Learning with Rigorous Support Vector Machines

Jinbo Bi[1] and Vladimir N. Vapnik[2]

[1] Department of Mathematical Sciences,
Rensselaer Polytechnic Institute, Troy NY 12180, USA
[2] NEC Labs America, Inc. Princeton NJ 08540, USA
bij2@rpi.edu, vlad@nec-labs.com

Abstract. We examine the so-called rigorous support vector machine (RSVM) approach proposed by Vapnik (1998). The formulation of RSVM is derived by explicitly implementing the structural risk minimization principle with a parameter H used to directly control the VC dimension of the set of separating hyperplanes. By optimizing the dual problem, RSVM finds the optimal separating hyperplane from a set of functions with VC dimension approximate to $H^2 + 1$. RSVM produces classifiers equivalent to those obtained by classic SVMs for appropriate parameter choices, but the use of the parameter H facilitates model selection, thus minimizing VC bounds on the generalization risk more effectively. In our empirical studies, good models are achieved for an appropriate $H^2 \in [5\%\ell, 30\%\ell]$ where ℓ is the size of training data.

1 Introduction

Support vector machines (SVMs) have proven to be a powerful and robust methodology for learning from empirical data. They originated from the concept in Vapnik-Chervonenkis (VC) theory which provides bounds on the generalization risk of a function f. Consider the learning problem of finding a function f, from a set of functions $\mathcal{F}$, which minimizes the expected risk functional $R[f] = \int L\left(y, f(\mathbf{x})\right) dP(\mathbf{x}, y)$ provided data $(\mathbf{x}, y)$ follows the distribution $P(\mathbf{x}, y)$. The loss functional $L(y, f(\mathbf{x}))$ measures the distance between the observation y and the prediction $f(\mathbf{x})$. However, $R[f]$ can not be directly calculated due to the unknown distribution P. Given ℓ training points $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_\ell, y_\ell)$ i.i.d. drawn from P, the empirical risk is computed as $R_{emp}[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, f(\mathbf{x}_i))$ to approximate $R[f]$. A typical form of VC bounds is stated as [13]: with probability $1 - \eta$, $R[f]$ is bounded from above by

$$R_{emp}[f] + \frac{h}{\ell} \cdot \mathcal{E}\left(1 + \sqrt{1 + \frac{4R_{emp}[f]}{\mathcal{E}h/\ell}}\right), \tag{1}$$

where $\mathcal{E} = 1 - \ln\left(\frac{h}{2\ell}\right) - \frac{1}{h}\ln\left(\frac{\eta}{4}\right)$, ℓ is the size of training data, and h is the VC dimension of $\mathcal{F}$. The second term of the bound (1) that controls the VC

confidence is basically a monotonically increasing function in terms of h for a fixed ℓ, and the ratio $\frac{h}{\ell}$ is the dominating factor in this term.

This paper focuses on binary classification problems. SVMs construct classifiers that generalize well by minimizing the VC bound. The classic SVM formulation (C-SVM) was derived based on a simplified version of the VC bound. In contrast, the SVM formulation examined in this paper is derived by explicitly implementing the structural risk minimization (SRM) principle without simplification. This approach can more effectively minimize the VC bound due to the easier tuning of the model parameter, so we name it "rigorous" SVM (RSVM). Instead of using a parameter C as in C-SVMs, RSVM uses a parameter H to provide an effective estimate of VC dimension h.

We follow Vapnik ([13], Chapter 10) in deriving the RSVM formulations in Section 2. Then we investigate basic characteristics of the RSVM approach (Section 3), compare RSVM to other SVM methods (Section 4), and solve RSVM by discussing strategies for choosing H and developing a decomposition algorithm (Section 5). Computational results are included in Section 6.

The following notation is used through this paper. Vectors are denoted by bold lower case letters such as $\mathbf{x}$, and presumed to be column vectors unless otherwise stated. The $\mathbf{x}'$ is the transpose of $\mathbf{x}$. Matrices are denoted by bold capital letters such as $\mathbf{Q}$. The $||\cdot||$ denotes the ℓ_2 norm of a vector. The inner product between two vectors such as $\mathbf{w}$ and $\mathbf{x}$ is denoted as $(\mathbf{w}\cdot\mathbf{x})$.

2 Review of RSVM Formulations

We briefly review the derivation of the RSVM approach in this section. Readers can consult [13] for a complete description. SVMs construct classifiers based on separating hyperplanes. A separating hyperplane $\{\mathbf{x} : (\mathbf{w}\cdot\mathbf{x}) + b = 0\}$ is in a canonical form if it satisfies $y_i\left((\mathbf{w}\cdot\mathbf{x}_i) + b\right) \geq 1,\ i = 1,\ldots,\ell$ [11]. The margin of separation is defined as the Euclidean distance between the separating hyperplane and either of the planes determined by $(\mathbf{w}\cdot\mathbf{x})+b = 1$ and $(\mathbf{w}\cdot\mathbf{x})+b = -1$. For a hyperplane of canonical form, the margin equals $1/||\mathbf{w}||$. For any such separating hyperplane characterized uniquely by a pair $(\mathbf{w}, b)$, a classifier can be constructed based on it as $g_{\mathbf{w},b}(\mathbf{x}) = \mathrm{sgn}\left((\mathbf{w}\cdot\mathbf{x}) + b\right)$.

Consider the set of classifiers $\mathcal{F} = \{g_{\mathbf{w},b} : ||\mathbf{w}|| \leq \frac{1}{\Delta}\}$ where Δ determines that any separating hyperplane $(\mathbf{w}, b)$ in this set separates training points $\mathbf{x}$ with a margin at least Δ. If input vectors $\mathbf{x}$ belong to a ball B_R of radius R, this set of classifiers defined on B_R has its VC dimension h bounded from above by $\frac{R^2}{\Delta^2} + 1$ [13] assuming that the dimension of $\mathbf{x}$ is larger than the ratio $\frac{R^2}{\Delta^2}$. This is often the case encountered in practice, especially for a kernel method. For instance, employing a RBF kernel corresponds to constructing hyperplanes in a feature space of infinite dimension. In real-world applications, a separating hyperplane does not always exist. To allow for errors, we use slack variables $\xi_i = \max\{0, 1 - y_i\left((\mathbf{w}\cdot\mathbf{x}_i) + b\right)\}$ [4], and the empirical risk is approximated by the ℓ_1 error metric $\frac{1}{\ell}\sum_{i=1}^{\ell}\xi_i$.

In C-SVMs, $R^2||\mathbf{w}||^2$ is regarded as a rough estimate of the VC dimension of $\mathcal{F}$ provided $\frac{1}{\Delta}$ can be attained at a $||\mathbf{w}||$. C-SVMs minimize the objective function $C\sum_{i=1}^{\ell}\xi_i + \frac{1}{2}||\mathbf{w}||^2$ on purpose to minimize the VC bound (1) with $R_{emp}[f]$ evaluated by $\frac{1}{\ell}\sum_{i=1}^{\ell}\xi_i$ and VC dimension h approximated by $R^2||\mathbf{w}||^2$. Comparing the objective function with the bound (1) yields that in order to achieve the goal, C should be chosen so that $1/C \approx \left(2R^2\mathcal{E}\left(1+\sqrt{1+\frac{4R_{emp}}{\mathcal{E}\hat{h}/\ell}}\right)\right)$ where R_{emp} and $\hat{h}$ are the smallest possible empirical risk and VC dimension respectively. Obviously, it is difficult to estimate this C due to no access to the R_{emp} and $\hat{h}$ beforehand. In practice, C is usually selected from a set of candidates according to cross-validation performance. The obtained C could be far from the desirable value if the candidate set is not well-selected.

Based on the bound (1), $R_{emp}[f]$ and h are the only two factors that a learning machine can control in order to minimize the bound. We thus do not directly minimize the bound as done in C-SVMs. Instead we regulate the two factors by fixing one and minimizing the other. RSVM restricts the set of functions $\mathcal{F}$ to one with VC dimension close to a pre-specified value, and minimizes the empirical risk by finding an optimal function from this set $\mathcal{F}$.

In RSVM formulations, the upper bound $R^2||\mathbf{w}||^2+1$ is used as an estimate of the VC dimension. If data is uniformly distributed right on the surface of the ball B_R, the VC dimension of $\mathcal{F}$ is exactly equal to $R^2||\mathbf{w}||^2+1$ according to the derivation of the bound [13]. However, data following such a distribution is not commonplace in real life. To make the estimation effective, we approximate the distribution by performing the transformation for each training point $\mathbf{x}_i$ as:

$$(\mathbf{x}_i - \bar{\mathbf{x}})/||\mathbf{x}_i - \bar{\mathbf{x}}|| \tag{2}$$

where $\bar{\mathbf{x}} = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathbf{x}_i$ is the mean. The transformed points live on the surface of the unit ball ($R=1$) centered at the origin. In C-SVMs, the VC dimension is commonly estimated using the same bound with the radius R of the smallest ball containing input data, which amounts to having most data points inside a unit ball after proper rescaling. The upper bound is closer to the true VC dimension when data points are on the surface of the ball than inside the ball. Hence with the transformation (2), $||\mathbf{w}||^2+1$ becomes a more accurate estimate of the VC dimension h. Then the VC dimension of $\mathcal{F}$ can be effectively controlled by restricting $||\mathbf{w}|| \le H$ with a given H. RSVM Primal is formulated in variables $\mathbf{w}$, b and $\boldsymbol{\xi}$ as [13]:

$$\begin{aligned}
\min \quad & E(\mathbf{w}, b, \boldsymbol{\xi}) = \textstyle\sum_{i=1}^{\ell}\xi_i && (3)\\
\text{s.t.} \quad & y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right) \ge 1-\xi_i,\ \xi_i \ge 0,\ i=1,\cdots,\ell, && (4)\\
& (\mathbf{w}\cdot\mathbf{w}) \le H^2. && (5)
\end{aligned}$$

Let γ be the Lagrange multiplier corresponding to the constraint (5), and α_i, s_i be the Lagrange multiplier to the constraints $y_i((\mathbf{w}\cdot\mathbf{x}_i)+b) \ge 1-\xi_i$ and $\xi_i \ge 0$ respectively. The index i is understood to run over $1,\cdots,\ell$ unless otherwise

noted. We can write the Lagrangian as

$$L = \sum \xi_i - \sum \alpha_i \left(y_i \left((\mathbf{w} \cdot \mathbf{x}) + b\right) - 1 + \xi_i\right) - \gamma \left(H^2 - (\mathbf{w} \cdot \mathbf{w})\right) - \sum s_i \xi_i, \quad (6)$$

and compute its derivatives with respect to the primal variables $\mathbf{w}$, b and $\boldsymbol{\xi}$. At optimality, these derivatives equal to 0. We thus have the optimality conditions:

$$\gamma \mathbf{w} = \sum \alpha_i y_i \mathbf{x}_i, \quad (7)$$

$$\sum \alpha_i y_i = 0, \quad (8)$$

$$0 \leq \alpha_i \leq 1, \quad \gamma \geq 0. \quad (9)$$

We derive the dual formulation based on the discussion of two cases: $\gamma = 0$ and $\gamma > 0$. By complementarity, either $\gamma = 0$ or $(\mathbf{w} \cdot \mathbf{w}) - H^2 = 0$ at optimality. Without loss of generality, we assume they are not both equal to 0 at optimality.
1. If $\gamma = 0$ or $(\mathbf{w} \cdot \mathbf{w}) < H^2$ at optimality, by the KKT conditions, the optimal solution to RSVM Primal is also optimal for the relaxation problem by removing the constraint (5) from RSVM Primal. The relaxation problem degenerates to a linear program, so the dual problem becomes a linear program as follows:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & \sum_{i=1}^{\ell} \alpha_i \\ \text{s.t.} & \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{x}_i = 0, \\ & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq 1. \end{array} \quad (10)$$

2. If $\gamma > 0$ or $(\mathbf{w} \cdot \mathbf{w}) = H^2$ at optimality, by Eq.(7), we have $\mathbf{w} = \frac{1}{\gamma} \sum_i \alpha_i y_i \mathbf{x}_i$. Substituting $\mathbf{w}$ into the Lagrangian, simplifying and adding in the dual constraints (8) and (9) yield the following optimization problem:

$$\begin{array}{ll} \max\limits_{\boldsymbol{\alpha}, \gamma} & W(\boldsymbol{\alpha}, \gamma) = \sum\limits_{i=1}^{\ell} \alpha_i - \dfrac{1}{2\gamma} \sum\limits_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \dfrac{\gamma H^2}{2} \\ \text{s.t.} & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq 1, \quad \gamma > 0. \end{array} \quad (11)$$

The optimal γ can be obtained by optimizing the unconstrained problem $\max_\gamma W(\alpha, \gamma)$. Set the derivative of W with respect to γ equal to 0. Solving the resulting equation produces two roots. The positive root $\frac{1}{H} \sqrt{\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}$ is the optimal γ for Problem (11). Substituting this optimal γ into $W(\boldsymbol{\alpha}, \gamma)$ and adding a minus sign to W yield the dual problem [13]:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & W(\boldsymbol{\alpha}) = H \sqrt{\sum\limits_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)} - \sum\limits_{i=1}^{\ell} \alpha_i \\ \text{s.t.} & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq 1. \end{array} \quad (12)$$

To perform capacity control, we should choose H such that $(\mathbf{w} \cdot \mathbf{w}) = H^2$ at optimality, which means the constraint (5) is *active*. Otherwise, RSVM corresponds

to just the training error minimization without capacity control. Therefore the second case, $\gamma > 0$, is of our concern. We refer to Problem (12) as RSVM Dual and assume the optimal γ is positive through later sections.

3 Characteristics of RSVM

In this section we discuss some characteristics of RSVM that are fundamental to its construction and optimization. Given a series of candidates for the parameter H, such that $0 < H_1 < H_2 < \cdots < H_t < \cdots$, we show that solving RSVM Primal (3) with respect to this series of values for H and choosing the best solution $(\mathbf{w}, b)$ actually yields a direct implementation of the induction principle of SRM. The following proposition characterizes this result. The C-SVM was also formulated following the SRM principle but not an explicit implementation.

Proposition 1. *Let $0 < H_1 < H_2 < \cdots < H_t < \cdots$. It follows the induction principle of SRM to solve RSVM (3) respectively with H_1, H_2, $\cdots$, H_t, $\cdots$ and choose the solution $(\mathbf{w}, b)$ that achieves the minimal value of the bound (1).*

Proof. Let $\mathcal{F}$ be the set consisting of all hyperplanes. We only need to prove that the series of subsets of $\mathcal{F}$, from each of which RSVM finds a solution, are nested with respect to $H_1, H_2, \ \cdots, \ H_t, \ \cdots$. In other words, they satisfy $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_t \subset \cdots$. Consider the two consecutive sets $\mathcal{F}_{t-1}$ and $\mathcal{F}_t$. It is clear that the set $\mathcal{F}_{t-1} = \{g_{\mathbf{w},b}(\mathbf{x}) : (\mathbf{w} \cdot \mathbf{w}) \leq H_{t-1}^2\}$ is a subset of $\mathcal{F}_t = \{g_{\mathbf{w},b}(\mathbf{x}) : (\mathbf{w} \cdot \mathbf{w}) \leq H_t^2\}$ for $H_{t-1} < H_t$. Recall that $g_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}((\mathbf{w} \cdot \mathbf{x}) + b)$. Then we verify that each element in the series has the structure:
1. $\mathcal{F}^t$ has a finite VC dimension $h_t \leq H_t^2 + 1$.
2. $\mathcal{F}_t$ contains the functions of the form $g_{\mathbf{w},b}(\mathbf{x})$, for which the loss function is an indicator function.

Similar to C-SVMs, RSVM constructs optimal hyperplanes by optimizing in the dual space. In general, solving the dual does not necessarily produce the optimal value of the primal unless there is no duality gap. In other words, the strong duality should be met, which requires that $W(\boldsymbol{\alpha}, \gamma) = E(\mathbf{w}, b, \boldsymbol{\xi})$ at the respective primal and dual optimal solutions. Equivalently, this imposes $W(\boldsymbol{\alpha}) = -E(\mathbf{w}, b, \boldsymbol{\xi})$ at the optimal RSVM Primal and Dual solutions.

Theorem 1. *There is no duality gap between Primal (3) and Dual (12).*

Proof. We use the following theorem [2]:
If (i) the problem $\min\{f(\mathbf{x}) : \mathbf{c}(\mathbf{x}) \leq 0, \mathbf{x} \in \mathbb{R}^n\}$ has a finite optimal value, (ii) the functions f and $\mathbf{c}$ are convex, and (iii) an interior point $\mathbf{x}$ exists, i.e., $\mathbf{c}(\mathbf{x}) < 0$, then there is no duality gap.
It is obvious that RSVM Primal satisfies the first two conditions. If $H > 0$, a feasible $\mathbf{w}$ exists for $(\mathbf{w} \cdot \mathbf{w}) < H^2$. With this $\mathbf{w}$, an interior point $(\mathbf{w}, b, \boldsymbol{\xi})$ can be constructed by choosing $\boldsymbol{\xi}$ large enough to satisfy $y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) > 1 - \xi_i, \ \xi_i > 0, \ i = 1, \cdots, \ell$.

RSVM Primal is a quadratically-constrained quadratic program that is a convex program. For a convex program, a local minimizer is also a global minimizer. If the solution is not unique, the set of global solutions is convex. Although the objective of RSVM Dual is not surely convex, RSVM Dual is in principle a convex program since it can be recast as a second-order cone program (SOCP) by substituting t for the square root term in the objective and adding a constraint to restrict the square root term no more than t. SOCPs are non-linear convex programs [7]. Therefore same as C-SVMs, RSVM does not get trapped at any local minimizer. We leave investigation of SOCPs in RSVM to future research.

Examining uniqueness of the solution can provide insights into the algorithm as shown for C-SVMs [5]. Since the goal is to construct a separating hyperplane characterized by $(\mathbf{w}, b)$, and the geometric interpretation of SVMs mainly rests on the primal variables, we provide Theorem 2 only addressing the conditions for the primal $\hat{\mathbf{w}}$ to be unique. In general, the optimal $\hat{\mathbf{w}}$ of RSVM is not necessarily unique, which is different from C-SVMs where even if the optimal solutions $(\mathbf{w}, b, \boldsymbol{\xi})$ may not be unique, they share the same optimal $\mathbf{w}$ [5]. Arguments about the offset $\hat{b}$ can be drawn similarly to those in [5], and will not be discussed here.

Theorem 2. *If the constraint* $(\mathbf{w} \cdot \mathbf{w}) \leq H^2$ *is active at any optimal solution to RSVM Primal, then the optimal* $\mathbf{w}$ *is unique.*

Proof. Realize that the optimal solution set of RSVM Primal is a convex set. Let $\hat{\mathbf{w}}$ be an optimal solution of RSVM, and $(\hat{\mathbf{w}} \cdot \hat{\mathbf{w}}) = H^2$. Assume that RSVM has another solution $\bar{\mathbf{w}}$ also satisfying $(\bar{\mathbf{w}} \cdot \bar{\mathbf{w}}) = H^2$. Then the middle point on the line segment connecting $\hat{\mathbf{w}}$ and $\bar{\mathbf{w}}$ is also optimal, but it cannot satisfy $(\mathbf{w} \cdot \mathbf{w}) = H^2$, contradicting the assumption.

Since RSVM Dual (12) is derived assuming $\gamma > 0$, solving it always produces a primal solution with $(\mathbf{w} \cdot \mathbf{w}) = H^2$. From the primal perspective, however, alternative solutions may exist satisfying $(\mathbf{w} \cdot \mathbf{w}) < H^2$, so Theorem 2 will not hold. Notice that such a solution is also optimal to the relaxation problem

$$\begin{array}{ll} \min & \sum_{i=1}^{\ell} \xi_i \\ \text{s.t.} & y_i \left((\mathbf{w} \cdot \mathbf{x}_i) + b \right) \geq 1 - \xi_i, \ i = 1, \cdots, \ell, \\ & \xi_i \geq 0, \ i = 1, \cdots, \ell. \end{array} \tag{13}$$

If the relaxation problem (13) has a unique solution $\bar{\mathbf{w}}$ and let $\bar{H} = ||\bar{\mathbf{w}}||$, there exist only two cases: 1. if $H < \bar{H}$, the constraint (5) must be active at any RSVM optimal solution, and thus Theorem 2 holds; 2. if $H \geq \bar{H}$, Primal (3) has only one solution $\bar{\mathbf{w}}$. In both cases, the optimal $\mathbf{w}$ of Primal (3) is unique. We hence conclude with Theorem 3.

Theorem 3. *If the relaxation problem (13) has a unique solution, then for any* $H > 0$*, RSVM (3) has a unique optimal* $\mathbf{w}$*.*

One of the principal characteristics of SVMs is the use of kernels [11]. It is clear that RSVM can construct nonlinear classifiers by substituting the kernel $k(\mathbf{x}_i, \mathbf{x}_j)$ for the inner product $(\mathbf{x}_i \cdot \mathbf{x}_j)$ in Dual (12). By using a kernel k, we map

the original data $\mathbf{x}$ to $\Phi(\mathbf{x})$ in a feature space so that $k(\mathbf{x}_i, \mathbf{x}_j) = (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j))$. From the perspective of primal, solving Dual (12) with inner products replaced by kernel entries corresponds to constructing a linear function $f(\mathbf{x}) = (\mathbf{w} \cdot \Phi(\mathbf{x})) + b$ in feature space. Similarly, by optimality conditions, $\mathbf{w} = \frac{1}{\gamma} \sum \alpha_i y_i \Phi(\mathbf{x}_i)$, and the function $f(\mathbf{x}) = \frac{1}{\gamma} \sum \alpha_i y_i k(\mathbf{x}_i, \mathbf{x})$ with $\gamma = \frac{1}{H} \sqrt{\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)}$. Notice that the transformation (2) now has to be taken in the feature space, i.e., $(\Phi(\mathbf{x}_i) - \bar{\Phi})/||\Phi(\mathbf{x}_i) - \bar{\Phi}||$ where $\bar{\Phi}$ denotes the mean of all $\Phi(\mathbf{x}_i)$s. We verify that this transformation can be implicitly performed by defining a kernel associated with k as $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\hat{k}(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\hat{k}(\mathbf{x}_i, \mathbf{x}_i)\hat{k}(\mathbf{x}_j, \mathbf{x}_j)}}$ (normalizing) where $\hat{k}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{\ell} \sum_{q=1}^{\ell} k(\mathbf{x}_i, \mathbf{x}_q) - \frac{1}{\ell} \sum_{p=1}^{\ell} k(\mathbf{x}_p, \mathbf{x}_j) + \frac{1}{\ell^2} \sum_{p,q=1}^{\ell} k(\mathbf{x}_p, \mathbf{x}_q)$ (centering).

A question naturally arises as how Dual (12) behaves in case $\gamma = 0$. Denote the optimal solution to Dual (12) by $\hat{\boldsymbol{\alpha}}$. Define $S(\hat{\boldsymbol{\alpha}}) = \sum_{i,j=1}^{\ell} \hat{\alpha}_i \hat{\alpha}_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$, and $S(\boldsymbol{\alpha}) \geq 0$ for any $\boldsymbol{\alpha}$ due to the positive semi-definiteness of k. As shown in the expression of the optimal γ, once $S(\hat{\boldsymbol{\alpha}}) > 0$, the optimal $\gamma > 0$. Hence in the case of $\gamma = 0$, $S(\hat{\boldsymbol{\alpha}})$ has to be 0. Many solvers for nonlinear programs use the KKT conditions to construct termination criteria. To evaluate the KKT conditions of Dual (12), the derivative of $W(\alpha)$ with respect to each α_i needs to be computed:

$$\nabla_i W = \frac{H y_i \left(\sum_j \alpha_j y_j k(\mathbf{x}_i, \mathbf{x}_j) \right)}{\sqrt{S(\boldsymbol{\alpha})}} - 1. \tag{14}$$

Realize that the derivative is not well-defined if $S(\boldsymbol{\alpha}) = 0$. Hence no solution can be obtained for Dual (12) if H is so large that $\gamma = 0$.

4 Comparison with Other SVMs

We compare RSVM to other SVM formulations in this section to identify their relationships. These approaches include the C-SVM with a parameter C [4, 13], the geometric Reduced convex Hull approach (RHSVM) with a parameter D [1, 3], and the ν-SVM classification with a parameter ν [12]. The comparison reveals the equivalence of these approaches for properly-selected parameter choices. We emphasize the equivalence of the normal vector $\mathbf{w}$ constructed by these approaches. Two $\mathbf{w}$ vectors are said to be "equivalent" if they are precisely the same or only scale differently.

Let $(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{w}})$ be optimal to the RSVM Dual and Primal. Denote the corresponding solutions of the C-SVM, RHSVM, and ν-SVM respectively by $(\boldsymbol{\alpha}^C, \mathbf{w}^C)$, $(\boldsymbol{\alpha}^D, \mathbf{w}^D)$ and $(\boldsymbol{\alpha}^\nu, \mathbf{w}^\nu)$. We obtain the following three propositions.

Proposition 2. *If* $C = \frac{1}{\hat{\gamma}} = \frac{\sqrt{S(\hat{\boldsymbol{\alpha}})}}{H}$, *then* $\boldsymbol{\alpha}^C = \frac{\hat{\boldsymbol{\alpha}}}{\hat{\gamma}}$ *is a solution to C-SVM.*

Proof. Consider Problem (11). Note that this problem is equivalent to RSVM Dual (12). We rewrite the objective function $W(\boldsymbol{\alpha}, \gamma) =$

$$\gamma \left(\sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \frac{H^2}{2} \right)$$

where $\boldsymbol{\alpha}$ has been rescaled by dividing by γ. Since H is a pre-specified constant in the above parentheses, for any fixed $\gamma \geq 0$, solving Problem (11) is equivalent to solving the following problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2} \sum\limits_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum\limits_{i=1}^{\ell} \alpha_i \\ \text{s.t.} & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq \frac{1}{\gamma}, \ i = 1, \cdots, \ell. \end{array} \tag{15}$$

Multiplying the solution to Problem (15) by γ produces a solution to Problem (11). Realize that Problem (15) is exactly the dual C-SVM formulation with the parameter $C = \frac{1}{\gamma}$. Set $C = \frac{1}{\hat{\gamma}}$ where $\hat{\gamma} = \frac{\sqrt{S(\hat{\boldsymbol{\alpha}})}}{H}$ is optimal to Problem (11). With this C, C-SVM has a solution $\boldsymbol{\alpha}^C = \frac{\hat{\boldsymbol{\alpha}}}{\hat{\gamma}}$.

Proposition 3. *If $D = \frac{2}{\sum \hat{\alpha}_i}$, then $\boldsymbol{\alpha}^D = \frac{2\hat{\boldsymbol{\alpha}}}{\sum \hat{\alpha}_i}$ is a solution to RHSVM.*

Proof. Consider RSVM Dual (12). The equality constraint can be rewritten as $\sum_{y_i=1} \alpha_i = \sum_{y_i=-1} \alpha_i = \delta$ for $\delta = \frac{1}{2} \sum_{i=1}^{\ell} \alpha_i$. Now define $\boldsymbol{\beta} = \frac{\hat{\boldsymbol{\alpha}}}{\hat{\delta}}$ where $\hat{\delta} = \frac{1}{2} \sum \hat{\alpha}_i$, and then $\sum \beta_i = 2$. It can be shown by contradiction that $\boldsymbol{\beta}$ is an optimal solution to the following problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2} \sum\limits_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} & \sum_{y_i=1} \alpha_i = 1, \ \sum_{y_i=-1} \alpha_i = 1, \\ & 0 \leq \alpha_i \leq \frac{1}{\hat{\delta}}, \ i = 1, \cdots, \ell. \end{array} \tag{16}$$

Realize that Problem (16) is exactly the dual RHSVM formulation [1] with the parameter $D = \frac{1}{\hat{\delta}}$. Then $D = \frac{2}{\sum \hat{\alpha}}$, and $\boldsymbol{\alpha}^D = \boldsymbol{\beta} = \frac{\hat{\boldsymbol{\alpha}}}{\hat{\delta}} = \frac{2\hat{\boldsymbol{\alpha}}}{\sum \hat{\alpha}_i}$.

Proposition 4. *If $\nu = \frac{\sum \hat{\alpha}_i}{\ell}$, then $\boldsymbol{\alpha}^\nu = \frac{\hat{\boldsymbol{\alpha}}}{\ell}$ is a solution to ν-SVM.*

Proof. Consider Problem (15) with parameter γ equal to the $\hat{\gamma}$. Multiply the $\boldsymbol{\alpha}$ in (15) by $\frac{\hat{\gamma}}{\ell}$. Set $\nu = \frac{\hat{\gamma}}{\ell} \sum \alpha_i^C = \frac{\sum \hat{\alpha}_i}{\ell}$. Solving the dual ν-SVM formulation [12]:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2} \sum\limits_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq \frac{1}{\ell}, \ i = 1, \cdots, \ell, \\ & \sum_{i=1}^{\ell} \alpha_i = \nu. \end{array} \tag{17}$$

Table 1. Relations between RSVM, C-SVM, RHSVM, and ν-SVM. For appropriate parameter choices as defined in the table, the optimal separating hyperplanes produced by the four methods are parallel. $S(\hat{\boldsymbol{\alpha}}) = \sum_{i,j=1}^{\ell} \hat{\alpha}_i \hat{\alpha}_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$.

	Parameter	Dual	Primal
RSVM	H	$\hat{\boldsymbol{\alpha}}$	$\hat{\mathbf{w}}$
C-SVM	$\frac{H}{\sqrt{S(\hat{\boldsymbol{\alpha}})}}$	$\hat{\boldsymbol{\alpha}} \frac{H}{\sqrt{S(\hat{\boldsymbol{\alpha}})}}$	$\mathbf{w}^C = \hat{\mathbf{w}}$
RHSVM	$\frac{2}{\sum \hat{\alpha}_i}$	$\hat{\boldsymbol{\alpha}} \frac{2}{\sum \hat{\alpha}_i}$	$\mathbf{w}^D = \hat{\mathbf{w}} \frac{2\sqrt{S(\hat{\boldsymbol{\alpha}})}}{H \sum \hat{\alpha}_i}$
ν-SVM	$\frac{\sum \hat{\alpha}_i}{\ell}$	$\hat{\boldsymbol{\alpha}} \frac{1}{\ell}$	$\mathbf{w}^\nu = \hat{\mathbf{w}} \frac{\sqrt{S(\hat{\boldsymbol{\alpha}})}}{H\ell}$

yields a solution $\boldsymbol{\alpha}^\nu = \frac{\hat{\gamma}}{\ell} \boldsymbol{\alpha}^C = \frac{\hat{\boldsymbol{\alpha}}}{\ell}$.

We summarize the above results in Table 1 along with comparison of primal $\hat{\mathbf{w}}$. Solving the four formulations with their parameters chosen according to Table 1 yields equivalent solutions, namely, the same orientation of the optimal separating hyperplanes. In RSVM, the VC dimension is pre-specified approximate to $H^2 + 1$ prior training. In C-SVM, instead of pre-specified, the VC dimension can be evaluated via $(\mathbf{w}^C \cdot \mathbf{w}^C)$ only after a solution $\mathbf{w}^C$ has been obtained. For the other two approaches, it is not straightforward to estimate VC dimension based on their solutions or parameter values.

5 Choosing H and the Decomposition Scheme

According to duality analysis, the parameter H should be selected within an upper limit $\hat{H}$ or otherwise Dual (12) will not produce a solution. We focus on finding an upper bound $\hat{H}$ on valid choices of H. A choice of H is valid for RSVM if there exists an optimal RSVM solution satisfying $(\mathbf{w} \cdot \mathbf{w}) = H^2$.

To proceed with our discussion, we first define separability. A set of data $(\mathbf{x}_i, y_i), i = 1, \cdots, \ell$, is linearly separable (or strictly separable) if there exists a hyperplane $\{\mathbf{x} : f(\mathbf{x}) = 0\}$, such that $y_i f(\mathbf{x}_i) \geq 0$ (or $y_i f(\mathbf{x}_i) > 0$); otherwise, it is linearly inseparable. Note that linear separability can be extended to hyperplanes in feature space, and thus it is not linear in input space. In terms of RSVM (3), if the minimal objective value is 0 for a choice of H, meaning $\xi_i = 0$ for all i, the data are strictly separable, whereas for inseparable data, the objective E will never achieve 0 for any choice of H. Without loss of generality, we discuss the strictly separable case and the inseparable case.

1. For the strictly separable case, a valid H exists so that $E(\hat{\mathbf{w}}, \hat{b}, \hat{\boldsymbol{\xi}}) = 0$. By strong duality, the dual $W(\hat{\boldsymbol{\alpha}}) = H\sqrt{S(\hat{\boldsymbol{\alpha}})} - \sum \hat{\alpha}_i = 0$, so $H = \frac{\sum \hat{\alpha}_i}{\sqrt{S(\hat{\boldsymbol{\alpha}})}}$, which is well-defined since $S(\hat{\boldsymbol{\alpha}}) > 0$ for a valid H. Rescaling $\hat{\boldsymbol{\alpha}}$ by $\hat{\delta} = \frac{1}{2} \sum_{i=1}^{\ell} \hat{\alpha}_i$ does not change the fraction, and $H = \frac{2}{\sqrt{S(\boldsymbol{\beta})}}$ where $\boldsymbol{\beta} = \frac{\hat{\boldsymbol{\alpha}}}{\hat{\delta}}$. As shown in Proposition 3, $\boldsymbol{\beta}$ is optimal for RHSVM dual problem (16). Notice that the largest valid H can be evaluated by computing the smallest possible $S(\boldsymbol{\beta})$. We thus relax Problem

(16) by removing the upper bound on α, $\alpha_i \leq \frac{1}{\delta}$, to produce the smallest $S(\hat{\boldsymbol{\beta}})$. Now $\hat{\boldsymbol{\beta}}$ is a solution to the RHSVM dual for the linearly separable case [1] where RHSVM finds the closest points in the convex hulls of each class of data, and $\sqrt{S(\hat{\boldsymbol{\beta}})}$ is the distance between the two closest points. So $\frac{1}{2}\sqrt{S(\hat{\boldsymbol{\beta}})}$ is the maximum margin of the problem. We therefore have the following proposition.

Proposition 5. *For linearly separable problems, $H > \frac{1}{\Delta}$ is not valid for RSVM (3) where Δ is the maximum hard margin.*

2. For the inseparable case, we can solve the relaxation problem (13) to have a solution $\bar{\mathbf{w}}$. Then $||\bar{\mathbf{w}}||$ is a potential upper bound on valid choices of H. In a more general case, if a kernel is employed, the point $\mathbf{x}_i$ in Problem (13) has to be replaced by its image $\Phi(\mathbf{x}_i)$ which is often not explicitly expressed, so it is impossible to directly solve Problem (13). We instead solve the problem with the substitution of $\mathbf{w} = \sum \beta_i y_i \Phi(\mathbf{x}_i)$, and the problem becomes

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},b,\boldsymbol{\xi}} & E(\boldsymbol{\beta}, b, \boldsymbol{\xi}) = \sum_{i=1}^{\ell} \xi_i \\ \text{s.t.} & y_i \left(\sum_{j=1}^{\ell} \beta_j y_j k(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ i = 1, \cdots, \ell. \end{array} \tag{18}$$

This is a linear program in terms of $\boldsymbol{\beta}$ not $\mathbf{w}$. Let $\mathcal{B}$ be the entire set of optimal solutions to Problem (18). Denote the $\mathbf{w}$ constructed based on any solution in $\mathcal{B}$ by $\mathbf{w}_{\boldsymbol{\beta}}$. If the supremum of $||\mathbf{w}_{\boldsymbol{\beta}}||$ $(= \sqrt{S(\boldsymbol{\beta})}\ )$ on $\mathcal{B}$ is finite, then it is an upper bound on valid H. We show this result in the following proposition.

Proposition 6. *If $\hat{H} = \sup\limits_{\boldsymbol{\beta} \in \mathcal{B}} ||\mathbf{w}_{\boldsymbol{\beta}}|| < \infty$, $H > \hat{H}$ is not valid for RSVM (3).*

Proof. Assume $H > \hat{H}$ is valid. We show that by contradiction, there exists another solution $\hat{\boldsymbol{\beta}}$ that is optimal to Problem (18) but not included in $\mathcal{B}$.

If H is valid, then $\gamma > 0$. The optimal $\hat{\mathbf{w}}$ to the Primal (3) with $\mathbf{x}_i$ replaced by $\Phi(\mathbf{x}_i)$ can be expressed as $\hat{\mathbf{w}} = \sum \hat{\beta}_i y_i \Phi(\mathbf{x}_i)$ where $\hat{\boldsymbol{\beta}} = \frac{\hat{\boldsymbol{\alpha}}}{\gamma}$ and $\hat{\boldsymbol{\alpha}}$ is optimal for Dual (12). Let the respective optimal objective values of Problem (3) and (18) be $\hat{E}$ and $\bar{E}$. Since the feasible region of Primal (3) is a subset of the feasible region of Problem (18), $\hat{E} \geq \bar{E}$. However, any $\mathbf{w}_{\boldsymbol{\beta}}$ is feasible to Primal (3) for $H > \hat{H}$, and thus optimal for Primal (3), so $\hat{E} = \bar{E}$. Then $\hat{\boldsymbol{\beta}}$ is also an optimal solution to Problem (18) but not included in $\mathcal{B}$ since $||\hat{\mathbf{w}}|| = H > \hat{H}$.

We devise our strategies for choosing H based on the above two propositions. Since H^2 is used to provide an estimate of the VC dimension and VC dimension is typically a positive integer no greater than ℓ, we consider just positive integers in $[0, \ell] \cap [0, \hat{H}^2]$ where $\hat{H}$ is calculated depending on the linear separability. Actually $\hat{H}$ can be obtained with small computational cost by solving either a hard margin C-SVM (separable) or a linear program (18) (inseparable). Moreover, previous research [13, 10] suggested that $\frac{h}{\ell} \in [0.05, 0.25]$ might be a good choice for the capacity control. We recommend selecting integers first from a small range, such

as $[0.05\ell, 0.25\ell] \cap [0, \hat{H}^2]$, as candidates for H^2. If it does not produce desirable performance, the range can be augmented to include choices in $[0, \ell] \cap [0, \hat{H}^2]$.

We next explore the possibility of large-scale RSVM learning by developing a decomposition scheme for RSVM based on the one proposed for C-SVMs [6, 8]. A decomposition algorithm consists of two steps. First, select the working set B of q variables. Second, decompose the problem and optimize $W(\boldsymbol{\alpha})$ on B. The algorithm repeats the two steps until the termination criteria are met. We show that the decomposition algorithm can be carried over on RSVM with small extra cost of computation as compared with the algorithm for C-SVMs.

For notational convenience, we switch to the matrix vector product notation here. Define the matrix $\mathbf{Q}$ as $\mathbf{Q}_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$. Then $W(\boldsymbol{\alpha}) = H\sqrt{S(\boldsymbol{\alpha})} - \mathbf{e}'\boldsymbol{\alpha}$ where $S(\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\mathbf{Q}\boldsymbol{\alpha}$ and $\mathbf{e}$ is a vector of ones of appropriate dimension. Let variables $\boldsymbol{\alpha}$ be separated into a working set B and the remaining set N. We properly arrange $\boldsymbol{\alpha}$, $\mathbf{y}$ and $\mathbf{Q}$ with respect to B and N so that

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_B \\ \boldsymbol{\alpha}_N \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} \mathbf{y}_B \\ \mathbf{y}_N \end{pmatrix}, \ \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{BB} & \mathbf{Q}_{BN} \\ \mathbf{Q}_{NB} & \mathbf{Q}_{NN} \end{pmatrix}.$$

Decompose $S(\boldsymbol{\alpha})$ to the sum of three terms $S_{BB} = \boldsymbol{\alpha}'_B\mathbf{Q}_{BB}\boldsymbol{\alpha}_B$, $S_{BN} = 2(\mathbf{Q}_{BN}\boldsymbol{\alpha}_N)'\boldsymbol{\alpha}_B$, and $S_{NN} = \boldsymbol{\alpha}'_N\mathbf{Q}_{NN}\boldsymbol{\alpha}_N$, and rewrite $\mathbf{e}'\boldsymbol{\alpha} = \mathbf{e}'\boldsymbol{\alpha}_B + \mathbf{e}'\boldsymbol{\alpha}_N$. Since the $\boldsymbol{\alpha}_N$ are fixed, $\mathbf{p}_{BN} = \mathbf{Q}_{BN}\boldsymbol{\alpha}_N$, S_{NN} and $\mathbf{e}'\boldsymbol{\alpha}_N$ are constant. The $\mathbf{e}'\boldsymbol{\alpha}_N$ can be omitted from $W(\boldsymbol{\alpha})$ without changing the solution. Dual (12) can be reformulated as the following subproblem in variables $\boldsymbol{\alpha}_B$:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}_B} & H\sqrt{\boldsymbol{\alpha}'_B\mathbf{Q}_{BB}\boldsymbol{\alpha}_B + 2\mathbf{p}'_{BN}\boldsymbol{\alpha}_B + S_{NN}} - \mathbf{e}'\boldsymbol{\alpha}_B \\ \text{s.t.} & \mathbf{y}'_B\boldsymbol{\alpha}_B = -\mathbf{y}'_N\boldsymbol{\alpha}_N, \\ & \mathbf{0} \le \boldsymbol{\alpha}_B \le \mathbf{e}. \end{array} \tag{19}$$

Note that S_{NN} can not be omitted as in C-SVMs since it stays inside the square root. Typically the working set B consists of merely a few variables, and N contains the majority of variables. Computing $\mathbf{p}_{BN}$ and S_{NN} consumes significant time. If the kernel matrix is large and not stored in memory, computing $\mathbf{Q}_{NN}$ and $\mathbf{Q}_{BN}$ collapses the efficiency of the decomposition. Let $(\mathbf{Q}\boldsymbol{\alpha})_B$ be the vector consisting of the first q components (in the working set B) of $\mathbf{Q}\boldsymbol{\alpha}$. Then $(\mathbf{Q}\boldsymbol{\alpha})_B = \mathbf{Q}_{BB}\boldsymbol{\alpha}_B + \mathbf{Q}_{BN}\boldsymbol{\alpha}_N$. The key to our scheme is the use of the two equations:

$$\mathbf{p}_{BN} = (\mathbf{Q}\boldsymbol{\alpha})_B - \mathbf{Q}_{BB}\boldsymbol{\alpha}_B, \tag{20}$$

$$S_{NN} = S(\boldsymbol{\alpha}) - S_{BB} - S_{BN}, \tag{21}$$

in computing $\mathbf{p}_{BN}$ and S_{NN} instead of a direct evaluation. We keep track of the value of $S(\boldsymbol{\alpha})$ after solving each subproblem. Compared with the algorithm for C-SVMs, the update of $S(\boldsymbol{\alpha})$ and the evaluation of S_{NN} introduce the extra computation which, however, takes only a few arithmetic operations. See our implementation[3] for more details of the algorithm.

[3] A preliminary solver for RSVM written in C++ is available at *http://www.cs.rpi.edu/~bij2/rsvm.html.*

6 Experimental Studies

The goals of our experiments were to demonstrate the performance of RSVM, discover knowledge for choosing proper H, and compare RSVM to other SVM approaches. We conducted our experiments on the MNIST hand-written digit database (60,000 digits, 784 variables), the Wisconsin Breast Cancer (569 observations, 30 variables) and Adult-4 (4781 examples, 123 variables) benchmark datasets.[4] For the digit dataset, we want to distinguish odd numbers from even numbers. The proportion of positive examples to negative examples is roughly even in the digit data, but the proportions in the Adult-4 and Breast Cancer data are 1188/3593 and 212/357, respectively. We randomly took 200 examples from Breast Cancer and 1000 examples from Adult-4, respectively, for training such that the ratios of positive to negative examples of training data are the same as of the entire data. The remaining examples were used for test.

The data were preprocessed in the following way: examples were centered to have mean $\mathbf{0}$ by subtracting the mean of the training examples; then each variable (totally $28 \times 28 = 784$ variables) was scaled to have standard deviation 1; after that, each example was normalized to have ℓ_2-norm equal 1. Note that the test data should be blinded to the learning algorithm. Hence the test data were preprocessed using the mean of training data and the standard deviation of each variable computed based on training data. We simply used the inner product (a linear kernel) in all our experiments.

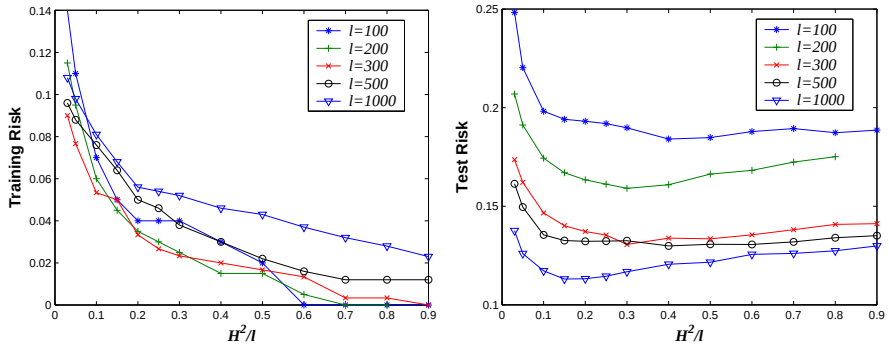

Fig. 1. Curves of the error rates versus the ratio H^2/ℓ for various choices of ℓ: *left*, the training risk; *right*, the test risk.

We first performed a series of experiments on the digit dataset. The first ℓ digits of the database were adopted as the training dataset, and ℓ was respectively

[4] MNIST data was downloaded from *http://yann.lecun.com/exdb/mnist/*. The Breast Cancer and Adult-4 datasets were obtained respectively from UC-Irvine data repository and *http://www.research.microsoft.com/~jplatt* [9].

Table 2. Results obtained using the training sets of $\ell = 200$ (left) and $\ell = 1000$ (right) digits. The N_SV stands for the number of support vectors. R_{trn} and R_{tst} are the percentages of errors on the training and test datasets, respectively. Numbers in the column of the parameter D should be multiplied by 10^{-2}.

H^2/ℓ	N_SV	R_{trn}	R_{tst}	C	D	ν	N_SV	R_{trn}	R_{tst}	C	D	ν
0.03	184	11.5	20.7	0.075	1.097	0.912	599	10.8	13.8	0.124	0.345	0.579
0.05	167	9.5	19.1	0.129	1.258	0.795	518	9.8	12.6	0.247	0.414	0.482
0.1	140	6.0	17.4	0.298	1.623	0.616	431	8.1	11.7	0.636	0.534	0.374
0.15	127	4.5	16.7	0.476	1.890	0.529	394	6.8	11.3	1.021	0.600	0.333
0.2	119	3.5	16.3	0.712	2.215	0.452	378	5.6	11.3	1.407	0.650	0.308
0.25	114	3.0	16.0	0.982	2.528	0.396	363	5.4	11.5	1.906	0.711	0.281
0.3	105	2.5	16.0	1.21	2.732	0.366	351	5.2	11.7	2.378	0.756	0.265
0.4	100	1.5	16.1	1.71	3.153	0.317	333	4.6	12.1	3.323	0.830	0.241
0.5	100	1.5	16.6	2.25	3.567	0.280	325	4.3	12.2	4.255	0.890	0.225
0.6	95	0.5	16.8	2.90	4.079	0.245	322	3.7	12.6	5.190	0.942	0.212
0.7	96	0	17.2	3.64	4.649	0.215	320	3.2	12.6	6.208	0.996	0.200
0.8	94	0	17.5	4.50	5.326	0.188	318	2.8	12.7	7.233	1.046	0.191

equal to 100, 200, 300, 500, 1000. The last 10,000 digits comprised the test set for all the experiments. Figure 1 presents the performance of RSVM obtained on distinct sizes of training data with a large spread of choices of H^2. The training risk monotonically decreases as H^2 increases for all the choices of ℓ. The corresponding test risk curve, however, has the minimum point as shown in Figure 1(right). Although the optimal ratios H^2/ℓ are different for various sizes of training data, they are roughly located around $[0.05, 0.30]$ except when $\ell = 100$, it is a little off. In this case, we may want to explore the full range of valid H, $[0, \min\{\hat{H}^2, \ell\}]$, where $\hat{H}^2 = 90$ obtained by solving the hard-margin C-SVM for $\ell = 100$. Table 2 provides in detail the results obtained by RSVM for $\ell = 200$ and $\ell = 1000$. We applied the heuristic of choosing the ratio H^2/ℓ in $[0.05, 0.30]$ into the subsequent experiments on the Adult-4 and Breast Cancer data. Results are summarized in Table 3 which shows that this heuristic is useful since good models are achieved with H^2/ℓ chosen in this much smaller range.

As shown in Table 2, the ratio H^2/ℓ was chosen from 0.03 to 0.8 in the experiments on digits. The corresponding value of C for C-SVM spreads within a small range, for instance, from 0.124 to 7.233 for $\ell = 1000$. As in Table 3, the ratio was chosen in a smaller range for experiments with Adult-4. But the value of C jumped from small numbers to very big numbers. Hence it is hard to pre-determine the proper range for C-SVM and to evaluate what happens in training when using a C far beyond the acceptable range. In addition, the proper range of C (not only the best C) is problem-dependent by cross referencing the results for $\ell = 200$ and $\ell = 1000$ in Table 2 and results in Table 3. Hence it is not straightforward to distinguish if C takes a proper value so that H^2 is valid.

Because of the geometric motivation for RHSVM, the parameter D scales with the size of training data. From Table 2, RHSVM used rather small values

of D in our experiments, especially for a large training set and a small h (or small H^2). Since D is the upper bound on each α, too small D may cause computation unstable. As ν is the lower bound on the fraction of support vectors as well as the upper bound on the fraction of error examples, the range of ν is conceptually $[0\%, 100\%]$. Hence it does not bear the problem of potentially lying in a wrong range. But results from our experiments suggest that ν should be tuned carefully at the lower end because small variation (from 0.2 to 0.191) on ν may cause a large change on h (from 700 to 800), especially on large datasets. All parameters in these methods change monotonically when increasing H, so these methods can effectively trade off between the training risk and the VC dimension. They can perform similarly provided there is a cogent way to tune their parameters.

Table 3. Results obtained on Adult-4 (left, $\ell = 1000$) and Breast Cancer (right, $\ell = 200$) datasets. FP/FN$_r$ and FP/FN$_t$ represent false positive versus false negative rates respectively for training and test.

H^2/ℓ	R_{trn}	FP/FN$_r$	R_{tst}	FP/FN$_t$	C	R_{trn}	FP/FN$_r$	R_{tst}	FP/FN$_t$	C
0.03	15.1	8.1/17.3	16.7	11.4/18.5	3.2e-1	2.5	0/4.0	2.4	0.7/3.4	2.1e-2
0.05	13.7	10.8/14.6	16.8	14.5/17.6	6.4e-1	2.5	0/4.0	2.4	0.7/3.4	3.7e-1
0.1	13.9	4.4/17.0	18.0	10.9/20.4	1.0e+5	1.5	1.3/1.6	2.2	0.7/3.0	1.3e+0
0.15	15.4	6.0/18.5	19.3	12.8/21.4	1.2e+5	1.5	1.3/1.6	2.4	0.7/3.4	2.6e+0
0.2	13.8	6.0/16.4	19.8	22.0/19.1	1.5e+5	1.0	2.7/0	2.4	0.7/3.4	3.8e+0
0.25	13.4	4.0/16.5	19.5	12.7/21.7	1.6e+5	1.0	2.7/0	3.8	2.1/4.7	4.8e+0
0.3	12.5	2.4/15.8	18.5	9.5/21.1	1.7e+5	1.0	2.7/0	3.5	2.1/4.3	5.5e+0

7 Conclusion

We have described the RSVM approach and examined how to tune its parameter H and how to train it in large-scale. We compared RSVM with other SVM approaches, and the comparison revealed the relationships between each other of these methods. This work made efforts to address the derivation of SVM algorithms from the fundamentals. C-SVMs minimize the VC bound in a straightforward way with a constant C used to approximate a varying term in bound (1). The RSVM approach uses a parameter H to directly estimate the VC dimension of the hypothesis space. The bound (1) can be effectively minimized by minimizing the training risk with a given H, and finding the H which results in the minimum value of the bound. To date, no appropriate parameter range has been proposed for C-SVMs generally effective for problems of all kinds. On the contrary, a proper range for H can be easily determined by solving simple optimization problems as discussed in Section 5. Furthermore, we can shrink the parameter range even more by examining integer values in the range $[0.05\ell, 0.30\ell] \cap [0, \hat{H}]$ first. Based on our empirical observation, the resulting models based on this range were not far from the best model.

One important open problem is to develop fast and efficient solvers for the RSVM dual problem. We may convert RSVM dual to a SOCP since SOCPs can be solved with the same complexity as the C-SVM dual quadratic program. Our preliminary investigation shows that large-scale RSVM learning is possible by means of a decomposition scheme. A SMO-like algorithm [9], (a decomposition scheme with $q = 2$,) may provide a more efficient implementation for RSVM.

Acknowledgements

The material is mainly based on research supported by NEC Labs America, Inc. Many thanks to the reviewers for their valuable comments.

References

1. K. P. Bennett and E. J. Bredensteiner. Duality and geometry in SVM classifiers. In *Proc. 17th International Conf. on Machine Learning*, pages 57–64, San Francisco, CA, 2000. Morgan Kaufmann.
2. D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1999.
3. J. Bi and K. P. Bennett. Duality, geometry, and support vector regression. In *Advances in Neural Information Processing Systems, Volumn 14*, Cambridge, MA., 2001. MIT Press.
4. B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.
5. C. J. C. Burges and D. J. Crisp. Uniqueness theorems for kernel methods. *Technical Report MSR-TR-2002-11*, 2002.
6. T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press, 1999.
7. M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.
8. E. Osuna, R. Freund, and F. Girosi. Improved training algorithm for support vector machines. In *Proceedings of IEEE Neural Networks for Signal Processing VII Workshop*, pages 276–285, Piscataway, NY, 1997. IEEE Press.
9. J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208. MIT Press, 1999.
10. B. Schölkopf, C.J.C. Burges, and V. N. Vapnik. Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings of First International Conference on Knowledge Discovery & Data Mining*, Menlo Park, 1995. AAAI Press.
11. B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press, Cambridge, MA., 2002.
12. B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.
13. V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.