathbf{P}(Yf(X) \le 0) \le \hat{\mathbf{E}}_n \phi(Yf(X)) + \frac{2}{\gamma} R_n(\mathcal{F}) + K\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $K = (B\bar{s}+1)r$, $\bar{s} = \max_j s_j$, and where the Rademacher complexity is given by

$$R_n(\mathcal{F}_s) = \frac{1}{n}\mathbf{E}_\epsilon \mathbf{E}_{D_n} \sup_{\|w\|\le B} \sup_{|b|\le r} \sup_{\sigma \preceq s} \sum_{i=1}^n \epsilon_i \left(\left\langle w, \sigma * x^{(i)} \right\rangle + b\right).$$

The *empirical* Rademacher complexity is bounded by

$$\begin{aligned}
\hat{R}_n(\mathcal{F}_s) &= \frac{1}{n}\mathbf{E}_\epsilon \sup_{\|w\|\le B} \sup_{|b|\le r} \sup_{\sigma \preceq s} \sum_{i=1}^n \epsilon_i \left(\left\langle w, \sigma * x^{(i)} \right\rangle + b\right) \\
&= \frac{1}{n}\mathbf{E}_\epsilon \sup_{\|w\|\le B} \sup_{|b|\le r} \sup_{\sigma \preceq s} \left[\left\langle w, \sum_{i=1}^n \epsilon_i \sigma * x^{(i)} \right\rangle + \sum_{i=1}^n \epsilon_i b\right] \\
&\le \frac{1}{n}\mathbf{E}_\epsilon \sup_{\|w\|\le B} \sup_{\sigma \preceq s} \left\langle w, \sum_{i=1}^n \epsilon_i \sigma * x^{(i)} \right\rangle + \frac{1}{n}\mathbf{E}_\epsilon \sup_{|b|\le r} \sum_{i=1}^n \epsilon_i b \\
&\le \frac{n_+}{n}\mathbf{E}_\epsilon \sup_{\|w\|\le B} \sup_{\sigma \preceq s} \left\langle w, \frac{1}{n_+} \sum_{i\in I_+} \epsilon_i \sigma * x^{(i)} \right\rangle + \frac{n_-}{n}\mathbf{E}_\epsilon \sup_{\|w\|\le B} \sup_{\sigma \preceq s} \left\langle w, \frac{1}{n_-} \sum_{i\in I_-} \epsilon_i \sigma * x^{(i)} \right\rangle \\
&+ \frac{1}{n}\mathbf{E}_\epsilon \sup_{|b|\le r} \sum_{i=1}^n \epsilon_i b \ .
\end{aligned}$$

Consider the first of the three terms which bound the Rademacher complexity. We proceed to bound it in the following steps.

$$\begin{aligned}
\frac{1}{n_+}\mathbf{E}_\epsilon \sup_{\|w\|\le B} \sup_{\sigma \preceq s} \left\langle w, \sum_{i\in I_+} \epsilon_i \sigma * x^{(i)} \right\rangle &\le \frac{1}{n_+}\mathbf{E}_\epsilon \sup_{\|w\|\le B} \sup_{\sigma \preceq s} \|w\| \left\| \sum_{i\in I_+} \epsilon_i \sigma * x^{(i)} \right\| \\
&= \frac{B}{n_+}\mathbf{E}_\epsilon \sup_{\sigma \preceq s} \left[\sum_{j=1}^d \left(\sum_{i\in I_+} \epsilon_i \sigma_j x_j^{(i)}\right)^2\right]^{1/2} \\
&= \frac{B}{n_+}\mathbf{E}_\epsilon \left[\sum_{j=1}^d s_j^2 \left(\sum_{i\in I_+} \epsilon_i x_j^{(i)}\right)^2\right]^{1/2} \\
&\overset{(a)}{\le} \frac{B}{n_+} \left[\sum_{j=1}^d s_j^2 \mathbf{E}_\epsilon \left(\sum_{i\in I_+} \epsilon_i x_j^{(i)}\right)^2\right]^{1/2} \\
&= \frac{B}{\sqrt{n_+}} \sqrt{\sum_{j=1}^d s_j^2 v_j^+} \ .
\end{aligned}$$

[2] Slightly improved since we assume bounded functions in this work.

where Jensen's inequality was used in (a). A similar bound can also be derived for the second and third terms in the inequality. In particular

$$\frac{1}{n}\mathbf{E}_{\epsilon}\sup_{|b|\leq r}\sum_{i=1}^{n}\epsilon_i b \leq \frac{r}{\sqrt{n}} \ .$$

From McDiarmid's inequality[3], it follows that the empirical Rademacher complexity is concentrated around its mean. Specifically, with probability at least $1-\delta/2$

$$R_n(\mathcal{F}) \leq \hat{R}_n(\mathcal{F}) + K\sqrt{\frac{2\ln(4/\delta)}{n}}.$$

We thus conclude that with probability at least $1-\delta$, for every $f \in \mathcal{F}_s$,

$$\begin{aligned}\mathbf{P}(Yf(X)\leq 0) \leq \hat{\mathbf{E}}_n\phi(Yf(X)) + \frac{2B}{\gamma}\left(\frac{\sqrt{n_+}}{n}\sqrt{\sum_{j=1}^{d}s_j^2 v_j^+} + \frac{\sqrt{n_-}}{n}\sqrt{\sum_{j=1}^{d}s_j^2 v_j^-}\right) \\ + \frac{2r}{\gamma\sqrt{n}} + \left(\frac{2}{\gamma}+1\right)K\sqrt{\frac{2\ln(4/\delta)}{n}}.\end{aligned} \tag{21}$$

Next, we eliminate the dependence on $s=(s_1,\ldots,s_d)$. The basic idea is to construct a grid over the d-dimensional space $\mathbb{R}_+^d$, obtain a bound for each point of the grid, and then use the union bound to obtain a result for the full (infinite) grid .

Let $\{p_i\}$, $i\in\mathbb{N}$, be a set of positive numbers such that $\sum_i p_i = 1$, where for concreteness we set $p_i = 1/i(i+1)$, $i\in\mathbb{N}$. For each $1\leq j\leq d$ let $a_{i_j}^j = \sigma_0 2^{i_j}$, $i_j\in\mathbb{N}$, where σ_0 serves as the size of the smallest grid spacing. For each $(i_1,\ldots,i_d)$ and $i=\zeta(i_1,\ldots,i_d)$ (defined in (20)), set $\boldsymbol{a}_i = (a_{i_1}^1,\ldots,a_{i_d}^d)$, and denote by $\mathcal{M}(\boldsymbol{a}_i)$ the domain

$$\mathcal{M}(\boldsymbol{a}_i) = \left\{\sigma:\ \sigma_1\leq a_{i_1}^1,\ldots,\sigma_d\leq a_{i_d}^d\right\}.$$

From (21) and the union bound we have that with probability at least $1-\delta$ for all f with $\sigma\in\mathcal{M}(\boldsymbol{a}_i)$,

$$\begin{aligned}P(Yf(X)\leq 0) \leq \hat{E}_n\phi(Yf(X)) + \frac{2B}{\gamma}\left(\frac{\sqrt{n_+}}{n}\sqrt{\sum_{j=1}^{d}\left(a_{i_j}^j\right)^2 v_j^+} + \frac{\sqrt{n_-}}{n}\sqrt{\sum_{j=1}^{d}\left(a_{i_j}^j\right)^2 v_j^-}\right) \\ + \frac{2r}{\gamma\sqrt{n}} + \left(\frac{2}{\gamma}+1\right)K\sqrt{\frac{2\ln(4/p_i\delta)}{n}}.\end{aligned} \tag{22}$$

For each σ and j let $i_j(\sigma)$ be the smallest index for which $a_{i_j(\sigma)}^j \geq \sigma_j$. By the definition of $\tilde{\sigma}_j$, it follows that for each j, $i_j(\sigma)\leq\log_2(\tilde{\sigma}_j/\sigma_0)$, and $a_{i_j(\sigma)}^j \leq \tilde{\sigma}_j$. Let $i(\sigma) =$

[3] In future work one may employ tighter bounding techniques, such as the Entropy method, to improve the results.

$\zeta(i_1(\sigma), \ldots, i_d(\sigma)) \leq \prod_{j=1}^{d} i_j(\sigma)$. This implies that

$$\begin{aligned} \log(1/p_{i(\sigma)}) &\leq 2 \log 2i(\sigma) \\ &\leq 2 \log \left[2 \prod_{j=1}^{d} \log_2(\tilde{\sigma}_j/\sigma_0) \right] \\ &\leq 2 \sum_{j=1}^{d} \log \left(2^{1/d} \log_2(\tilde{\sigma}_j/\sigma_0) \right) . \end{aligned}$$

Combining these results with (22) and using the fact that $a^j_{i_j} \leq \tilde{\sigma}_j$ completes the proof. □

Appendix B. The Dual optimization problem

The functions $\frac{\tilde{w}_j^2}{\tilde{\sigma}_j}$ are convex only when the constraints on $\tilde{\sigma}$ are satisfied. If we were to formulate the dual problem by assigning the constraints $\tilde{\sigma} \succeq \mathbf{0}$ Lagrange multipliers there would be no consideration to the domain of $\frac{\tilde{w}_j^2}{\tilde{\sigma}_j}$. Therefore problem (17), in which the constraints $\tilde{\sigma} \succeq \mathbf{0}$ are made implicit (Boyd and Vandenberghe, 2004, pp.257–258) by modifying the objective function to be infinite when these constraints are violated, was proposed. The solutions of problems (17) and (16) are identical.

The Lagrange multipliers $\mu \in \mathbb{R}^n, \mu_+, \mu_-, \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}^n$ are defined as the multipliers of the respective constraints.

The Lagrangian is

$$L(\tilde{w}, \tilde{\sigma}, b, \xi; \mu, \mu_+, \mu_-, \mu_1, \mu_2) = \begin{cases} \mathbf{1}^T \xi + \mu^T \left(\mathbf{1} - \xi - Y \left(X^T \tilde{w} + b\mathbf{1} \right) \right) & \\ +\mu_+ \left((v^+)^T \tilde{\sigma} - R_+ \right) + \mu_- \left((v^-)^T \tilde{\sigma} - R_- \right) & \\ +\mu_1 \left(\sum_{j=1}^{d} \frac{\tilde{w}_j^2}{\tilde{\sigma}_j} - 1 \right) - \mu_2^T \xi & \text{if} \quad \tilde{\sigma} \succeq \mathbf{0} \\ \infty & \text{otherwise.} \end{cases}$$

The dual function is

$$\begin{aligned} \eta(\mu, \mu_+, \mu_-, \mu_1, \mu_2) &= \min_{\tilde{w}, \tilde{\sigma} \succeq \mathbf{0}, b, \xi} L(\tilde{w}, \tilde{\sigma}, b, \xi, \mu, \mu_+, \mu_-, \mu_1, \mu_2) \\ &= \min_{\tilde{w}, \tilde{\sigma} \succeq \mathbf{0}} \mu_1 \sum_{j=1}^{n} \frac{\tilde{w}_j^2}{\tilde{\sigma}_j} - \mu^T Y X^T \tilde{w} + (\mu_+ v^+ + \mu_- v^-)^T \tilde{\sigma} + \min_b \left(-b \mu^T y \right) \\ &+ \min_{\tilde{\xi}} (\mathbf{1} - \mu - \mu_2)^T \xi + \mathbf{1}^T \mu - \mu_1 - R_+ \mu_+ - R_- \mu_- \\ &= \min_{\tilde{w}, \tilde{\sigma} \succeq \mathbf{0}} \sum_{j=1}^{d} h_j(\tilde{w}_j, \tilde{\sigma}_j) + \begin{cases} \mathbf{1}^T \mu - \mu_1 - R_+ \mu_+ - R_- \mu_- & \text{if} \quad \mu^T y = 0, \\ & \mu + \mu_2 = \mathbf{1} \\ -\infty & \text{otherwise} \end{cases} , \end{aligned}$$

where $h_j(\tilde{w}_j, \tilde{\sigma}_j) = \mu_1 \frac{\tilde{w}_j^2}{\tilde{\sigma}_j} - \left(\sum_{i=1}^{n} \mu_i y^{(i)} x_j^{(i)} \right) \tilde{w}_j + (\mu_+ v_j^+ + \mu_- v_j^-) \tilde{\sigma}_j$.

The constant (in this context) μ_1 is nonnegative because it is a Lagrange multiplier of an inequality constraint. We separate minimizing $h_j(\tilde{w}_j, \tilde{\sigma}_j)$ into two cases:

1. $\mu_1 = 0$. The functions $h_j(\tilde{w}_j, \tilde{\sigma}_j)$ are linear. Consequently their minimal value is $-\infty$ unless $\mu = 0$ and then their value is 0.

2. $\mu_1 > 0$. The gradient of $h_j(\tilde{w}_j, \tilde{\sigma}_j)$ is

$$\nabla h_j = \begin{bmatrix} 2\mu_1 \frac{\tilde{w}_j}{\tilde{\sigma}_j} - \left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right) \\ -\mu_1 \left(\frac{\tilde{w}_j}{\tilde{\sigma}_j}\right)^2 + \mu_+ v_j^+ + \mu_- v_j^- \end{bmatrix} \quad .$$

We minimize $h_j(\tilde{w}_j, \tilde{\sigma}_j \geq 0)$ first over variable $\tilde{w}_j$ by solving $\nabla_{\tilde{w}_j} h_j = 0$. The result is

$$\tilde{w}_j^\star = \frac{\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}}{2\mu_1} \tilde{\sigma}_j \quad .$$

Thus

$$h_j(\tilde{w}_j^\star, \tilde{\sigma}_j \geq 0) = \left(\mu_+ v_j^+ + \mu_- v_j^- - \frac{\left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right)^2}{4\mu_1} \right) \tilde{\sigma}_j \quad .$$

Lemma 22 *The conditions for the minima of the functions $h_j(\tilde{w}_j, \tilde{\sigma}_j)$, $j = 1, \ldots, d$ to be bounded below are*

$$\left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right)^2 - 4\mu_1(\mu_+ v_j^+ + \mu_- v_j^-) \leq 0,\ j = 1, \ldots, d$$

respectively. If these conditions are met, the minimum value is 0.

Proof.

If $\mu_+ v_j^+ + \mu_- v_j^- - \frac{\left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right)^2}{4\mu_1} > 0$ then $\tilde{\sigma}^\star = 0 \Rightarrow h(\tilde{w}_j^\star, \tilde{\sigma}_j^\star) = 0$.

If $\mu_+ v_j^+ + \mu_- v_j^- - \frac{\left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right)^2}{4\mu_1} = 0$ then $h(\tilde{w}_j^\star, \tilde{\sigma}_j^\star) = 0$.

If $\mu_+ v_j^+ + \mu_- v_j^- - \frac{\left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right)^2}{4\mu_1} < 0$ then $\tilde{\sigma}^\star \to \infty \Rightarrow h(\tilde{w}_j^\star, \tilde{\sigma}_j^\star) \to -\infty$.
Consequently $h_j(\tilde{w}_j, \tilde{\sigma}_j)$, $j = 1, \ldots, d$ are bounded below if

$$\mu_+ v_j^+ + \mu_- v_j^- - \frac{\left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right)^2}{4\mu_1} \geq 0, \quad j = 1, \ldots, d.$$

This condition is equivalent to the condition

$$\left(\sum_{i=1}^n \mu(i) y^{(i)} x_j^{(i)}\right)^2 - 4\mu_1(\mu_+ v_j^+ + \mu_- v_j^-) \leq 0, \quad j = 1, \ldots, d$$

because $\mu_1 > 0$. □

In summary, the dual function is

$$\eta(\mu,\mu_1,\mu_2,\mu_3)=\begin{cases} \mathbf{1}^T\mu-\mu_1-R_+\mu_+-R_-\mu_- & \text{if } \mu^Ty=0,\quad \mu+\mu_2=\mathbf{1},\quad \mu_1>0 \\ & \qquad \left(\sum_{i=1}^n\mu(i)y^{(i)}x_j^{(i)}\right)^2-4\mu_1(\mu_+v_j^++\mu_-v_j^-)\leq 0,\forall j \\ 0 & \text{if } \mu=0,\quad \mu_1=0,\quad \mu_2=C,\quad \mu_3=\mathbf{1} \\ -\infty & \text{otherwise.} \end{cases}$$

Note that when $\mu_1=0$ the value of the dual function is equal to $\mathbf{1}^T\mu$. Additionally the conditions $\left(\sum_{i=1}^n\mu(i)y^{(i)}x_j^{(i)}\right)^2-4\mu_1(\mu_+v_j^++\mu_-v_j^-)\leq 0$, $j=1,\ldots,d$ are satisfied.

The dual optimization problem is

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T\mu-\mu_1-R_+\mu_+-R_-\mu_- \\ \text{subject to} & \left(\sum_{i=1}^n\mu(i)y^{(i)}x_j^{(i)}\right)^2-4\mu_1(\mu_+v_j^++\mu_-v_j^-)\leq 0,\quad j=1,\ldots,d \\ & \mu^Ty=0 \\ & \mu+\mu_2=\mathbf{1} \\ & \mu,\mu_1,\mu_+,\mu_-,\mu_2\succeq\mathbf{0}, \end{array}\tag{23}$$

where $\mu,\mu_1,\mu_+,\mu_-,\mu_2$ are the optimization variables. The variable μ_2 appears only in the constraints $\mu+\mu_2=\mathbf{1}$ and $\mu_2\succeq\mathbf{0}$. Therefore we can combine these constraints into the constraint $\mu\preceq\mathbf{1}$. Thus problem

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T\mu-\mu_1-R_+\mu_+-R_-\mu_- \\ \text{subject to} & \left(\sum_{i=1}^n\mu(i)y^{(i)}x_j^{(i)}\right)^2-4\mu_1(\mu_+v_j^++\mu_-v_j^-)\leq 0,\quad j=1,\ldots,d \\ & \mu^Ty=0 \\ & \mathbf{0}\preceq\mu\preceq\mathbf{1} \\ & \mu_1,\mu_+,\mu_-\geq 0, \end{array}\tag{24}$$

is equivalent to problem (23).

Appendix C. The dual of the dual optimization problem

The primal variables $\tilde{w},b,\xi$ of problem (15) can be determined from the Lagrange multipliers of the dual problem (16). In order deliniete this point, the dual problem associated with problem (16) is formulated. It is termed as the dual dual problem associated with problem (15). From a comparison between the dual dual problem and the primal problem the relations between the Lagrange multipliers of the dual problem (16) and the primal variables of problem (15) are uncovered.

Consider the dual problem (17). Denote the Lagrange multipliers of the d equality constraints, the quadratic over linear inequality constraints, the single equality constraint the lower bound box constraint on μ, the upper bound box constraint on μ and the nonnegativity constraints as $\breve{w},\breve{\sigma}\in\mathbb{R}^d,\breve{b}\in\mathbb{R},\lambda,\breve{\xi}\in\mathbb{R}^n,\lambda_+,\lambda_-\in\mathbb{R}$ respectively.

The Lagrangian, subject to $\mu_1\geq 0$, is

$$
\begin{aligned}
&L(\mu, \mu_+, \mu_-, s; \breve{w}, \breve{\sigma}, \breve{b}, \lambda, \breve{\xi}, \lambda_+, \lambda_-) \\
&= -\mathbf{1}^T\mu + \mu_1 + R_+\mu_+ + R_+\mu_- + \breve{w}^T\left(XY\mu - s\right) + \sum_{j=1}^{d}\breve{\sigma}_j\left[\frac{s_j^2}{4\mu_1} - \left(\mu_+v_j^+ + \mu_-v_j^-\right)\right] \\
&+b\left(\mu^Ty\right) - \lambda^T\mu - \xi^T\left(\mu - \mathbf{1}\right) - \lambda_1\mu_1 - \lambda_+\mu_+ - \lambda_-\mu_- \\
&= h(s, \mu_1) + \left(YX^T\breve{w} + by - \mathbf{1} + \xi - \lambda\right)^T\mu \\
&+\left(R_+ - \sum_{j=1}^{d}\breve{\sigma}_jv_j^+ - \lambda_+\right)\mu_+ + \left(R_- - \sum_{j=1}^{d}\breve{\sigma}_jv_j^- - \lambda_-\right)\mu_- + \mathbf{1}^T\xi,
\end{aligned}
$$

where $h(s, \mu_1) = \sum_{j=1}^{d}\breve{\sigma}_j\frac{s_j^2}{4\mu_1} - w^Ts + (1 - \lambda_1)\mu_1$ and otherwise is equal to infinity. The dual function is

$$
\eta(\breve{w}, \breve{\sigma}, \breve{b}, \lambda, \breve{\xi}, \lambda_+, \lambda_-) = \min_{\mu_1\geq 0} h(s, \mu_1) + \begin{cases} \mathbf{1}^T\breve{\xi} & \text{if} \quad Y\left(X^T\breve{w} + \breve{b}\mathbf{1}\right) - \mathbf{1} + \breve{\xi} - \lambda = \mathbf{0} \\ & \qquad R_+ - \sum_{j=1}^{d}\breve{\sigma}_jv_j^+ - \lambda_+ = 0 \\ & \qquad R_- - \sum_{j=1}^{d}\breve{\sigma}_jv_j^- - \lambda_- = 0 \\ -\infty & \text{otherwise.} \end{cases}
$$

Note that if any $\sigma_j = 0$, then the minimum of the function is minus infinity. Given $\tilde{\sigma} \succ \mathbf{0}$ it is clear that $\mu_1 = 0$ can't be a minimizer since the value of $h(s, 0)$ is infinity. The gradient of $h(s, \mu_1)$ according to the variable s is $\nabla_s h(s, \mu_1) = \frac{1}{2\mu_1}\check{\Sigma}s - \breve{w}$. Thus $s^\star = 2\mu_1\Sigma^{-1}\breve{w}$. Substituting the minimizer $s^\star$ into $h(s, \mu_1)$ we get $h(s^\star, \mu_1) = \left(1 - \lambda_1 - \breve{w}\check{\Sigma}\breve{w}\right)\mu_1$. The minimum of $h(s^\star, \mu_1)$ is zero if $1 - \lambda_1 - \breve{w}\check{\Sigma}\breve{w} \geq 0$ and minus infinity otherwise.

Therefore the dual problem associated with problem (17) is

$$
\begin{array}{ll}
\text{minimize} & \mathbf{1}^T\breve{\xi} \\
\text{subject to} & Y\left(X^T\breve{w} + \breve{b}\mathbf{1}\right) - \mathbf{1} + \breve{\xi} - \lambda = \mathbf{0} \\
& R_+ - \sum_{j=1}^{d}\breve{\sigma}_jv_j^+ - \lambda_+ = 0 \\
& R_- - \sum_{j=1}^{d}\breve{\sigma}_jv_j^- - \lambda_- = 0 \\
& 1 - \lambda_1 - \breve{w}\check{\Sigma}\breve{w} \geq 0 \\
& \breve{\sigma}, \breve{\xi}, \lambda, \lambda_1, \lambda_+, \lambda_- \succeq \mathbf{0}
\end{array}
\tag{25}
$$

with the variables $\breve{w}, \breve{\sigma}, \breve{b}, \breve{\xi}, \lambda, \lambda_1, \lambda_+, \lambda_-$. Problem (25) is equivalent to

$$
\begin{array}{ll}
\text{minimize} & \mathbf{1}^T\breve{\xi} \\
\text{subject to} & Y\left(X^T\breve{w} + \breve{b}\mathbf{1}\right) \succeq \mathbf{1} - \breve{\xi} \\
& \breve{w}\check{\Sigma}\breve{w} \leq 1 \\
& R_+ \geq \sum_{j=1}^{d}\breve{\sigma}_jv_j^+ \\
& R_- \geq \sum_{j=1}^{d}\breve{\sigma}_jv_j^- \\
& \breve{\sigma}, \breve{\xi} \succeq \mathbf{0}.
\end{array}
\tag{26}
$$

A comparison between the dual dual problem (26) and the primal problem (11) reveals that the problems are identical by the relations $\tilde{w} = \breve{w}, \tilde{\sigma} = \breve{\sigma}, b = \breve{b}$. Therefore the primal solution is the lagrange multipliers of the appropriate constraints of the solution of the dual problem (17). □

Appendix D. Convexification of problems with convex constraints and a convex over a linear objective function

Lemma 23 *Consider a problem of the form*

$$\begin{array}{ll} \text{minimize} & f_0(x)/(c^Tx+d) \\ \text{subject to} & f_i(x) \leq 0, \qquad\qquad i=1,\dots,m, \end{array} \tag{27}$$

where $f_0(x), f_1(x), \dots, f_m(x)$ are convex functions, and the domain of the objective function is defined as $\{x|c^Tx+d>0\}$.

Optimization problem (27) is a quasiconvex optimization problem which can be transformed into the following equivalent convex optimization problem

$$\begin{array}{ll} \text{minimize} & g_0(y,t) \\ \text{subject to} & g_i(y,t) \leq 0, \qquad i=1,\dots,m \\ & c^Ty+dt=1, \end{array} \tag{28}$$

where g_i is the perspective of f_i. The variables are $y \in \mathbb{R}^n$ and $t \in \mathbb{R}$. The domain of the objective function and inequality constraint functions restricts $t>0$ due to the definition of the domain of a perspective of a function.

Proof: To show equivalence, we first note that if x is feasible in (27),i.e. $f_i(x) \leq 0$, $i=1,\dots,m$ then the pair $(y,t)=(tx,t)$ is feasible in (28)

$$g_i(y,t)=g_i(tx,t)=tf_i\left(\frac{tx}{t}\right)=tf_i(x)\leq 0,$$

with the same objective value

$$g_0(y,t)=g_0(tx,t)=\frac{tf_0\left(\frac{tx}{t}\right)}{1}=\frac{tf_0(x)}{t(c^Tx+d)}=\frac{f_0(x)}{c^Tx+d}.$$

Note that the demand $c^Tx+d>0$ on the domain of x is satisfied because $c^Ty+dt=t(c^Tx+d)=1$. It follows that the optimal value of (27) is greater than or equal to the optimal value of (28).

Conversely, if (y,t) is feasible in (28), i.e. $g_i(y,t)\leq 0$, $i=1,\dots,m$, then $x=\frac{y}{t}$ is feasible in (27)

$$f_i(x)=f_i\left(\frac{y}{t}\right)=\frac{g_i(y,t)}{t}\leq 0$$

with the same objective value

$$\frac{f_0(x)}{c^Tx+d}=\frac{f_0(\frac{y}{t})}{c^T\left(\frac{y}{t}\right)+d}=\frac{tf_0(\frac{y}{t})}{c^Ty+dt}=\frac{g_0(y,t)}{1}=g_0(y,t).$$

Therefore the optimal value of (28) is greater than or equal to the optimal value of (27). Putting both parts together, we can conclude that the optimal values are the same.

Problem (28) is convex because the equality constraint is affine, $f_i(x)$, $i = 0, 1, \ldots, m$, are convex functions, and the perspective of a convex function is also a convex function. $\square$