Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Neurocomputing 55 (2003) 79–108

NEUROCOMPUTING

www.elsevier.com/locate/neucom

A geometric approach to support vector regression

Jinbo Bi*, Kristin P. Bennett

Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Received 14 March 2002; accepted 8 January 2003

Abstract

We develop an intuitive geometric framework for support vector regression (SVR). By examining when ε-tubes exist, we show that SVR can be regarded as a classification problem in the dual space. Hard and soft ε-tubes are constructed by separating the convex or reduced convex hulls, respectively, of the training data with the response variable shifted up and down by ε. A novel SVR model is proposed based on choosing the max-margin plane between the two shifted data sets. Maximizing the margin corresponds to shrinking the effective ε-tube. In the proposed approach, the effects of the choices of all parameters become clear geometrically. The kernelized model corresponds to separating the convex or reduced convex hulls in feature space. Generalization bounds for classification can be extended to characterize the generalization performance of the proposed approach. We propose a simple iterative nearest-point algorithm that can be directly applied to the reduced convex hull case in order to construct soft ε-tubes. Computational comparisons with other SVR formulations are also included.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Support vector machines; Kernel methods; Regression; Nearest-point algorithms

1. Introduction

Support vector machines (SVMs) [21,22,7] are a robust methodology for inference with minimal parameter choices. Intuitive geometric formulations exist for the classification case addressing both the error metric and capacity control [4,6]. For linearly separable classification, the primal SVM finds the separating plane with the maximum hard margin between the classes. As shown in Fig. 1 (left), the equivalent dual SVM

* Corresponding author.
E-mail address: bij2@rpi.edu (J. Bi).

0925-2312/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0925-2312(03)00380-1

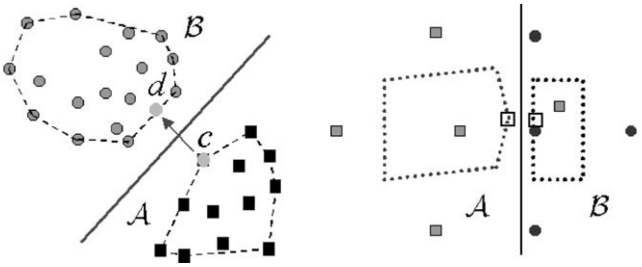

Fig. 1. SVM Classification; left: hard margin case; right: soft margin case.

computes the closest points in the convex hulls of the data from each class. The optimal separating plane bisects these two points. For the inseparable case, the primal SVM optimizes the soft margin of separation between the two classes. As shown in Fig. 1 (right), the corresponding dual SVM finds the closest points in the reduced convex hulls. In this paper, we derive analogous arguments for SVM regression (SVR).

We provide a geometric explanation for SVR with the ε-insensitive loss function. From the primal perspective, a linear function with no residuals greater than ε corresponds to an ε-tube constructed about the data in the space of the data attributes and the response variable [21] (see e.g. Fig. 3(a)). In Section 2, for a fixed $\varepsilon > 0$ we examine the question, "When does a "perfect" or "hard" ε-tube exist?". As shown by duality analysis, the existence of a hard ε-tube depends on the separability of two sets. The two sets consist of the training data augmented with the response variable shifted up and down by ε. In the dual space, regression becomes the classification problem of distinguishing between these two sets. The geometric formulations developed for the classification cases [4,6] become applicable to the regression cases [5].

Let us briefly review the geometric interpretation of classic SVMs for binary classification of both linearly separable and inseparable data. For the separable case, finding the maximum margin plane between the two classes $\mathscr{A}$ and $\mathscr{B}$ is equivalent to bisecting the closest points in the convex hulls of each class. As in Fig. 1 (left), the convex hull method finds the closest points $\mathbf{c}$ and $\mathbf{d}$, respectively, in the convex hulls of $\mathscr{A}$ and $\mathscr{B}$, then bisects them to obtain the optimal separating plane. A similar interpretation can be achieved for the inseparable case by using reduced convex hulls. Reduced convex hulls are a subset of the original convex hull created by reducing the influence of any single data point. Solving the reduced convex hull formulation is exactly equivalent to solving the classic inseparable SVM classification for appropriate choices of parameters. The optimal plane bisects the closest points in the reduced convex hulls as in Fig. 1 (right).

In this work we show the same geometric intuition can be applied to the ε-insensitive regression problem by converting the regression problem to a classification problem. Fig. 2 (left) illustrates a simple regression problem. In Fig. 2 (middle), the class of points (circles) that overestimate the data is constructed by shifting each data point up by ε and the class of points (triangles) that underestimate the data is constructed by shifting each data point down by ε. Then a maximum margin separating plane

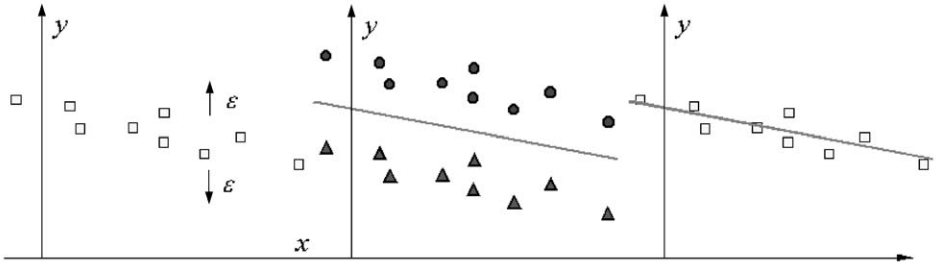

Fig. 2. SVM Regression; left: original data; middle: shifted data and separating plane; right: regression plane.

is constructed between the two shifted data sets. Fig. 2 (right) demonstrates how the separating plane is the final regression function. Fig. 2 assumes that a hard ε-tube exists with zero ε-insensitive error. As in classification, the reduced convex hull approach is used for the soft ε-tube case when the ε-insensitive error is positive.

The primary contributions of this work include the following:

- A novel geometric interpretation of SVR from the dual perspective is developed along with a mathematically rigorous derivation of the geometric concepts.
- New SVR models, H-SVR and RH-SVR, are proposed based on separating the convex hulls or reduced convex hulls of the augmented training data. We call the resulting formulation convex hull SVR (H-SVR) and reduced convex hull SVR (RH-SVR).[1] Much like in SVM classification, to compute a hard ε-tube, H-SVR computes the closest points in the convex hulls of the augmented classes. The corresponding maximum margin (max-margin) plane defines the effective ε-tube. The size of the margin determines how much the effective ε-tube shrinks. Similarly, to compute a soft ε-tube, RH-SVR finds the closest points in the reduced convex hulls of the two augmented sets. For properly tuned parameters, the methods perform similarly but not necessarily identically to traditional ε-SVR [9] and ν-SVR [19].
- RH-SVR eliminates the parameter C, for which the geometric role or interpretation is not known for the formulations of ε-SVR and ν-SVR. The geometric roles of the two parameters of RH-SVR, ν and ε, are very clear, facilitating model selection, especially for nonexperts. Like ν-SVR, RH-SVR shrinks the ε-tube and has a parameter ν controlling the robustness of the solution. The parameter ε acts as an upper bound on the size of the allowable insensitivity to error. Thus the effective size of the tube, denoted as $\hat{\varepsilon}$, is determined by both ε and ν.
- A bound on generalization performance is derived by analyzing the connection between the primal regression problem and dual classification problem. The regression model obtained by solving the RH-SVR has the property that a point having residual larger than ε in the primal space corresponds to a misclassified point in the dual classification space, and a point having residual larger than $\hat{\varepsilon}$ corresponds to a margin error in the dual space. Therefore bounds derived for classification can be adopted

[1] Note that the algorithms were originally denoted as C-SVR and RC-SVR in [5]. We change the acronym to avoid confusion with the classic SVR.

to analogously analyze the generalization performance of RH-SVR. The proposed RH-SVR model directly minimizes this bound.

- One of the major benefits of H-SVR and RH-SVR is that they can be solved by fast iterative nearest-point algorithms such as those used in the SVM classification. Prior nearest-point algorithms (NPA) [13,14] were very effective but limited to the convex hull case for classification. We propose a generic NPA applicable to both the regular and reduced convex hull cases. Consequently, the approach is applicable to both the hard ε-tube and soft ε-tube cases of regression problems. We can therefore adapt the NPA to classification and regression SVMs. The resulting algorithm is efficient, easy to implement, and requires no optimization solvers.

The remainder of this paper is organized as follows. In Section 2, we examine when the ε-tubes exist using theorems of alternative and duality analysis. Section 3 is dedicated to the construction of hard ε-tubes about data by the regression models produced by separating the convex hulls of shifted data. We present how to construct the soft ε-tubes if ε is not large enough to make a hard tube exist in Section 4. Following that, we kernelize our approaches, H-SVR and RH-SVR, in Section 5. Then some further characteristics about RH-SVR are discussed in Section 6. Section 7 explores the generalization performance of the RH-SVR. A generic nearest-point algorithm is proposed and analyzed in Section 8 for both the hard-tube and soft-tube cases in regression. Experimental studies for demonstration of RH-SVR behaviors and comparison with ε-SVR and ν-SVR are included in Section 9. The last section summarizes the conclusions and discusses potential extensions.

2. When does the ε-tube exist?

SVR constructs a regression model that minimizes some empirical risk measure regularized to control capacity. Let $\mathbf{x}$ be the n predictor variables and y the dependent response variable. In [21], Vapnik proposed using the ε-insensitive loss function $L^{\varepsilon}(\mathbf{x}, y, f) = |y - f(\mathbf{x})|_{\varepsilon} = \max(0, |y - f(\mathbf{x})| - \varepsilon)$, in which an example is an error if its residual $|y - f(\mathbf{x})|$ is greater than ε. Plotting the points in $(\mathbf{x}, y)$ space as in Fig. 3(a), we see that for a "perfect" regression model the data fall in a hard ε-tube about the regression line. Let $(\mathbf{x}_i, y_i)$ be an example where $\mathbf{x}_i \in \mathbb{R}^n$ is the ith predictor vector, y_i is the corresponding response, and $i = 1, 2, \ldots, \ell$. The training data are then $(\mathbf{X}, \mathbf{y})$ where the matrix $\mathbf{X} = (\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_{\ell})' \in \mathbb{R}^{\ell \times n}$ and $\mathbf{y} \in R^{\ell}$ is the response. A hard ε-tube for a fixed $\varepsilon > 0$ is defined as a plane $y = \mathbf{w}'\mathbf{x} + b$ satisfying $-\varepsilon\mathbf{e} \leqslant \mathbf{y} - \mathbf{X}\mathbf{w} - b\mathbf{e} \leqslant \varepsilon\mathbf{e}$ where $\mathbf{e}$ is an ℓ-dimensional vector of ones. Vectors and matrices are denoted as bold lower-case and capital letters, respectively, and $\mathbf{x}'$ or $\mathbf{X}'$ is the transpose of a vector $\mathbf{x}$ or a matrix $\mathbf{X}$.

When does a hard ε-tube exist? Clearly, for ε large enough such a tube always exists for finite data. The smallest hard tube, the ε_0-tube, can be found by optimizing:

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\varepsilon} & \varepsilon \\ \text{s.t.} & -\varepsilon\mathbf{e} \leqslant \mathbf{y} - \mathbf{X}\mathbf{w} - b\mathbf{e} \leqslant \varepsilon\mathbf{e}. \end{array} \tag{1}$$

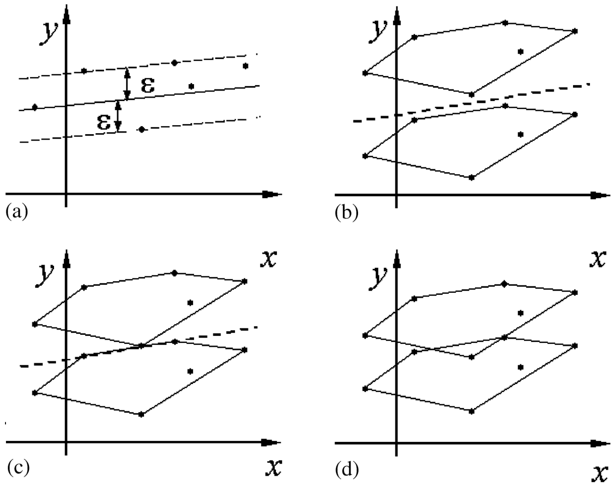

Fig. 3. The (a) primal hard ε_0-tube, and dual cases: (b) dual strictly separable $\varepsilon > \varepsilon_0$, (c) dual separable $\varepsilon = \varepsilon_0$, and (d) dual inseparable $\varepsilon < \varepsilon_0$.

Note that the smallest tube is typically not the ε-SVR solution. Let $\mathscr{D}^+$ and $\mathscr{D}^-$ be formed by augmenting the data with the response variable respectively increased and decreased by ε, i.e., $\mathscr{D}^+ = \{\begin{pmatrix} \mathbf{x}_i \\ y_i+\varepsilon \end{pmatrix}, i = 1,\ldots,\ell\}$ and $\mathscr{D}^- = \{\begin{pmatrix} \mathbf{x}_i \\ y_i-\varepsilon \end{pmatrix}, i = 1,\ldots,\ell\}$. The vector $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix}$ is in the augmented space. Consider the simple problem in Fig. 3(a). For any fixed $\varepsilon > 0$, there are three possible cases: $\varepsilon > \varepsilon_0$ in which strict hard ε-tubes exist, $\varepsilon = \varepsilon_0$ in which only ε_0-tubes exist, and $\varepsilon < \varepsilon_0$ in which no hard ε-tubes exist. A strict hard ε-tube with no points on the edges of the tube only exists for $\varepsilon > \varepsilon_0$. Fig. 3(b–d) illustrates what happens in the dual space for each case. The convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ are drawn along with the max-margin plane in (b) and the supporting plane in (c) for separating the convex hulls.

Clearly, the existence of the tube is directly related to the separability of $\mathscr{D}^+$ and $\mathscr{D}^-$. If $\varepsilon > \varepsilon_0$, then a strict hard tube exists and the convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ are strictly separable. We use the following definitions of separation of convex sets.

Definition 1 (Separable and strictly separable convex sets; Bazaraa et al. [3]). Let $\mathscr{U}$ and $\mathscr{V}$ be nonempty convex sets. A plane $H = \{x : \mathbf{w}'\mathbf{x} = \alpha\}$ is said to separate $\mathscr{U}$ and $\mathscr{V}$ if $\mathbf{w}'\mathbf{x} \geqslant \alpha, \forall \mathbf{x} \in \mathscr{U}$ and $\mathbf{w}'\mathbf{x} \leqslant \alpha, \forall \mathbf{x} \in \mathscr{V}$. H is said to strictly separate $\mathscr{U}$ and $\mathscr{V}$ if $\mathbf{w}'\mathbf{x} \geqslant \alpha + \Delta$ for $\mathbf{x} \in \mathscr{U}$, and $\mathbf{w}'\mathbf{x} \leqslant \alpha - \Delta$ for each $\mathbf{x} \in \mathscr{V}$ where Δ is a positive scalar.

There are infinitely many possible ε-tubes when $\varepsilon > \varepsilon_0$. One can see that the max-margin plane separating $\mathscr{D}^+$ and $\mathscr{D}^-$ corresponds to one such ε-tube. In fact, this plane forms an $\hat{\varepsilon}$-tube where $\varepsilon > \hat{\varepsilon} \geqslant \varepsilon_0$. If $\varepsilon = \varepsilon_0$, then the convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ are separable but not strictly separable. The plane that separates the two convex hulls forms the ε_0-tube. In the last case, where $\varepsilon < \varepsilon_0$, the convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ intersect. No hard ε-tubes or max-margin planes exist.

It is easy to show by construction that if a hard ε-tube exists for a given $\varepsilon > 0$ then the convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ will be separable. If a hard ε-tube exists, then there exists $(\mathbf{w}, b)$ such that

$$\begin{aligned} &\mathbf{Xw} + b\mathbf{e} \leqslant \mathbf{y} + \varepsilon\mathbf{e}, \\ &-\mathbf{Xw} - b\mathbf{e} \leqslant -\mathbf{y} + \varepsilon\mathbf{e}. \end{aligned} \tag{2}$$

Gale's Theorem of the alternative can be used to precisely characterize the ε-tube. By Gale's Theorem [15], the system $\mathbf{Ax} \leqslant \mathbf{c}$ has a (or has no) solution if and only if the alternative system $\mathbf{A}'\mathbf{y} = 0$, $\mathbf{c}'\mathbf{y} = -1$, $\mathbf{y} \geqslant 0$ has no (or has a) solution.

Theorem 2 (Conditions for existence of hard ε-tube). *A hard ε-tube exists for a given $\varepsilon > 0$ if and only if the following system in $(\mathbf{u}, \mathbf{v})$ has no solution*:

$$\begin{aligned} &(\mathbf{y} + \varepsilon\mathbf{e})'\mathbf{u} - (\mathbf{y} - \varepsilon\mathbf{e})'\mathbf{v} < 0, \\ &\mathbf{X}'\mathbf{u} = \mathbf{X}'\mathbf{v}, \quad \mathbf{e}'\mathbf{u} = \mathbf{e}'\mathbf{v} = 1, \quad \mathbf{u} \geqslant 0, \quad \mathbf{v} \geqslant 0. \end{aligned} \tag{3}$$

Proof. A hard ε-tube exists if and only if system (2) has a solution. By Gale's Theorem of the alternative [15], System (2) has a solution if and only if the following alternative system has no solution:

$$\begin{aligned} &\mathbf{X}'\mathbf{u} = \mathbf{X}'\mathbf{v}, \quad (\mathbf{y} + \varepsilon\mathbf{e})'\mathbf{u} - (\mathbf{y} - \varepsilon\mathbf{e})'\mathbf{v} = -1, \\ &\mathbf{e}'\mathbf{u} = \mathbf{e}'\mathbf{v}, \quad \mathbf{u} \geqslant 0, \quad \mathbf{v} \geqslant 0. \end{aligned}$$

Rescaling by $1/\sigma$ where $\sigma = \mathbf{e}'\mathbf{u} = \mathbf{e}'\mathbf{v} > 0$ yields the result. □

Examining the geometry of the alternative system can give us insight into the problem. Note that if ε is large enough, $\varepsilon \geqslant \varepsilon_0$, then $(\mathbf{y} + \varepsilon\mathbf{e})'\mathbf{u} - (\mathbf{y} - \varepsilon\mathbf{e})'\mathbf{v} \geqslant 0$ for any $(\mathbf{u}, \mathbf{v})$ satisfying $\mathbf{X}'\mathbf{u} = \mathbf{X}'\mathbf{v}, \mathbf{e}'\mathbf{u} = \mathbf{e}'\mathbf{v} = 1$, $\mathbf{u}, \mathbf{v} \geqslant 0$. Clearly the alternative system investigates the convex combinations $\begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}+\varepsilon\mathbf{e})' \end{pmatrix}\mathbf{u}$ and $\begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}-\varepsilon\mathbf{e})' \end{pmatrix}\mathbf{v}$, and thus the system explores the property of convex hulls of shifted data $\mathscr{D}^+$ and $\mathscr{D}^-$. The relation between existence of a hard tube and separability of two convex hulls is summarized in Corollary 3.

Corollary 3 (Relation between existence of hard ε-tube and separability). *A hard ε-tube exists for a given $\varepsilon > 0$ if and only if the convex hull of points $\begin{pmatrix} \mathbf{x}_i \\ y_i+\varepsilon \end{pmatrix}$, $i=1,\ldots,\ell$, and the convex hull of points $\begin{pmatrix} \mathbf{x}_i \\ y_i-\varepsilon \end{pmatrix}$, $i = 1,\ldots,\ell$ are separable.*

Proof. A hard ε-tube exists if and only if system (2) has a solution $(\mathbf{w}, b)$. Then for any convex combination $\begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}+\varepsilon\mathbf{e})' \end{pmatrix}\mathbf{u}$ of points $\begin{pmatrix} \mathbf{x}_i \\ y_i+\varepsilon \end{pmatrix}$, $i=1,2,\ldots,\ell$ in $\mathscr{D}^+$ where $\mathbf{e}'\mathbf{u}=1$, $\mathbf{u} \geqslant 0$, we have $(\mathbf{y}+\varepsilon\mathbf{e})'\mathbf{u}-\mathbf{w}'(\mathbf{X}'\mathbf{u})-b \geqslant 0$. Similarly for $\mathscr{D}^-$, any convex combination $\begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}-\varepsilon\mathbf{e})' \end{pmatrix}\mathbf{v}$ of points $\begin{pmatrix} \mathbf{x}_i \\ y_i-\varepsilon \end{pmatrix}$, $i = 1,2,\ldots,\ell$, satisfies $(\mathbf{y} - \varepsilon\mathbf{e})'\mathbf{v} - \mathbf{w}'(\mathbf{X}'\mathbf{v}) - b \leqslant 0$. Then by Definition 1, the plane $y - \mathbf{w}'\mathbf{x} - b = 0$, i.e., $y = \mathbf{w}'\mathbf{x} + b$ in the middle of the ε-tube separates the two convex hulls. Note the separating plane and the plane constructing the ε-tube are the same. If no separating plane exists, then there is no tube. So as a consequence of this theorem, if $\mathscr{D}^+$ and $\mathscr{D}^-$ are separable, then a hard ε-tube exists. □

3. Constructing the hard ε-tube

For any $\varepsilon > \varepsilon_0$ infinitely many possible ε-tubes exist. Which ε-tube should be used? The linear program (1) can be solved to find the smallest ε_0-tube, but this corresponds to just doing empirical risk minimization and may result in poor generalization due to overfitting. We know capacity control or structural risk minimization is fundamental to the success of SVM classification and regression.

We take our inspiration from SVM classification. In hard-margin SVM classification, the dual SVM formulation constructs the max-margin plane by finding the two closest points in the convex hulls of the two classes. The max-margin plane is the plane bisecting these two points. We know that the existence of the tube in primal space is linked to the separability of the shifted sets, $\mathscr{D}^+$ and $\mathscr{D}^-$, in dual space. The key insight is that the regression problem can be regarded as a classification problem between $\mathscr{D}^+$ and $\mathscr{D}^-$. The two sets $\mathscr{D}^+$ and $\mathscr{D}^-$ defined as in Section 2 both contain the same number of data points. The only significant difference occurs along the y dimension as the response variable y is shifted up by ε in $\mathscr{D}^+$ and down by ε in $\mathscr{D}^-$. For $\varepsilon > \varepsilon_0$, the max-margin separating plane corresponds to a hard $\hat{\varepsilon}$-tube where $\varepsilon > \hat{\varepsilon} \geqslant \varepsilon_0$. The resulting tube is smaller than ε but not necessarily the smallest tube. Fig. 3(b) shows the max-margin plane found for $\varepsilon > \varepsilon_0$. Fig. 3(a) shows that the corresponding linear regression function for this simple example turns out to be the ε_0 tube. As in classification, we will have the hard and soft ε-tube cases. The soft ε-tube with $\varepsilon \leqslant \varepsilon_0$ is used to obtain good generalization when there are outliers.

We now apply the dual convex hull method to constructing the max-margin plane for our augmented sets $\mathscr{D}^+$ and $\mathscr{D}^-$ assuming they are strictly separable, i.e. $\varepsilon > \varepsilon_0$. The problem is illustrated in detail in Fig. 4. The convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$

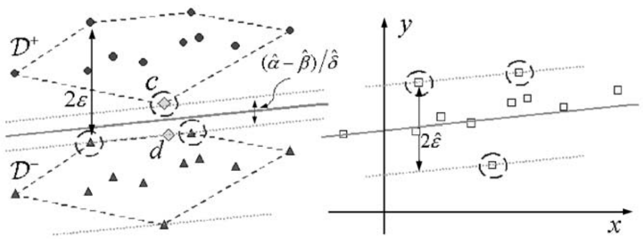

Fig. 4. The solution $\hat{\varepsilon}$-tube found by H-SVR can have $\hat{\varepsilon} < \varepsilon$. Left: Dots are in $\mathscr{D}^+$. Triangles are in $\mathscr{D}^-$. Support vectors are circled. Right: Squares are original data.

are defined as

$$conv(\mathscr{D}^+) = \left\{ \sum_{\mathbf{z}_i^+ \in \mathscr{D}^+} u_i \mathbf{z}_i^+ \;\middle|\; \sum u_i = 1, u_i \geqslant 0 \right\},$$
$$conv(\mathscr{D}^-) = \left\{ \sum_{\mathbf{z}_i^- \in \mathscr{D}^-} v_i \mathbf{z}_i^- \;\middle|\; \sum v_i = 1, v_i \geqslant 0 \right\}. \tag{4}$$

The closest points of the convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ can be found by solving the following dual H-SVR quadratic program in $(\mathbf{u}, \mathbf{v})$:

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v}} \quad & \frac{1}{2} \left\| \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}+\varepsilon\mathbf{e})' \end{pmatrix} \mathbf{u} - \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}-\varepsilon\mathbf{e})' \end{pmatrix} \mathbf{v} \right\|^2 \\ \text{s.t.} \quad & \mathbf{e}'\mathbf{u} = 1, \quad \mathbf{e}'\mathbf{v} = 1, \\ & \mathbf{u} \geqslant 0, \quad \mathbf{v} \geqslant 0. \end{aligned} \tag{5}$$

Let the closest points in the convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ be $\mathbf{c} = \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}+\varepsilon\mathbf{e})' \end{pmatrix} \hat{\mathbf{u}}$ and $\mathbf{d} = \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}-\varepsilon\mathbf{e})' \end{pmatrix} \hat{\mathbf{v}}$, respectively. The max-margin separating plane bisects these two points. The normal $(\hat{\mathbf{w}}, \hat{\delta})$ of the plane is the difference between them, i.e., $\hat{\mathbf{w}} = \mathbf{X}'\hat{\mathbf{u}} - \mathbf{X}'\hat{\mathbf{v}}$, $\hat{\delta} = (\mathbf{y}+\varepsilon\mathbf{e})'\hat{\mathbf{u}} - (\mathbf{y}-\varepsilon\mathbf{e})'\hat{\mathbf{v}}$. The threshold, $\hat{b}$, is the distance from the origin to the point halfway between the two closest points along the normal: $\hat{b} = \hat{\mathbf{w}}'((\mathbf{X}'\hat{\mathbf{u}} + \mathbf{X}'\hat{\mathbf{v}})/2) + \hat{\delta}((\mathbf{y}'\hat{\mathbf{u}} + \mathbf{y}'\hat{\mathbf{v}})/2)$. The separating plane has the equation $\hat{\mathbf{w}}'\mathbf{x} + \hat{\delta} y - \hat{b} = 0$. Rescaling this plane by $-\hat{\delta}$ yields the regression function.

Dual H-SVR (5) is in the dual space. The corresponding Primal H-SVR is

$$\begin{aligned} \min_{\mathbf{w},\delta,\alpha,\beta} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + \tfrac{1}{2}\delta^2 - (\alpha - \beta) \\ \text{s.t.} \quad & \mathbf{X}\mathbf{w} + \delta(\mathbf{y} + \varepsilon\mathbf{e}) - \alpha\mathbf{e} \geqslant 0, \\ & \mathbf{X}\mathbf{w} + \delta(\mathbf{y} - \varepsilon\mathbf{e}) - \beta\mathbf{e} \leqslant 0. \end{aligned} \tag{6}$$

Dual H-SVR (5) can be derived by taking the Wolfe or Lagrangian dual [15] of Primal H-SVR (6) and simplifying. In addition, by the KKT conditions, the primal solutions $\hat{\mathbf{w}}$ and $\hat{\delta}$ are exactly the ones defined above as $\begin{pmatrix} \hat{\mathbf{w}} \\ \hat{\delta} \end{pmatrix} = \mathbf{c} - \mathbf{d}$. The primal formulation works in the regression space. Once the optimal solution $(\hat{\mathbf{w}}, \hat{\delta}, \hat{\alpha}, \hat{\beta})$ is obtained, we have the regression model.

It has been shown that the optimal plane from H-SVR bisects the $\hat{\varepsilon}$-tube [5]. The boundary planes of the margin for class $\mathscr{D}^+$ and class $\mathscr{D}^-$ determine the lower and upper edges of the $\hat{\varepsilon}$-tube, respectively. The support vectors from $\mathscr{D}^+$ and $\mathscr{D}^-$ correspond to the points along the lower and upper edges of the $\hat{\varepsilon}$-tube as in Fig. 4. The following theorem shows that maximizing the margin corresponds to reducing the effective tube size.

Theorem 4 (H-SVR constructs $\hat{\varepsilon}$-tube; Bi and Bennett [5]). *Let the max-margin plane obtained by H-SVR (5) be* $\hat{\mathbf{w}}'\mathbf{x} + \hat{\delta} y - \hat{b} = 0$ *where* $\hat{\mathbf{w}} = \mathbf{X}'\hat{\mathbf{u}} - \mathbf{X}'\hat{\mathbf{v}}$, $\hat{\delta} = (\mathbf{y}+\varepsilon\mathbf{e})'\hat{\mathbf{u}} - (\mathbf{y}-\varepsilon\mathbf{e})'\hat{\mathbf{v}}$, *and* $\hat{b} = \hat{\mathbf{w}}'((\mathbf{X}'\hat{\mathbf{u}} + \mathbf{X}'\hat{\mathbf{v}})/2) + \hat{\delta}((\mathbf{y}'\hat{\mathbf{u}} + \mathbf{y}'\hat{\mathbf{v}})/2)$. *If* $\varepsilon > \varepsilon_0$, *then the plane* $y = \bar{\mathbf{w}}'\mathbf{x} + \bar{b}$

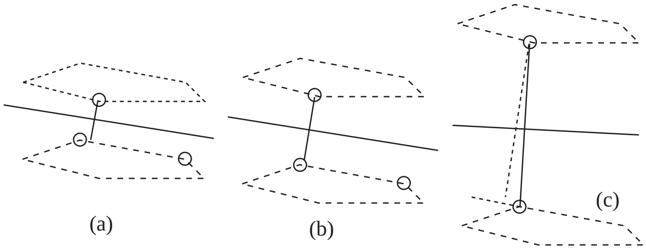

Fig. 5. In (a) and (b), the regression model remains the same for different values of ε. The regression model changes in (c) when ε becomes too large. Support vectors are circled.

corresponds to an $\hat{\varepsilon}$-tube of the training data $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, \ell$ where $\bar{\mathbf{w}} = -\hat{\mathbf{w}}/\hat{\delta}$, $\bar{b} = \hat{b}/\hat{\delta}$ and $\hat{\varepsilon} = \varepsilon - (\hat{\alpha} - \hat{\beta})/(2\hat{\delta}) < \varepsilon$.

The effective tube size is reduced by the size of the maximal margin. The regression model and the effective $\hat{\varepsilon}$-tube defined by this model depend on the choice of ε. However, within a certain range of ε, they will remain the same as shown in Figs. 5(a) and (b). When ε becomes too large as in Fig. 5(c), support vectors and the corresponding $(\hat{\mathbf{w}}, \hat{\delta})$ will change, and so will the regression model.

4. Constructing the soft ε-tube

In real-life applications, outliers almost always exist. Making ε large to fit outliers may result in poor overall accuracy. For $\varepsilon < \varepsilon_0$, a hard ε-tube does not exist. In soft-margin classification, outliers were handled in the dual space by using reduced convex hulls. Reduced convex hulls limit the influence of any given point by reducing the upper bound on the multiplier for each point to $D < 1$. The same strategy works for soft ε-tubes. Formally the reduced convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ are defined as

$$\mathscr{U} = \left\{ \sum_{\mathbf{z}_i^+ \in \mathscr{D}^+} u_i \mathbf{z}_i^+ \,\middle|\, \sum u_i = 1, 0 \leqslant u_i \leqslant D \right\},$$

$$\mathscr{V} = \left\{ \sum_{\mathbf{z}_i^- \in \mathscr{D}^-} v_i \mathbf{z}_i^- \,\middle|\, \sum v_i = 1, 0 \leqslant v_i \leqslant D \right\}. \tag{7}$$

See Fig. 6 for an example with $D = 1/2$. Instead of taking the full convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$, we reduce the convex hulls away from the difficult boundary cases. RH-SVR computes the closest points in the reduced convex hulls of the shifted data points of $\mathbf{z}_i = \binom{\mathbf{x}_i}{y_i}$ along the y dimension respectively up and down by ε. Then the RH-SVR formulation becomes

$$\begin{array}{ll} \min\limits_{\mathbf{u},\mathbf{v}} & \dfrac{1}{2} \left\| \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}+\varepsilon\mathbf{e})' \end{pmatrix} \mathbf{u} - \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}-\varepsilon\mathbf{e})' \end{pmatrix} \mathbf{v} \right\|^2 \\ \text{s.t.} & \mathbf{e}'\mathbf{u} = 1, \quad \mathbf{e}'\mathbf{v} = 1, \\ & 0 \leqslant \mathbf{u} \leqslant D\mathbf{e}, \quad 0 \leqslant \mathbf{v} \leqslant D\mathbf{e}. \end{array} \tag{8}$$

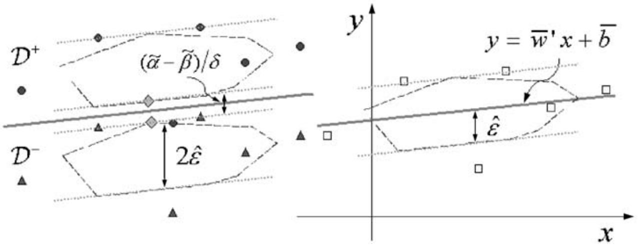

Fig. 6. Soft $\hat\varepsilon$-tube found by RH-SVR: left: dual, right: primal space.

If $D<1$, Eq. (7) defines the reduced convex hulls; if $D\geqslant 1$, it becomes equivalent to defining the convex hulls because if the summation of nonnegative coefficients equals 1, each of the coefficients cannot be larger than 1. The parameter D hence determines whether or not the convex hulls are reduced and the extent of the reduction. We consider Eq. (7) as both cases and distinguish them only by different D values.

Now let us derive the same result as in Theorem 4 for the more general formulation (8). As in the separable case, the dual RH-SVR problem (8) corresponds to computing a soft ε-tube in the primal regression space by duality theorems. The following soft-tube version of Primal H-SVR (9) has RH-SVR (8) as its Wolfe Dual:

$$\begin{aligned}
\min_{\mathbf{w},\delta,\alpha,\beta,\boldsymbol{\xi},\boldsymbol{\eta}}\quad & \tfrac{1}{2}\|\mathbf{w}\|^2+\tfrac{1}{2}\delta^2-(\alpha-\beta)+D(\mathbf{e}'\boldsymbol{\xi}+\mathbf{e}'\boldsymbol{\eta})\\
\text{s.t.}\quad & \mathbf{X}\mathbf{w}+\delta(\mathbf{y}+\varepsilon\mathbf{e})-\alpha\mathbf{e}+\boldsymbol{\xi}\geqslant 0,\quad \boldsymbol{\xi}\geqslant 0,\\
& \mathbf{X}\mathbf{w}+\delta(\mathbf{y}-\varepsilon\mathbf{e})-\beta\mathbf{e}-\boldsymbol{\eta}\leqslant 0,\quad \boldsymbol{\eta}\geqslant 0.
\end{aligned}\tag{9}$$

As in the SVM classification case, we assume D is sufficiently small such that the two convex sets are strictly separable. RH-SVR separates them by bisecting the closest points in each of the sets. The next theorem proves that after rescaling, the separating surface is exactly the regression function and the effective tube size is ε reduced by the margin of separation. This result is proved using the Separation Theorem A.2 in the appendix.

Theorem 5 (RH-SVR constructs soft $\hat\varepsilon$-tube). *Let the soft max-margin plane obtained by RH-SVR (8) be* $\hat{\mathbf{w}}'\mathbf{x}+\hat\delta y-\hat b=0$ *where* $\hat{\mathbf{w}}=\mathbf{X}'\hat{\mathbf{u}}-\mathbf{X}'\hat{\mathbf{v}}$, $\hat\delta=(\mathbf{y}+\varepsilon\mathbf{e})'\hat{\mathbf{u}}-(\mathbf{y}-\varepsilon\mathbf{e})'\hat{\mathbf{v}}$, *and* $\hat b=\hat{\mathbf{w}}'((\mathbf{X}'\hat{\mathbf{u}}+\mathbf{X}'\hat{\mathbf{v}})/2)+\hat\delta((\mathbf{y}'\hat{\mathbf{u}}+\mathbf{y}'\hat{\mathbf{v}})/2)$. *Then the plane* $y=\bar{\mathbf{w}}'\mathbf{x}+\bar b$ *corresponds to a soft* $\hat\varepsilon=\varepsilon-(\tilde\alpha-\tilde\beta)/(2\hat\delta)<\varepsilon$ *tube of training data* $\mathbf{z}_i=\binom{\mathbf{x}_i}{y_i}$, $i=1,2,\ldots,\ell$, *or equivalently, an* $\hat\varepsilon$*-tube about the convex set* $\mathscr{S}=\left\{\sum_{\mathbf{z}_i\in\mathscr{D}} s_i\mathbf{z}_i \mid \sum s_i=1, 0\leqslant s_i\leqslant D\right\}$ *where* $\bar{\mathbf{w}}=-\hat{\mathbf{w}}/\hat\delta$, $\bar b=\hat b/\hat\delta$ *and* $\tilde\alpha=\hat{\mathbf{w}}'\mathbf{X}'\hat{\mathbf{u}}+\hat\delta(\mathbf{y}+\varepsilon\mathbf{e})'\hat{\mathbf{u}}$, $\tilde\beta=\hat{\mathbf{w}}'\mathbf{X}'\hat{\mathbf{v}}+\hat\delta(\mathbf{y}-\varepsilon\mathbf{e})'\hat{\mathbf{v}}$.

Proof. First we prove that $\hat\delta>0$. Let the closest points be $\mathbf{c}=\begin{pmatrix}\mathbf{X}'\\(\mathbf{y}+\varepsilon\mathbf{e})'\end{pmatrix}\hat{\mathbf{u}}\in\mathscr{U}$ and $\mathbf{d}=\begin{pmatrix}\mathbf{X}'\\(\mathbf{y}-\varepsilon\mathbf{e})'\end{pmatrix}\hat{\mathbf{v}}\in\mathscr{V}$. By the KKT conditions, $\binom{\hat{\mathbf{w}}}{\hat\delta}=\mathbf{c}-\mathbf{d}$. For any point $\mathbf{z}=\binom{\mathbf{x}}{y}$ in the reduced convex hull $\mathscr{S}$ of training data, $\mathbf{z}^+=\binom{\mathbf{x}}{y+\varepsilon}\in\mathscr{U}$, and $\mathbf{z}^-=\binom{\mathbf{x}}{y-\varepsilon}\in\mathscr{V}$.

By the Separation Theorem A.2 in the appendix,

$$\begin{aligned}&(\mathbf{c}-\mathbf{d})'\mathbf{z}^{+} = \hat{\mathbf{w}}'\mathbf{x} + \hat{\delta}(y+\varepsilon) \geqslant (\mathbf{c}-\mathbf{d})'\mathbf{c},\\ &(\mathbf{c}-\mathbf{d})'\mathbf{z}^{-} = \hat{\mathbf{w}}'\mathbf{x} + \hat{\delta}(y-\varepsilon) \leqslant (\mathbf{c}-\mathbf{d})'\mathbf{d}.\end{aligned} \tag{10}$$

Define $\tilde{\alpha}=(\mathbf{c}-\mathbf{d})'\mathbf{c}=\hat{\mathbf{w}}'\mathbf{X}'\hat{\mathbf{u}}+\hat{\delta}(\mathbf{y}+\varepsilon\mathbf{e})'\hat{\mathbf{u}}$ and $\tilde{\beta}=(\mathbf{c}-\mathbf{d})'\mathbf{d}=\hat{\mathbf{w}}'\mathbf{X}'\hat{\mathbf{v}}+\hat{\delta}(\mathbf{y}-\varepsilon\mathbf{e})'\hat{\mathbf{v}}$. Then $\tilde{\alpha}-\tilde{\beta}=\|\mathbf{c}-\mathbf{d}\|^2>0$. Subtract the second inequality from the first inequality: $2\hat{\delta}\varepsilon \geqslant \tilde{\alpha}-\tilde{\beta}$, that is, $\hat{\delta} \geqslant (\tilde{\alpha}-\tilde{\beta})/(2\varepsilon) > 0$. Rescale the inequalities by $-\hat{\delta}<0$, reverse the signs, and let $\bar{\mathbf{w}}=-\hat{\mathbf{w}}/\hat{\delta}$. Inequalities (10) become $\bar{\mathbf{w}}'\mathbf{x}-y \leqslant \varepsilon-\tilde{\alpha}/\hat{\delta}, \bar{\mathbf{w}}'\mathbf{x}-y \geqslant -\varepsilon-\tilde{\beta}/\hat{\delta}$. Let $\bar{b}=\hat{b}/\hat{\delta}=(\tilde{\alpha}+\tilde{\beta})/(2\hat{\delta})$. Substituting $\bar{b}$ into the inequalities yields

$$\bar{\mathbf{w}}'\mathbf{x}+\bar{b}-y \leqslant \left(\varepsilon-\frac{\tilde{\alpha}-\tilde{\beta}}{2\hat{\delta}}\right), \quad \bar{\mathbf{w}}'\mathbf{x}+\bar{b}-y \geqslant -\left(\varepsilon-\frac{\tilde{\alpha}-\tilde{\beta}}{2\hat{\delta}}\right).$$

Denoting $\hat{\varepsilon}=\varepsilon-(\tilde{\alpha}-\tilde{\beta})/(2\hat{\delta})$, we end up with $\bar{\mathbf{w}}'\mathbf{x}+\bar{b}-y \leqslant \hat{\varepsilon}$, $\bar{\mathbf{w}}'\mathbf{x}+\bar{b}-y \geqslant -\hat{\varepsilon}$, for any $(\mathbf{x},y)$ in the reduced convex hull $\mathscr{S}$ of training data. Hence the plane $y=\bar{\mathbf{w}}'\mathbf{x}+\bar{b}$ is in the middle of the $\hat{\varepsilon}=\varepsilon-(\tilde{\alpha}-\tilde{\beta})/(2\hat{\delta})<\varepsilon$ tube. □

Notice that Theorem 4 is the special case of this theorem when $D=1$. The thresholds $\tilde{\alpha}$ and $\tilde{\beta}$ determine the planes parallel to the regression plane and through the closest points. In the separable case where $D=1$, $\tilde{\alpha}$ and $\tilde{\beta}$ are optimal for the primal H-SVR (6). However, in the inseparable case, these planes are parallel but not necessarily identical to the planes obtained by Primal RH-SVR (9).

5. Kernelizing H-SVR and RH-SVR

In this section, we kernelize H-SVR and RH-SVR. The objective functions of H-SVR (5) and RH-SVR (8) can be rewritten as

$$\tfrac{1}{2}(\mathbf{u}-\mathbf{v})'(\mathbf{X}\mathbf{X}'+\mathbf{y}\mathbf{y}')(\mathbf{u}-\mathbf{v})+2\varepsilon\mathbf{y}'(\mathbf{u}-\mathbf{v})+2\varepsilon^2. \tag{11}$$

The matrix $\mathbf{X}\mathbf{X}'$ is induced by the inner product in the vector space $\mathbb{R}^n$. Once an algorithm can be cast in terms of inner products, it is possible to substitute a kernel k for the inner product. By computing a kernel function, we implicitly map the input vector $\mathbf{x}$ to $\Phi(\mathbf{x})$ in a (usually higher dimensional) feature space which has the structure of a reproducing kernel Hilbert space through the nonlinear mapping $\Phi(\cdot)$. The mapping operator $\Phi(\cdot)$ is a vector of d functions such that $k(\mathbf{x}_i,\mathbf{x}_j)=\Phi(\mathbf{x}_i)'\Phi(\mathbf{x}_j)$, where d could be infinite. Nonlinear H-SVR and RH-SVR can then be constructed by minimizing the objective function (11) with the inner product matrix $\mathbf{X}\mathbf{X}'$ replaced by the kernel matrix $\mathbf{K}$ whose ijth entry is $k(\mathbf{x}_i,\mathbf{x}_j)$. Define $\Phi(\mathbf{X})=(\Phi(\mathbf{x}_1)\Phi(\mathbf{x}_2)\cdots\Phi(\mathbf{x}_\ell))'$, then $\mathbf{K}=\Phi(\mathbf{X})\Phi(\mathbf{X})'$. The dual reduced convex hull problem in feature space becomes

$$\begin{aligned}\min_{\mathbf{u},\mathbf{v}} \quad & \tfrac{1}{2}(\mathbf{u}-\mathbf{v})'(\mathbf{K}+\mathbf{y}\mathbf{y}')(\mathbf{u}-\mathbf{v})+2\varepsilon\mathbf{y}'(\mathbf{u}-\mathbf{v})\\ \text{s.t.} \quad & \mathbf{e}'\mathbf{u}=1, \quad \mathbf{e}'\mathbf{v}=1,\\ & 0 \leqslant \mathbf{u} \leqslant D\mathbf{e}, \quad 0 \leqslant \mathbf{v} \leqslant D\mathbf{e},\end{aligned} \tag{12}$$

where the objective function is equivalent to

$$\frac{1}{2}\left\|\begin{pmatrix}\Phi(\mathbf{X})'\\(\mathbf{y}+\varepsilon\mathbf{e})'\end{pmatrix}\mathbf{u}-\begin{pmatrix}\Phi(\mathbf{X})'\\(\mathbf{y}-\varepsilon\mathbf{e})'\end{pmatrix}\mathbf{v}\right\|^2$$

after adding the constant $2\varepsilon^2$. In the dual classification space, the response y serves as one more dimension of the attribute space. If $\hat{k}(\mathbf{x}_i,\mathbf{x}_j)$ is a kernel function in the regression problem, then the function $k\left(\begin{pmatrix}\mathbf{x}_i\\y_i\end{pmatrix},\begin{pmatrix}\mathbf{x}_j\\y_j\end{pmatrix}\right)=\hat{k}(\mathbf{x}_i,\mathbf{x}_j)+y_iy_j$ is a kernel [7] in the dual augmented space that maps $\mathbf{z}\mapsto\Phi(\mathbf{z})=\begin{pmatrix}\Phi(\mathbf{x})\\y\end{pmatrix}$. Formulation (12) actually computes the closest points, respectively, in the convex sets

$$\mathscr{U}=\left\{\sum_{\mathbf{z}_i^+\in\mathscr{D}^+}u_i\Phi(\mathbf{z}_i^+)\;\middle|\;\sum u_i=1, 0\leqslant u_i\leqslant D\right\},$$

$$\mathscr{V}=\left\{\sum_{\mathbf{z}_i^-\in\mathscr{D}^-}v_i\Phi(\mathbf{z}_i^-)\;\middle|\;\sum v_i=1, 0\leqslant v_i\leqslant D\right\}. \tag{13}$$

In primal regression space, bisecting the two closest points corresponds to constructing a linear plane $f(x)=\mathbf{w}'\Phi(\mathbf{x})+b$ in the feature space. The final regression model can be obtained through the following theorem.

Theorem 6 (Construct $\hat{\varepsilon}$-tube in feature space). *If* $(\hat{\mathbf{u}},\hat{\mathbf{v}})$ *is the solution to problem* (12), *then the resulting regression model is* $f(\mathbf{x})=\sum_{i=1}^{\ell}(\bar{v}_i-\bar{u}_i)k(\mathbf{x}_i,\mathbf{x})+\bar{b}$ *which constructs an* $\hat{\varepsilon}$*-tube in feature space, where* $\bar{u}_i=\hat{u}_i/\hat{\delta}$, $\bar{v}_i=\hat{v}_i/\hat{\delta}$, $\hat{\delta}=(\hat{\mathbf{u}}-\hat{\mathbf{v}})'\mathbf{y}+2\varepsilon$, *the intercept term* $\bar{b}=(\hat{\mathbf{u}}-\hat{\mathbf{v}})'\mathbf{K}(\hat{\mathbf{u}}+\hat{\mathbf{v}})/(2\hat{\delta})+(\hat{\mathbf{u}}+\hat{\mathbf{v}})'\mathbf{y}/2$, *and* $\hat{\varepsilon}=-(\hat{\mathbf{u}}-\hat{\mathbf{v}})'\mathbf{K}(\hat{\mathbf{u}}-\hat{\mathbf{v}})/(2\hat{\delta})+(\hat{\mathbf{v}}-\hat{\mathbf{u}})'\mathbf{y}/2$.

Proof. Replacing the matrix $\mathbf{X}$ by $\Phi(\mathbf{X})$ in RH-SVR primal (9) yields the primal formulation to problem (12). Then the primal finds the max-margin plane $\hat{\mathbf{w}}'\Phi(\mathbf{x})+\hat{\delta}y-\hat{b}=0$ in feature space. By the KKT conditions, the primal solution is $\begin{pmatrix}\hat{\mathbf{w}}\\\hat{\delta}\end{pmatrix}=\mathbf{c}-\mathbf{d}$ where $\mathbf{c}=\begin{pmatrix}\Phi(\mathbf{X})'\\(\mathbf{y}+\varepsilon\mathbf{e})'\end{pmatrix}\hat{\mathbf{u}}$ and $\mathbf{d}=\begin{pmatrix}\Phi(\mathbf{X})'\\(\mathbf{y}-\varepsilon\mathbf{e})'\end{pmatrix}\hat{\mathbf{v}}$ are the closest points in the sets of Eq. (13). Following the same line of argument as in Theorem 5, we define $\tilde{\alpha}=(\mathbf{c}-\mathbf{d})'\mathbf{c}$ and $\tilde{\beta}=(\mathbf{c}-\mathbf{d})'\mathbf{d}$, and use Separation Theorem A.2. Then we have

$$\hat{\mathbf{w}}'\Phi(\mathbf{x})+\hat{\delta}(y+\varepsilon)\geqslant\tilde{\alpha},\quad \hat{\mathbf{w}}'\Phi(\mathbf{x})+\hat{\delta}(y-\varepsilon)\leqslant\tilde{\beta}$$

and $\hat{\delta}>0$. Substitute $\hat{\mathbf{w}}=\Phi(\mathbf{X})'(\hat{\mathbf{u}}-\hat{\mathbf{v}})$, and rescale by $-\hat{\delta}$. Let $\bar{b}=\dfrac{\tilde{\alpha}+\tilde{\beta}}{2\hat{\delta}}$, then the inequalities become

$$\sum_{i=1}^{\ell}(\bar{v}_i-\bar{u}_i)k(\mathbf{x}_i,\mathbf{x})+\bar{b}-y\leqslant\left(\varepsilon-\frac{\tilde{\alpha}-\tilde{\beta}}{2\hat{\delta}}\right),$$

$$\sum_{i=1}^{\ell}(\bar{v}_i-\bar{u}_i)k(\mathbf{x}_i,\mathbf{x})+\bar{b}-y\geqslant-\left(\varepsilon-\frac{\tilde{\alpha}-\tilde{\beta}}{2\hat{\delta}}\right).$$

Therefore the resulting regression model $f(\mathbf{x})$ satisfies $|f(\mathbf{x}) - y| \leqslant \hat{\varepsilon}$ for any $(\mathbf{x}, y)$ whose image is in the convex set $\mathscr{S} = \left\{\sum_{\mathbf{z}_i \in \mathscr{D}} s_i \Phi(\mathbf{z}_i) \mid \sum s_i = 1, 0 \leqslant s_i \leqslant D\right\}$. The $f(\mathbf{x})$ constructs an $\hat{\varepsilon}$-tube in feature space where $\hat{\varepsilon}=\varepsilon-(\tilde{\alpha}-\tilde{\beta})/(2\hat{\delta})=-(\hat{\mathbf{u}}-\hat{\mathbf{v}})'\mathbf{K}(\hat{\mathbf{u}}-\hat{\mathbf{v}})/(2\hat{\delta}) + (\hat{\mathbf{v}} - \hat{\mathbf{u}})'\mathbf{y}/2$. □

6. Characterizing RH-SVR

We now investigate some characteristics of RH-SVR. RH-SVR has two parameters, D and ε, which must be selected by the user or by cross-validation. The parameter D determines the robustness of the solution by reducing the convex hull, and thus limits the influence of any single point. As in v-SVR, we can parameterize D by v. Let $D = 1/(\ell v)$, where ℓ is the number of points. Fig. 6 illustrates the case for $\ell = 6$ points, $v = 1/3$, and $D = 1/2$. In this example, every point in the reduced convex hull must depend on at least two data points since $\sum_{i=1}^{\ell} u_i = 1$ and $0 \leqslant u_i \leqslant 1/2$. In general, every point in the reduced convex hull can be written as the convex combination of at least $\lceil 1/D \rceil = \lceil \ell v \rceil$ points. Since these points are exactly the support vectors and there are two reduced convex hulls, $2\lceil \ell v \rceil$ is a lower bound on the number of support vectors in RH-SVR.

Let us denote the max-margin plane obtained by solving problem (12) as the classification function $f^{\mathrm{c}}(\mathbf{x}, y)$. The *functional margin* denoted as ρ is defined as the distance from either of the two boundary planes to the classification plane. The classification function $f^{\mathrm{c}}(\mathbf{x}, y)$ has functional margin $\rho = (\tilde{\alpha} - \tilde{\beta})/2$. After scaling by $-\hat{\delta}$, the same plane is regarded as the regression function $y = f(\mathbf{x})$. Although the regression plane is exactly the same as the classification plane after rescaling, the functional margin scales differently. It is now $(\tilde{\alpha} - \tilde{\beta})/(2\hat{\delta})$. Based on our analysis in Theorem 5 for linear models and Theorem 6 for nonlinear models, the above functional margin is just $\varepsilon - \hat{\varepsilon}$ where $\hat{\varepsilon}$ is the effective size of the tube. The *geometric margin*, however, does not vary with rescaling, so it characterizes the real geometric view of models. The *geometric margin* is defined as the Euclidean distance from either of the boundary planes to the classification plane. It can be calculated by dividing the functional margin by the length of the normal vector to the plane. Recall that the model used in the RH-SVR approach is $y = \bar{\mathbf{w}}'\Phi(\mathbf{x}) + \bar{b}$. So the geometric margin is $(\varepsilon - \hat{\varepsilon})/(\sqrt{\|\bar{\mathbf{w}}\|^2 + 1})$.

We adopt the same definitions of *error* and *margin error* as in [19]. If a point is an *error*, it means the point is misclassified by the decision function $\operatorname{sgn} \circ f^{\mathrm{c}}$, in other words, either $f^{\mathrm{c}}(\mathbf{x}_i, y_i+\varepsilon) < 0$ or $f^{\mathrm{c}}(\mathbf{x}_i, y_i-\varepsilon) > 0$. The *margin errors* are either errors or lie within the margin. The empirical probability of a margin error is the fraction of margin errors in the training data set

$$R^{\rho}_{\mathrm{emp}}[f^{\mathrm{c}}] := \frac{1}{\ell}\,|\{i \mid f^{\mathrm{c}}(\mathbf{x}_i, y_i + \varepsilon) < \rho \quad \text{or} \quad f^{\mathrm{c}}(\mathbf{x}_i, y_i - \varepsilon) > -\rho\}|. \tag{14}$$

At most a fraction v of training points can have coefficients up to the upper bound $1/(\ell v)$ for each of the two reduced convex hulls since $\sum_{i=1}^{\ell} u_i = 1$ and $0 \leqslant u_i \leqslant 1/(\ell v)$. Totally, there will be at most $2v$ fraction of the points with coefficients equal to $1/(\ell v)$.

Suppose $\tilde{\alpha}$ and $\tilde{\beta}$ are primal optimal, i.e., $\tilde{\alpha} = \hat{\alpha}$, $\tilde{\beta} = \hat{\beta}$. By complementarity

$$\begin{aligned} &\hat{\mathbf{u}}'(\mathbf{X}\hat{\mathbf{w}} + \hat{\delta}(\mathbf{y} + \varepsilon\mathbf{e}) - \hat{\alpha}\mathbf{e} + \boldsymbol{\xi}) = 0, \quad (D\mathbf{e} - \mathbf{u})'\boldsymbol{\xi} = 0, \\ &\hat{\mathbf{v}}'(\mathbf{X}\hat{\mathbf{w}} + \hat{\delta}(\mathbf{y} - \varepsilon\mathbf{e}) - \hat{\beta}\mathbf{e} - \boldsymbol{\eta}) = 0, \quad (D\mathbf{e} - \mathbf{v})'\boldsymbol{\eta} = 0, \end{aligned} \tag{15}$$

a point with u_i or $v_i = 1/(\ell v)$ may be a margin error since the associated ξ_i or η_i may be > 0. Therefore at most $2v$ of all points can be margin errors, so $R^{\rho}_{\text{emp}}[f^{\text{c}}]$ is bounded by $2v$.

Hence, $2v$ controls the number of support vectors and the number of margin errors. Moreover, the above argument actually holds true for the upper and lower edges of the tube separately. So v is the upper bound on the fractions of underestimated errors as well as overestimated errors. By choosing v sufficiently large, the inseparable case with $\varepsilon \leqslant \varepsilon_0$ is transformed to a separable case where once again our nearest-points problem is well defined. We can summarize these results in the following proposition.

Proposition 7. *Assume $\mathscr{U} \cap \mathscr{V} = \emptyset$ for RH-SVR. Then the following statements hold*:

(1) $2v$ *is an upper bound on the fraction of points with error greater than* $\hat{\varepsilon}$.
(2) $2v$ *is a lower bound on the fraction of support vectors.*

This proposition also holds for v-SVR under the assumption that the optimal ε is greater than zero—an equivalent condition to the reduced convex hulls not intersecting. Consider the dual v-SVR formulation in [19] expressed using their notation:

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & \tfrac{1}{2}(\alpha - \alpha^*)'\mathbf{K}(\alpha - \alpha^*) + \mathbf{y}'(\alpha - \alpha^*) \\ \text{s.t.} \quad & \mathbf{e}'\alpha = \mathbf{e}'\alpha^*, \quad \mathbf{e}'\alpha + \mathbf{e}'\alpha^* \leqslant Cv, \\ & 0 \leqslant \alpha \leqslant \tfrac{C}{\ell}\mathbf{e}, \leqslant \alpha^* \leqslant \tfrac{C}{\ell}\mathbf{e}. \end{aligned} \tag{16}$$

To see the relationship between RH-SVR and v-SVR, one can rescale the problem by defining $\mathbf{u} = 2\alpha/(Cv)$ and $\mathbf{v} = 2\alpha^*/(Cv)$.

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v}} \quad & \tfrac{1}{2}(\mathbf{u} - \mathbf{v})'\mathbf{K}(\mathbf{u} - \mathbf{v}) + \tfrac{2}{Cv}\mathbf{y}'(\mathbf{u} - \mathbf{v}) \\ \text{s.t.} \quad & \mathbf{e}'\mathbf{u} = \mathbf{e}'\mathbf{v}, \quad \mathbf{e}'\mathbf{u} + \mathbf{e}'\mathbf{v} \leqslant 2, \\ & 0 \leqslant \mathbf{u} \leqslant \tfrac{2}{\ell v}\mathbf{e}, \quad 0 \leqslant \mathbf{v} \leqslant \tfrac{2}{\ell v}\mathbf{e}. \end{aligned} \tag{17}$$

For reasonable choices of C and v, we can assume that at optimality $\mathbf{e}'\mathbf{u} + \mathbf{e}'\mathbf{v} = 2$ and by complementarity $\varepsilon > 0$ for v-SVR. Thus the constraints of v-SVR and RH-SVR are identical for $D = 2/(\ell v)$. The principal difference lies in the objective function. The Hessian of the objective function is $\mathbf{K}$ for v-SVR and $\mathbf{K} + \mathbf{yy}'$ for RH-SVR. Thus v-SVR does not find the closest points of the reduced convex hulls of the augmented datasets. In some sense the v-SVR method is regularized in the feature space, whereas RH-SVR is regularized in the feature and the response space. It is an open question as to the effect of this extra regularization. The parameter v in v-SVR is the same as $2v$ in RH-SVR. By comparing the objectives of v-SVR (17) and RH-SVR (12), C should be chosen in ν-SVR so that $2/(C\nu)$ is equal to $\frac{1}{2}(\hat{\mathbf{u}} - \hat{\mathbf{v}})'\mathbf{y} + 2\varepsilon$ in RH-SVR in order for both methods to obtain the same solution $(\hat{\mathbf{u}}, \hat{\mathbf{v}})$.

To summarize the geometric properties, the solution of RH-SVR depends on the user's choice of the maximum tube size ε. RH-SVR shrinks the tube by maximizing the margin of separation of the shifted data. The size of the reduced convex hull, controlled by ν, affects the size of the margin. So the effective size of the tube depends on both ε and ν. For given data, ε_0 is the smallest value to form a hard tube. Consider the following cases. If $\varepsilon > \varepsilon_0$, the convex hulls themselves are strictly separable. It is not necessary to reduce them. The optimal regression function will be solely dependent on ε. If $\varepsilon \leqslant \varepsilon_0$, or if users prefer robust solutions in the separable case, the convex hulls are reduced so that the reduced convex hulls are strictly separable. The optimal planes depend on both ε and ν, because ν alters the shape of the convex sets. The number of support vectors depends on the interplay of ν and ε. There are at least $2\nu\ell$ support vectors. A larger ε allows a smaller ν to be used.

7. Analyzing generalization performance

In this section, a generalization error bound is derived for the geometric SVR model. We analyze the bound on the probability that a point has residual $|y - f(\mathbf{x})|$ larger than ε. Geometrically, this corresponds to the point being located outside of the ε-tube. Although the bound is not based on typical metrics for regression such as mean squared error, the result is still useful because if the probability of $|y - f(\mathbf{x})| > \varepsilon$ for a point $(\mathbf{x}, y)$ is high under the given tolerance of error ε, then the model needs improvement. The basic idea to derive the bound is that since solving regression problems, by our approach, has been converted to solving classification problems, we can take advantage of theoretical results concerning the generalization performance of a classifier [23,22,2,20]. Fig. 7 shows that a misclassification error in the dual classification space corresponds to a point outside of ε-tube in the primal regression space, and a margin error corresponds to a point outside of the $\hat{\varepsilon}$-tube. We show the result in Theorem 8.

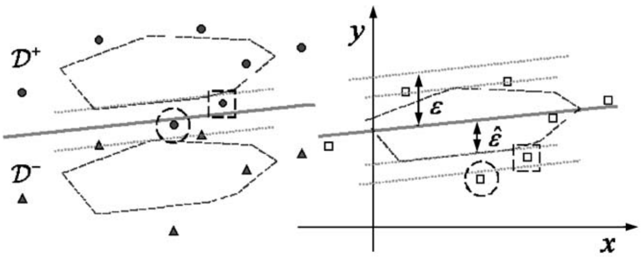

Fig. 7. The circled point in $\mathscr{D}^+$ is an error in the left graph, and has distance larger than ε from the regression plane on the right. The point with a rectangle is a margin error in the left graph, and distance larger than $\hat{\varepsilon}$ from the regression plane on the right.

Theorem 8 (Connection between regression and classification). *Let $f^{\mathrm{c}}(\mathbf{x}, y)$ be the classification function obtained by bisecting the two closest points in problem* (12), *and $f(\mathbf{x})$ be the corresponding regression function.* (i) *A point $\binom{\mathbf{x}}{y+\varepsilon}$ or $\binom{\mathbf{x}}{y-\varepsilon}$ is an error if and only if the point $\binom{\mathbf{x}}{y}$ is outside of the ε-tube constructed around $y = f(\mathbf{x})$*; (ii) *a point $\binom{\mathbf{x}}{y+\varepsilon}$ or $\binom{\mathbf{x}}{y-\varepsilon}$ is a margin error if and only if the point $\binom{\mathbf{x}}{y}$ is outside of the $\hat{\varepsilon}$-tube constructed around $y = f(\mathbf{x})$.*

Proof. The classification function $f^{\mathrm{c}}(\mathbf{x}, y)$ obtained by bisecting the closest points in problem (12) is $\hat{\mathbf{w}}'\Phi(\mathbf{x}) + \hat{\delta}y - \hat{b} = 0$. (For notational convenience, we do not use the expression in terms of kernel entries here.) As in Theorem 6, the corresponding regression function is $f(\mathbf{x}) = \bar{\mathbf{w}}'\Phi(\mathbf{x}) + \bar{b}$.

(i) If a point $\binom{\mathbf{x}_i}{y_i+\varepsilon}$ in class $\mathscr{D}^+$ is misclassified as in class $\mathscr{D}^-$ by $\operatorname{sgn}\circ f^{\mathrm{c}}$, it means $\hat{\mathbf{w}}'\Phi(\mathbf{x}_i) + \hat{\delta}(y_i + \varepsilon) - \hat{b} < 0$. Scale by $-\hat{\delta}$ (remember $\hat{\delta} > 0$), reverse the sign, and then we have $\bar{\mathbf{w}}'\Phi(\mathbf{x}_i) + \bar{b} - y_i > \varepsilon$, i.e., $f(\mathbf{x}_i) - y_i > \varepsilon$, so the point $\binom{\mathbf{x}_i}{y_i}$ is outside of the ε-tube. Similarly, if a point $\binom{\mathbf{x}_i}{y_i-\varepsilon}$ in $\mathscr{D}^-$ has been misclassified, then we have $y_i - f(\mathbf{x}_i) > \varepsilon$. The opposite direction also holds true. If a point is out of the ε-tube in regression space, one of the corresponding shifted points has to be an error in classification.

(ii) If a point such as $\binom{\mathbf{x}_i}{y_i+\varepsilon}$ in $\mathscr{D}^+$ is a margin error, based on the definition of margin errors, we have $\hat{\mathbf{w}}'\Phi(\mathbf{x}_i) + \hat{\delta}(y_i + \varepsilon) - \hat{b} < \rho$. Then after scaling by $-\hat{\delta}$, and substituting $\rho = (\tilde{\alpha} - \tilde{\beta})/2$, it becomes $\bar{\mathbf{w}}'\Phi(\mathbf{x}_i) + \bar{b} - y_i > \varepsilon - (\tilde{\alpha} - \tilde{\beta})/(2\hat{\delta}) = \hat{\varepsilon}$. Following the same line of reasoning, similar results hold true for margin errors $\binom{\mathbf{x}_i}{y_i-\varepsilon}$ on the other side of the tube. If a point is out of the $\hat{\varepsilon}$-tube in regression space, then one of the shifted points has to be a margin error in classification. □

The empirical fraction of margin errors is always less than or equal to $2v$. Let $\hat{\alpha}$ and $\hat{\beta}$ be optimal for Primal RH-SVR (9). Let $\tilde{\alpha}$ and $\tilde{\beta}$ be the values determined by the closest points as in Theorem 5. For $D = 1/(\ell v)$ for $0 \leqslant v \leqslant 0.5$, if $\tilde{\alpha} = \hat{\alpha}$ and $\tilde{\beta} = \hat{\beta}$, then by complementarity $2v$ is the upper bound on the fraction of margin errors. This fact is also true if $\tilde{\alpha} \neq \hat{\alpha}$ and $\tilde{\beta} \neq \hat{\beta}$. By complementarity for the shifted data $\mathscr{D}^+$, support vectors with $u_i < 1/(\ell v)$, called nonbound support vectors [17], have to locate precisely on the plane determined by $\hat{\alpha}$. The support vectors with $u_i = 1/(\ell v)$ may be margin errors, which means they are likely to be closer to the classification plane than nonbound support vectors or even on the wrong side. The closest point, which is the convex combination of all support vectors, is closer to the classification plane than the nonbound support vectors. Since all planes are parallel, the plane determined by $\tilde{\alpha}$ is closer to the classification plane than the one determined by $\hat{\alpha}$. Similarly the plane determined by $\tilde{\beta}$ is closer to the classification plane than the plane determined by $\hat{\beta}$. We then conclude that $\tilde{\alpha} - \tilde{\beta} \leqslant \hat{\alpha} - \hat{\beta}$. Consequently, taking the planes determined by $\tilde{\alpha}$

and $\tilde{\beta}$ as margin planes will yield fewer margin errors. Therefore 2ν still bounds from above the empirical fraction of margin errors.

There is a class of bounds on the generalization error of classifiers using the empirical fraction of margin errors and covering numbers to measure the classifier capacity. Schölkopf et al. proposed a generalization error bound for ν-SVM classification in [19] by applying theorems from [1,24]. They also hypothesized that better and perhaps more complicated bounds could be obtained by estimating the radius or even optimizing the choice of the center of the ball. Similar techniques can be adopted to analyze the generalization error of the RH-SVR regression approach based on the connection established in Theorem 8. Notice that the intercept term b in the general case can be incorporated into the kernel matrix $\mathbf{K}$ with minor modification.

Proposition 9 (Bound on generalization error). *Suppose RH-SVR uses a kernel of the form $k(\cdot,\cdot)$ such that $k(\mathbf{x},\mathbf{x})=1$ (for example, the RBF kernel). Let the response y be normalized to $[-1,1]$. Then all the data points $\begin{pmatrix}\Phi(\mathbf{x}_i)\\ y_i\end{pmatrix}$ in feature space live in a ball of radius $\sqrt{2}$ centered at the origin. With probability at least $1-\Delta$, the RH-SVR regression function $f(\mathbf{x})=\sum_i(\bar{v}_i-\bar{u}_i)k(\mathbf{x}_i,\mathbf{x})$ has $R^{\varepsilon}[f]$, the probability that $|y-f(\mathbf{x})|>\varepsilon$ for any $(\mathbf{x},y)$, bounded by*

$$R^{\varepsilon}[f]\leqslant 2\nu+\sqrt{\frac{2}{\ell}\left(\frac{8c^2(\|\mathbf{w}\|^2+1)}{(\varepsilon-\hat{\varepsilon})^2}\log_2(2\ell)-1+\ln\left(\frac{2}{\Delta}\right)\right)}. \tag{18}$$

Proof. The proof follows the one in [19] by applying the following two results. Suppose $\rho>0$, $0<\Delta<\frac{1}{2}$, and P is a probability distribution on the input space from which the training set is drawn. With probability at least $1-\Delta$ for every f in some function class $\mathscr{F}$, the probability of error of the classification decision function $\operatorname{sgn}\circ f$ on an independent test set denoted as $R[f]$ is bounded according to [1] by

$$R[f]\leqslant R^{\rho}_{\text{emp}}[f]+\sqrt{\frac{2}{\ell}\left(\ln\mathscr{N}\left(\mathscr{F},l_\infty^{2\ell},\frac{\rho}{2}\right)+\ln\left(\frac{2}{\Delta}\right)\right)}. \tag{19}$$

Let B_R be the ball of radius R around the origin in a certain Hilbert space F (feature space). Then the covering number $\mathscr{N}$ of the class of functions

$$\mathscr{F}=\{\mathbf{x}\mapsto\mathbf{w}'\mathbf{x}\,|\,\|\mathbf{w}\|\leqslant 1,\mathbf{x}\in B_R\}$$

at scale ρ when $\ell\geqslant 2$ satisfies $\log_2\mathscr{N}(\mathscr{F},l_\infty^{\ell},\rho)\leqslant(c^2R^2/\rho^2)\log_2\ell-1$ where $c<103$ is a constant [24]. Consequently,

$$\log_2\mathscr{N}(\mathscr{F},l_\infty^{2\ell},\rho/2)\leqslant\frac{4c^2R^2}{\rho^2}\log_2 2\ell-1. \tag{20}$$

In order to apply the second result, the geometric margin should be taken instead of the functional margin due to the condition $\|\mathbf{w}\|\leqslant 1$ in the class of functions $\mathscr{F}$. As we analyzed before in Section 6, the geometric margin of our method is $(\varepsilon-\hat{\varepsilon})/(\sqrt{\|\mathbf{w}\|^2+1})$. Notice that $R^{\rho}_{\text{emp}}[f]\leqslant 2\nu$. Substituting $\rho^2=(\varepsilon-\hat{\varepsilon})^2/(\|\mathbf{w}\|^2+1)$ into Eq. (20), and then substituting the resulting term into Eq. (19), we obtain the bound Eq. (18). □

When users take a large ε value, it means the tolerance of error is high, so the probability of a point outside of the ε-tube becomes low, which is exactly as shown in bound (18). From the perspective of Primal RH-SVR (9), RH-SVR minimizes the norm of the normal to the plane and maximizes the functional margin, so RH-SVR actually minimizes this generalization bound by maximizing the geometric margin of the classification function.

8. The nearest-point algorithm

The RH-SVR can be solved by applying nearest-point algorithms (NPAs) for finding the closest points in the reduced convex hulls. Prior NPAs of Keerthi et al. [13] and Kowalczyk [14] were limited to finding the closest points in convex hulls. We now demonstrate how the NPA can be extended to the problem of finding the closest points in the reduced convex hulls. The key to our NPA is the geometric conditions for optimality, thus we begin with a discussion of the optimality conditions.

8.1. Optimality conditions

Finding the closest points in two convex sets $\mathscr{U}$ and $\mathscr{V}$ can be briefly rewritten as

$$\begin{aligned} \min_{\mathbf{z}^+,\mathbf{z}^-} \quad & \tfrac{1}{2}\|\mathbf{z}^+ - \mathbf{z}^-\|^2 \\ \text{s.t.} \quad & \mathbf{z}^+ \in \mathscr{U}, \quad \mathbf{z}^- \in \mathscr{V}. \end{aligned} \tag{21}$$

In our case, $\mathscr{U}$ and $\mathscr{V}$ are the convex hulls or reduced convex hulls of the augmented training data. The convex hull generated by a finite set of points in $\mathbb{R}^n$ is actually a polyhedron. A polyhedron can be expressed by linear constraints. An optimization problem with linear cost function over a polyhedron such as $\min\{\mathbf{a}'\mathbf{z} \mid \mathbf{z} \in \mathscr{U}\}$ in Theorem A.2 is just a linear program. If a linear program has an optimal feasible solution, it has an optimal basic feasible solution which is an extreme point of the polyhedron. An extreme point is defined as a point in the set that does not lie within the line segment connecting any two distinct points in the set. When $D \geqslant 1$ the extreme points of $\mathscr{U}$ and $\mathscr{V}$ are subsets of $\mathscr{D}^+$ and $\mathscr{D}^-$, respectively. So the following corollary of Theorem A.2 can be easily derived. We describe it without proof.

Corollary 10 (Separation of $\mathscr{U}$ and $\mathscr{V}$ with $D \geqslant 1$). *Let $\mathscr{U}$ and $\mathscr{V}$ be defined as in Eq. (7). Then $(\mathbf{c} \in \mathscr{U}, \mathbf{d} \in \mathscr{V})$ is the solution to problem (21) if and only if*

$$\min\{\mathbf{a}'\mathbf{z}_i \mid \mathbf{z}_i \in \mathscr{D}^+\} \geqslant \mathbf{a}'\mathbf{c} \quad \textit{and} \quad \max\{\mathbf{a}'\mathbf{z}_j \mid \mathbf{z}_j \in \mathscr{D}^-\} \leqslant \mathbf{a}'\mathbf{d}, \tag{22}$$

where $\mathbf{a} = \mathbf{c} - \mathbf{d}$.

Since $\mathscr{D}^+$ and $\mathscr{D}^-$ are finite, the solution $(\mathbf{c}, \mathbf{d})$ to problem (21) can be verified by simply iterating through all the points. A linear programming package is not required. When $D < 1$, the extreme points of $\mathscr{U}$ are not necessarily included in $\mathscr{D}^+$. Assume ℓv takes a positive integer value M, obviously, $1 \leqslant M \leqslant |\mathscr{D}|$. Since $D = 1/\ell v$, it implies that the closest points **c** and **d** each depend on at least M training points. We now show that an extreme point of $\mathscr{U}$, $\mathbf{z}=\sum u_i\mathbf{z}_i$, has all u_i's either equal to 0 or $1/M$, but never in between.

Theorem 11 (Vertices of reduced convex set). *Let $\mathscr{S}$ be a reduced convex set defined as*

$$\mathscr{S}=\left\{\sum_{i=1}^{\ell}u_i\mathbf{z}_i \;\middle|\; \sum_{i=1}^{\ell}u_i=1,\quad 0\leqslant u_i\leqslant\frac{1}{M}\right\}.$$

If $\mathbf{z}=\sum_{i=1}^{\ell}\gamma_i\mathbf{z}_i$ is an extreme point of $\mathscr{S}$, then $\gamma_i=0$ or $1/M, \forall i=1,\ldots,\ell$.

Proof. We use proof by contradiction. Suppose not all γ_i's are either 0 or $1/M$, i.e., $\exists\gamma_t$ such that $0<\gamma_t<1/M$. We show that there has to exist another $\gamma_s, (s\neq t)$, and $0<\gamma_s<1/M$. Otherwise, all coefficients $\gamma_i=0$, or $1/M, \forall i\neq t$. Let the number of coefficients equal to $1/M$ be Γ.

Case 1: If $\Gamma\geqslant M$, then $\sum\gamma_i=\sum_{i\neq t}\gamma_i+\gamma_t\geqslant 1+\gamma_t>1$;

Case 2: If $\Gamma\leqslant M-1$, then $\sum\gamma_i=\sum_{i\neq t}\gamma_i+\gamma_t\leqslant(M-1)1/M+\gamma_t<1$. This contradicts the condition of $\sum_{i=1}^{\ell}\gamma_i=1$. Hence $\exists\gamma_s,\gamma_t$, such that $s\neq t$, $0<\gamma_t,\gamma_s<1/M$. Let $\varDelta>0$ be small enough such that $0\leqslant\gamma_t^1,\gamma_t^2\leqslant 1/M$ where $\gamma_t^1=\gamma_t-\varDelta$, $\gamma_t^2=\gamma_t+\varDelta$, and $0\leqslant\gamma_s^1,\gamma_s^2\leqslant 1/M$ where $\gamma_s^1=\gamma_s+\varDelta$, $\gamma_s^2=\gamma_s-\varDelta$. Then the points $\mathbf{z}^1=\sum_{i\neq t,s}\gamma_i\mathbf{z}_i+\gamma_t^1\mathbf{z}_t+\gamma_s^1\mathbf{z}_s\in\mathscr{S}$, and $\mathbf{z}^2=\sum_{i\neq t,s}\gamma_i\mathbf{z}_i+\gamma_t^2\mathbf{z}_t+\gamma_s^2\mathbf{z}_s\in\mathscr{S}$. Therefore $\mathbf{z}=\frac{1}{2}\mathbf{z}^1+\frac{1}{2}\mathbf{z}^2$, which means $\mathbf{z}$ is not an extreme point of $\mathscr{S}$, a contradiction. □

By exploiting Theorem 11, the optimality conditions can be evaluated more efficiently. It is not necessary to explicitly solve the linear programs in Theorem A.2 even for the reduced convex hull cases. Denote the set containing all convex combinations of $\mathbf{z}_i$'s with coefficients either 0 or $1/M$ by $\mathscr{S}_E$. Each point in $\mathscr{S}_E$ will have exactly M nonzero coefficients of value $1/M$. By Theorem 11, the set $E_{\mathscr{S}}$ of extreme points of $\mathscr{S}$ is a subset of $\mathscr{S}_E$. We can evaluate optimality by sequentially examining the points in $\mathscr{S}_E$ since it is a finite set. The cardinality of $\mathscr{S}_E$, however, is $\binom{\ell}{M}=\ell!/(M!(\ell-M)!)$, a rather large number for ℓ and M large. Fortunately, we do not have to evaluate all $\binom{\ell}{M}$ points. Consider the reduced convex hull $\mathscr{U}$. We can calculate $\mathbf{a}'\mathbf{z}_i, \forall\mathbf{z}_i\in\mathscr{D}^+$ and select the M points with the smallest values of $\mathbf{a}'\mathbf{z}_i$. Denote these M points in $\mathscr{D}^+$ as $\{\mathbf{z}_r^+, r=1,\ldots,M\}$. A point $\hat{\mathbf{z}}$ in $\mathscr{U}_E$ can be formed by associating these M points with coefficients $1/M$ and others with 0, i.e., $\hat{\mathbf{z}}=\sum_{r=1}^{M}(1/M)\mathbf{z}_r^+$. Only this point has to be examined as shown in the following corollary. If $\hat{\mathbf{z}}$ satisfies $\mathbf{a}'\hat{\mathbf{z}}\geqslant\mathbf{a}'\mathbf{c}$, then **c** is currently optimal in $\mathscr{U}$; otherwise, **c** can be improved using $\hat{\mathbf{z}}$. Correspondingly, the M points in $\mathscr{D}^-$ with largest values of $\mathbf{a}'\mathbf{z}_j, \forall\mathbf{z}_j\in\mathscr{D}^-$ are denoted as $\{\mathbf{z}_r^-, r=1,\ldots,M\}$. The following corollary holds true for the reduced convex hull case.

Corollary 12 (Separation of $\mathscr{U}$ and $\mathscr{V}$ with $D=1/M$). *Let $\mathscr{U}$ and $\mathscr{V}$ be defined as in Eq.* (7), *then* $(\mathbf{c},\mathbf{d})$ *is the solution to problem* (21) *if and only if* $(1/M)\sum_{r=1}^{M}\mathbf{a}'\mathbf{z}_r^+\geqslant\mathbf{a}'\mathbf{c}$ *where* $\mathbf{a}'\mathbf{z}_1^+,\ldots,z\mathbf{a}'\mathbf{z}_M^+$ *are the M smallest values of* $\mathbf{a}'\mathbf{z}_i, \forall\mathbf{z}_i\in\mathscr{D}^+$ *and* $(1/M)\sum_{r=1}^{M}\mathbf{a}'\mathbf{z}_r^-$

$\leqslant \mathbf{a}'\mathbf{d}$ *where* $\mathbf{a}'\mathbf{z}_1^-, \ldots, \mathbf{a}'\mathbf{z}_M^-$ *are the* M *largest values of all* $\mathbf{a}'\mathbf{z}_j, \forall \mathbf{z}_j \in \mathscr{D}^-$, *where* $\mathbf{a} = \mathbf{c} - \mathbf{d}$.

Notice that Corollary 10 is the special case of Corollary 12 for $M = 1$. So Corollary 12 is a more general result, and it can be applied to either convex hulls or reduced convex hulls. By using kernels, we evaluate the separation of the regular or reduced convex hulls generated by the images $\Phi(\mathbf{z}_i)$ in feature space. All theoretical analysis above can be extended to feature space. Let us consider the convex hulls of $\mathscr{D}^+$ and $\mathscr{D}^-$ defined in feature space in Eq. (13). Then similarly the closest points $\mathbf{c}$ and $\mathbf{d}$ can be expressed as $\mathbf{c} = \sum_{\mathbf{z}_i \in \mathscr{D}^+} \hat{u}_i \Phi(\mathbf{z}_i)$, $\mathbf{d} = \sum_{\mathbf{z}_j \in \mathscr{D}^-} \hat{v}_j \Phi(\mathbf{z}_j)$, and still $\mathbf{a} = \mathbf{c} - \mathbf{d}$. As a consequence, the inner product such as $\mathbf{a}'\mathbf{z}_r^+$ in the previous theorem should be calculated in feature space as

$$\begin{aligned}\mathbf{a}'\Phi(\mathbf{z}_r^+) &= \left(\sum_{\mathbf{z}_i \in \mathscr{D}^+} \hat{u}_i \Phi(\mathbf{z}_i) - \sum_{\mathbf{z}_j \in \mathscr{D}^-} \hat{v}_j \Phi(\mathbf{z}_j)\right)' \Phi(\mathbf{z}_r^+) \\ &= \sum_{\mathbf{z}_i \in \mathscr{D}^+} \hat{u}_i \Phi(\mathbf{z}_i)' \Phi(\mathbf{z}_r^+) - \sum_{\mathbf{z}_j \in \mathscr{D}^-} \hat{v}_j \Phi(\mathbf{z}_j)' \Phi(\mathbf{z}_r^+) \\ &= \sum_{\mathbf{z}_i \in \mathscr{D}^+} \hat{u}_i k(\mathbf{z}_i, \mathbf{z}_r^+) - \sum_{\mathbf{z}_j \in \mathscr{D}^-} \hat{v}_j k(\mathbf{z}_j, \mathbf{z}_r^+).\end{aligned}$$

The kernel version of Theorem A.2 becomes

Corollary 13 (Separation of $\mathscr{U}$ and $\mathscr{V}$ in feature space). *Let* $\mathscr{U}$ *and* $\mathscr{V}$ *be defined as in Eq.* (13), *then* $(\mathbf{c}, \mathbf{d})$ *is the solution to problem* (21) *if and only if* $\min\{\mathbf{a}'\mathbf{z} \mid \mathbf{z} \in \mathscr{U}\} \geqslant \mathbf{a}'\mathbf{c}$ *and* $\max\{\mathbf{a}'\mathbf{z} \mid \mathbf{z} \in \mathscr{V}\} \leqslant \mathbf{a}'\mathbf{d}$ *where* $\mathbf{c} = \sum_{\mathbf{z}_i \in \mathscr{D}^+} \hat{u}_i \Phi(\mathbf{z}_i)$, $\mathbf{d} = \sum_{\mathbf{z}_j \in \mathscr{D}^-} \hat{v}_j \Phi(\mathbf{z}_j)$, *and* $\mathbf{a} = \mathbf{c} - \mathbf{d}$.

One of the significant advantages of kernel methods is that the image $\Phi(\mathbf{z})$ is never explicitly formed, and the mapping is implicitly done when calculating the kernel values, the inner products between the images. Notice that the calculation of $\mathbf{a}'\Phi(\mathbf{z}_r^+)$ uses only inner products of $\Phi(\mathbf{z})$'s. Thus $\mathbf{a}'\mathbf{c}$ and $\mathbf{a}'\mathbf{d}$ can also be expressed in terms of kernel values. We only need to know the coefficients $\hat{u}_i$ and $\hat{v}_j$ because the final regression function is $f(\mathbf{x}) = \sum_{i=1}^{\ell} (\bar{v}_i - \bar{u}_i) k(\mathbf{x}_i, \mathbf{x}) + \bar{b}$, and $\bar{v}_i = \hat{v}_i / \hat{\delta}$, $\bar{u}_i = \hat{u}_i / \hat{\delta}$. The following corollary is the kernel version of Corollary 12.

Corollary 14 (Optimality of separation between $\mathscr{U}$ and $\mathscr{V}$). *Let* $\mathscr{U}$ *and* $\mathscr{V}$ *be defined as in Eq.* (13), *then* $(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ *is the solution to problem* (12) *if and only if*

$$\frac{1}{M} \sum_{r=1}^{M} \left(\sum_{\mathbf{z}_i \in \mathscr{D}^+} \hat{u}_i k(\mathbf{z}_i, \mathbf{z}_r^+) - \sum_{\mathbf{z}_j \in \mathscr{D}^-} \hat{v}_j k(\mathbf{z}_j, \mathbf{z}_r^+) \right) \geqslant \mathbf{c}'\mathbf{c} - \mathbf{c}'\mathbf{d} \qquad (23)$$

and

$$\frac{1}{M}\sum_{r=1}^{M}\left(\sum_{\mathbf{z}_i\in\mathscr{D}^+}\hat{u}_i k(\mathbf{z}_i,\mathbf{z}_r^-)-\sum_{\mathbf{z}_j\in\mathscr{D}^-}\hat{v}_j k(\mathbf{z}_j,\mathbf{z}_r^-)\right)\leqslant \mathbf{c}'\mathbf{d}-\mathbf{d}'\mathbf{d}, \tag{24}$$

where

$$\mathbf{c}'\mathbf{c}=\sum_{\mathbf{z}_i,\mathbf{z}_j\in\mathscr{D}^+}\hat{u}_i\hat{u}_j k(\mathbf{z}_i,\mathbf{z}_j),$$

$$\mathbf{c}'\mathbf{d}=\sum_{\mathbf{z}_i\in\mathscr{D}^+,\mathbf{z}_j\in\mathscr{D}^-}\hat{u}_i\hat{v}_j k(\mathbf{z}_i,\mathbf{z}_j),$$

$$\mathbf{d}'\mathbf{d}=\sum_{\mathbf{z}_i,\mathbf{z}_j\in\mathscr{D}^-}\hat{v}_i\hat{v}_j k(\mathbf{z}_i,\mathbf{z}_j), \tag{25}$$

$\mathbf{z}_r^+, r=1,\ldots,M$ *are the* M *points with the smallest values of* $\mathbf{a}'\Phi(\mathbf{z}_i),\forall\mathbf{z}_i\in\mathscr{D}^+$, *and* $\mathbf{z}_r^-, r=1,\ldots,M$ *are the* M *points with the largest values of* $\mathbf{a}'\Phi(\mathbf{z}_j),\forall\mathbf{z}_j\in\mathscr{D}^-$.

8.2. The algorithm

Our NPA exploits the optimality conditions in Corollary 14. In each iteration the algorithm evaluates the optimality of the current tth solution $(\mathbf{u}^t,\mathbf{v}^t)$ by computing Eq. (23) and (24). Once a point is found violating the optimality conditions, it provides clues on how to update the current solution. We provide the basic ideas and procedures of the NPA in this article. There are possibilities for improvement that are left to future research.

Keerthi et al. proposed a nearest-point algorithm [12] which is a hybrid approach combining Gilbert's algorithm for minimizing a quadratic form on a convex set [10,11] and the Mitchell-Dem'yanov-Malozemov algorithm [16] for separable classification problems. The second algorithm is one way to speed up convergence of NPA. There exist other ways to speed up the NPA based on the analysis in [13,14]. For example, Keerthi et al. has reported that it helps to find the point within a triangle, instead of on a line segment, in one convex hull that is closest to the point in the other convex hull. In Kowalczyk's implementation, allowing the searching points outside of the line segment is an essential step to speed it up. To keep our generic description of the NPA simple, we will not include those discussions in the framework of NPA here.

Let $\hat{\mathbf{z}}^+=\sum_{r=1}^{M}(1/M)\Phi(\mathbf{z}_r^+)$ and $\hat{\mathbf{z}}^-=\sum_{r=1}^{M}(1/M)\Phi(\mathbf{z}_r^-)$. Denote $\mathbf{a}=\mathbf{c}-\mathbf{d}$. We seek a solution to satisfy the optimality conditions with a tolerance of $\tau>0$. Then Eqs. (23) and (24) become

$$\mathbf{a}'\hat{\mathbf{z}}^+\geqslant\mathbf{a}'\mathbf{c}-\tau, \tag{26}$$

$$\mathbf{a}'\hat{\mathbf{z}}^-\leqslant\mathbf{a}'\mathbf{d}+\tau. \tag{27}$$

If these optimality conditions are not satisfied, e.g. $\mathbf{a}'\hat{\mathbf{z}}^+<\mathbf{a}'\mathbf{c}-\tau$, then there exists a point $\mathbf{c}^{\text{new}}$ on the segment from $\mathbf{c}$ to $\hat{\mathbf{z}}^+$ such that the objective value is strictly decreased, i.e., $\|\mathbf{c}^{\text{new}}-\mathbf{d}\|^2<\|\mathbf{c}-\mathbf{d}\|^2$. A point on a line segment with end points $\mathbf{c}$

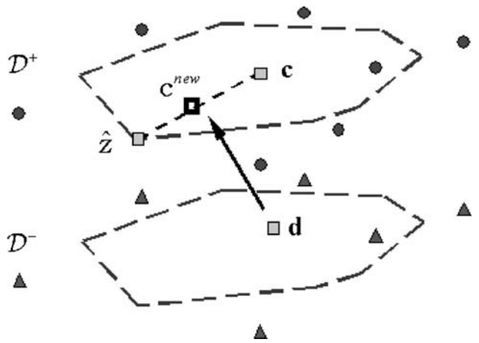

Fig. 8. The point $\mathbf{c}^{\text{new}}$ is the point on the line segment from $\mathbf{c}$ to $\hat{\mathbf{z}}$ closest to $\mathbf{d}$, and the distance between $\mathbf{c}^{\text{new}}$ and $\mathbf{d}$ is smaller than that between $\mathbf{c}$ and $\mathbf{d}$.

and $\hat{\mathbf{z}}^+$ can be expressed as $\lambda\mathbf{c} + (1-\lambda)\hat{\mathbf{z}}^+$, where $0 \leqslant \lambda \leqslant 1$. Fig. 8 illustrates this process. Correspondingly, in set $\mathscr{V}$, we can update $\mathbf{d}$ in a similar way. The following theorem applies to both cases.

Theorem 15 (Closest point of line segment). *Let $\mathscr{S}$ be a reduced convex hull, $\mathbf{p} \notin \mathscr{S}$, $\mathbf{z} \in \mathscr{S}$, and $\mathbf{a} = \mathbf{p} - \mathbf{z}$. If $\hat{\mathbf{z}} \in \mathscr{S}$ such that $\mathbf{a}'\hat{\mathbf{z}} > \mathbf{a}'\mathbf{z}$, then $\exists \mathbf{z}^{\text{new}} = \hat{\lambda}\mathbf{z} + (1-\hat{\lambda})\hat{\mathbf{z}}$, and $\mathbf{z}^{\text{new}} \neq \mathbf{z}$, such that $\|\mathbf{z}^{\text{new}} - \mathbf{p}\|^2 < \|\mathbf{z} - \mathbf{p}\|^2$ where*

$$\hat{\lambda} = \begin{cases} 0 & \text{if } (\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{p}-\hat{\mathbf{z}}) \leqslant 0, \\ \frac{(\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{p}-\hat{\mathbf{z}})}{\|\mathbf{z}-\hat{\mathbf{z}}\|^2} & \text{if } (\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{p}-\hat{\mathbf{z}}) > 0. \end{cases}$$

Proof. The point on the line segment closest to $\mathbf{p}$ can be found by solving the following problem for λ:

$$\min_{0\leqslant\lambda\leqslant 1} g(\lambda) = \|\lambda\mathbf{z} + (1-\lambda)\hat{\mathbf{z}} - \mathbf{p}\|^2. \tag{28}$$

The derivative of g is

$$g'(\lambda) = (\lambda\mathbf{z} + (1-\lambda)\hat{\mathbf{z}} - \mathbf{p})'(\mathbf{z} - \hat{\mathbf{z}}).$$

Setting $g'(\lambda) = 0$ yields $\hat{\lambda} = (\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{p}-\hat{\mathbf{z}})/\|\mathbf{z}-\hat{\mathbf{z}}\|^2$. The $\hat{\lambda}$ must be < 1 since

$$\begin{aligned} \|\mathbf{z}-\hat{\mathbf{z}}\|^2 &= (\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{z}-\hat{\mathbf{z}}) = (\mathbf{z}-\hat{\mathbf{z}})'((\mathbf{z}-\mathbf{p}) + (\mathbf{p}-\hat{\mathbf{z}})) \\ &= (\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{z}-\mathbf{p}) + (\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{p}-\hat{\mathbf{z}}) \\ &> (\mathbf{z}-\hat{\mathbf{z}})'(\mathbf{p}-\hat{\mathbf{z}}) \end{aligned}$$

by the condition of $\mathbf{a}'\hat{\mathbf{z}} > \mathbf{a}'\mathbf{z}$. Therefore we have $\mathbf{z}^{\text{new}} \neq \mathbf{z}$. Since $g(\lambda)$ is a strictly convex function, if $\hat{\lambda} < 0$, we just set $\hat{\lambda} = 0$. Then $\hat{\lambda}$ is feasible, and is the unique global minimizer of problem (28). Furthermore $\mathbf{z}, \hat{\mathbf{z}}$ are both in the convex set $\mathscr{S}$, so $\mathbf{z}^{\text{new}} \in \mathscr{S}$. □

The normal of the max-margin plane $\begin{pmatrix} \hat{\mathbf{w}} \\ \hat{\delta} \end{pmatrix}$ is exactly the difference between the closest points $\mathbf{c}-\mathbf{d}$, and the intercept term $\hat{b}$ in Theorems 5 and 6 is the distance from the origin to the point halfway between the two closest points $\mathbf{c}$ and $\mathbf{d}$ along the normal $\begin{pmatrix} \hat{\mathbf{w}} \\ \hat{\delta} \end{pmatrix}$, so $\hat{b}=\begin{pmatrix} \hat{\mathbf{w}} \\ \hat{\delta} \end{pmatrix}'(\mathbf{c}+\mathbf{d})/2=\frac{1}{2}(\mathbf{c}-\mathbf{d})'(\mathbf{c}+\mathbf{d})=(\|\mathbf{c}\|^2-\|\mathbf{d}\|^2)/2$. The outline of our NPA is presented in Algorithm 1. In order to use one subroutine to assess optimality for both sets, we add a minus sign to Eq. (27) so that evaluating optimality in $\mathscr{V}$ becomes a minimization problem too. The parentheses contain the algorithm parameters.

Algorithm 1. Nearest-point algorithm

```
main routine NPA ($M, \varepsilon, \tau$)
    Initialize $\mathbf{u}, \mathbf{v}$ as $u_i = v_i = \frac{1}{\ell}$, $i = 1, \ldots, \ell$
    repeat
        UpdateOneSet ($\mathbf{c}, \mathscr{D}^+, \mathbf{c}-\mathbf{d}$)
        UpdateOneSet ($\mathbf{d}, \mathscr{D}^-, \mathbf{d}-\mathbf{c}$)
    until ($\mathbf{u}, \mathbf{v}$) is optimal for both $\mathscr{U}$ and $\mathscr{V}$
    Compute $b$
    Compute $\delta$
    Scale $\mathbf{u}$ and $\mathbf{v}$ by $-\delta$, $b$ by $\delta$
    Return the obtained model
end

subroutine UpdateOneSet ($\mathbf{z}, \mathscr{S}, \mathbf{a}$)
    Find $M$ points $\mathbf{z}_i$ in $\mathscr{S}$ with smallest value of $\mathbf{a}'\Phi(\mathbf{z}_i)$, compute $\hat{\mathbf{z}}$
        using these $M$ points
    Check the optimality of $\hat{\mathbf{z}}$ by $\mathbf{a}'\hat{\mathbf{z}} \geqslant \mathbf{a}'\mathbf{z} - \tau$ (Equation (26) or (27))
    If not optimal, then update $\mathbf{z}$ using $\hat{\mathbf{z}}$ as given in Theorem 15
end
```

We have successfully tested this NPA on the Boston Housing problem discussed in the next section. Optimal solutions were found and the quality of the results was equivalent to that of those found using the commercial optimization package CPLEX. But our initial implementation is relatively slow. More efficient implementations are possible and strategies like those discussed in [12,14] can be employed to further speed up the algorithm. Better strategies for initializing $\mathbf{u}$ and $\mathbf{v}$ may also enhance the algorithm. We leave more extensive investigations of NPA to future work.

9. Experimental studies

The goal of our experimental study was to examine differences between SVR models. We used the CPLEX 6.6 optimization package to optimize three regression models: ε-SVR, ν-SVR, and RH-SVR. Note that it is not possible to apply NPA to all three models. First the difference between RH-SVR and ε-SVR is illustrated on a toy linear

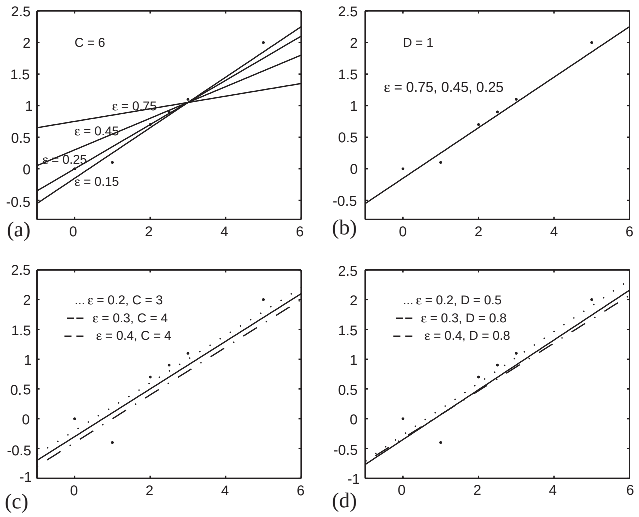

Fig. 9. Regression lines from (a) ε-SVR and (b) RH-SVR with distinct ε where $\varepsilon > \varepsilon_0$; regression lines from (c) ε-SVR and (d) RH-SVR where $\varepsilon \leqslant \varepsilon_0$.

problem. The data consist of (x, y): (0, 0), (1, 0.1), (2, 0.7), (2.5, 0.9), (3, 1.1) and (5, 2) in separable cases of $\varepsilon > \varepsilon_0$, and the point (1, 0.1) is replaced by $(1, -0.4)$ in inseparable cases of $\varepsilon \leqslant \varepsilon_0$ to make the figures legible. As shown in Figs. 9 (a) and (b), ε-SVR produces undesirable results for large ε, and RH-SVR constructs the same function for ε sufficiently large. For $\varepsilon \leqslant \varepsilon_0$, ε-SVR generates parallel lines with different ε, and RH-SVR adjusts the lines by slight rotation as presented in Figs. 9(c) and (d).

In Table 1, we compare RH-SVR, ε-SVR and ν-SVR on the Boston Housing problem. Following the experimental design in [18] we used the RBF kernel with $2\sigma^2 = 3.9$ for all three approaches, $C = 500 \cdot \ell$ for ε-SVR and ν-SVR, and $\varepsilon = 3.6$ for RH-SVR. RH-SVR, ε-SVR, and ν-SVR are computationally similar for good parameter choices. In ε-SVR, the actual size of the ε tube is fixed. In RH-SVR, ε is the maximum allowable tube width. Choosing ε is critical for ε-SVR but less so for RH-SVR.

Table 2 lists the effective tube sizes, numbers of SVs, and numbers of errors. Here, the errors we counted in experiments are those support vectors at upper bound $1/\ell\nu$. Both RH-SVR and ν-SVR can shrink or grow the tube according to desired robustness. But ν-SVR has no upper ε bound. Unlike ν-SVR, the actual sizes of ε-tubes constructed by RH-SVR do not vanish to 0 when increasing ν. From Theorem 6, this is to be expected. In this particular experiment, the results show that RH-SVR tends to take larger values of ν to produce the same effective tube size as ν-SVR, thus RH-SVR has correspondingly more support vectors and errors. Cross-referencing with the previous table, the effective tube sizes for the functions that generalize well are similar. From Section 6, we know that both ν-SVR and RH-SVR minimize objective functions over the same set of constraints, i.e., the same reduced convex hulls, but the objective

Table 1
Testing results for Boston Housing; MSE: average of mean squared errors of 25 testing points over 100 trials; STD: standard deviation.

	2ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RH-SVR	MSE	14.9	9.3	8.8	8.9	9.8	11.0	11.5	12.4	12.5
	STD	22.9	7.6	6.9	6.4	7.2	7.7	8.8	9.7	10.5
	ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν-SVR	MSE	9.6	8.9	9.5	10.8	10.9	11.0	11.2	11.1	11.2
	STD	5.8	7.9	8.3	8.2	8.3	8.4	8.5	8.4	8.4
	ε	0	1	2	3	4	5	6	7	8
ε-SVR	MSE	11.2	10.8	9.5	10.3	11.6	13.6	15.6	17.2	19.9
	STD	8.3	8.2	8.2	7.3	5.8	5.8	5.9	5.8	6.0

Table 2
Testing results for Boston Housing, actual tube size, fraction of SVs, and fraction of errors (i.e., points with u_i or $v_i = \frac{1}{\ell\nu}$.)

	2ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RH-SVR	$\hat{\varepsilon}$	2.86	2.03	1.66	1.53	1.34	1.10	0.72	0.77	0.72
	SVs	0.44	0.70	0.68	0.71	0.82	0.89	0.92	0.92	0.96
	Errors	0.05	0.00	0.24	0.24	0.19	0.36	0.64	0.80	0.88
	ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν-SVR	ε	2.70	1.99	1.43	1.08	0.77	0.48	0.22	0.00	0.00
	SVs	0.3	0.4	0.6	0.7	0.8	0.9	1.0	1.0	1.0
	Errors	0.0	0.1	0.2	0.2	0.3	0.4	0.5	0.5	0.5

functions are different on the terms concerning the response y. Hence in this example, a smaller reduced convex hull is required for RH-SVR to achieve the same effective tube as ν-SVR. By increasing ε, the number of support vectors used by RH-SVR can be reduced.

To probe the effects of ε on RH-SVR, different ε values were used and results are provided in Table 3. From the table, $\varepsilon = 3.6$ is the best choice for the experiments on Boston Housing data. The smaller ε is, the greater the convex hulls must be reduced. With $\varepsilon = 2.6$, the best choice of 2ν is 0.5 while $2\nu = 0.3$ is the best with $\varepsilon = 3.6$. We obtained poor results when taking $\varepsilon = 2.6$ and $2\nu = 0.1$ because in this case, the convex hulls were not reduced enough to be separable, and hence the nearest-point problem was not well-defined. Good solutions were obtained over a wide range of appropriate parameter values. For example, RH-SVR achieved small MSE values for 2ν in the range of $[0.2, 0.5]$ and $\varepsilon = 3.6$ as ν-SVR performed similarly when $\nu \in [0.1, 0.3]$. Hence the method is as robust as ν-SVR in this sense.

Table 3
Boston Housing results with different ε values

	2ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ε=2.6	MSE	37.3	11.2	10.7	9.6	8.9	10.6	11.5	12.5	13.8
	STD	72.3	7.6	7.3	7.4	8.4	9.1	9.3	9.8	10.8
ε=3.6	MSE	14.9	9.3	8.8	8.9	9.8	11.0	11.5	12.4	12.5
	STD	22.9	7.6	6.9	6.4	7.2	7.7	8.8	9.7	10.5
ε=5.0	MSE	9.6	9.5	10.7	11.0	13.4	15.8	16.6	18.0	23.5
	STD	5.1	6.7	7.3	7.7	8.4	9.7	10.3	11.7	11.8

Table 4
QSAR data sets

Data set	No. of obs.	No. of vars.	Preprocessed no. of vars.
Blood/brain barrier	62	694	569
Cancer	46	769	362
Cholecystokinin	66	626	350
HIV	64	620	561
Caco2	27	715	713

Table 5
The experimental results with QSAR data sets

Data set	RH-SVR		ν-SVR	
	Q^2	std(Q^2)	Q^2	std(Q^2)
Blood/brain barrier	0.290	0.079	0.297	0.074
Cancer	0.331	0.171	0.248	0.15
Cholecystokinin	0.414	0.068	0.459	0.087
HIV	0.441	0.077	0.390	0.077
Caco2	0.371	0.083	0.424	0.069

The RH-SVR method was also tested on several data sets arising from quantitative structure–activity relationships (QSAR) analysis. QSAR analysis plays an essential role in the rational drug design process [8]. Problems encountered in this area are usually very difficult inference problems because the dimension of the input space is very high while the available amount of training data is small. The datasets used in our experiments were created in an ongoing NSF-funded drug design and semi-supervised learning (DDASSL) project (see http://www.drugmining.com). We summarize the datasets in Table 4. Results in Table 5 are obtained based on the leave-one-out procedure with model selection for choosing proper parameter values in each fold.

The statistic Q^2 is defined as the MSE normalized by the variance of the response. We calculate the std(Q^2) as the standard deviation of squared errors divided by the variance of the response. The experiment is designed as follows: in each fold, one point is held out, the remaining data are further divided into training set (80% of data) and validation set (20% of data). Then we train the SVR machines on the training data and tune the parameters based on the validation data. The model obtained with the selected parameters is tested on the hold-out data point. Table 5 shows that RH-SVR and ν-SVR perform similarly with parameters tuned for each method. Based on our observation, both RH-SVR and ν-SVR prefer the fraction of SVs to be 20% as $2\nu = 0.2$ in RH-SVR and $\nu = 0.2$ in ν-SVR are selected in many folds. The selected ε values in model selection for RH-SVR vary in different folds.

10. Discussion

In this work we showed that regression can be viewed as a classification problem discriminating between the over and underestimates of the regression data in the $(\mathbf{x}, y)$ space. By examining when ε-tubes exist, we showed formally that in the dual space SVR can be regarded as a classification problem. Hard and soft ε-tubes are constructed by separating the convex or reduced convex hulls, respectively, of the training data with the response variable shifted up and down by ε. We proposed RH-SVR based on choosing the soft max-margin plane between the two shifted datasets. Like ν-SVM, RH-SVR shrinks the ε-tube. The max-margin determines how much the tube shrinks. Domain knowledge can be incorporated into the RH-SVR parameters ε and ν. The parameter C in ν-SVM and ε-SVR has been eliminated. Computational results demonstrate that no one SVR formulation is superior for good parameter choices. RH-SVR alone has a geometrically intuitive framework that allows users to easily grasp the model and its parameters.

Considering regression as classification opens up many interesting possibilities for analysis and algorithms. We showed how generalization error bounds for the classification case could be generalized to the regression case. The objective of RH-SVR directly minimizes this error bound. Since the dual formulation of RH-SVR corresponds geometrically to finding the closest points in the reduced convex hulls, RH-SVR can be solved by nearest-point algorithms. We outlined a simple NPA for RH-SVR. Initial results for this NPA are promising but we leave optimizing the NPA to future work. There may be possibilities of adapting other ideas from SV classification to this SVR. Similarly, it is possible to construct new regression methods by applying other descriptive algorithms to the augmented training data. Can the generalization models such as the LOO-bound or the SPAN bound used for parameter optimization in classification be extended to SVR? Can other classification methods be generalized to regression using similar strategies? These are all open questions for future research.

Acknowledgements

Many thanks to the helpful comments of the reviewers that improved this paper. This work was supported by NSF IRI-9702306 and NSF IIS-9979860.

Appendix

This appendix is dedicated to the separation theorems that are used in constructing the soft ε-tube for regression and deriving the nearest point algorithm. In order to prove the main theorem, we need the following lemma.

Lemma A.1 (Minimum distance from a point to a convex set; Bazaraa et al. [3]). *Let $\mathscr{S}$ be a nonempty, closed convex set in $\mathbb{R}^n$, and $\mathbf{p} \notin \mathscr{S}$. $\hat{\mathbf{x}}$ is the point in $\mathscr{S}$ closest to $\mathbf{p}$ if and only if $(\mathbf{p}-\hat{\mathbf{x}})'(\mathbf{x}-\hat{\mathbf{x}}) \leqslant 0, \forall \mathbf{x} \in \mathscr{S}$.*

Finding the closest points in two convex sets $\mathscr{U}$ and $\mathscr{V}$ can be stated as

$$\begin{aligned} \min_{\mathbf{z}^+,\mathbf{z}^-} \quad & \tfrac{1}{2}\|\mathbf{z}^+ - \mathbf{z}^-\|^2 \\ \text{s.t.} \quad & \mathbf{z}^+ \in \mathscr{U}, \quad \mathbf{z}^- \in \mathscr{V}. \end{aligned} \tag{A.1}$$

The sets $\mathscr{U}$ and $\mathscr{V}$ in our cases are the convex hulls or reduced convex hulls of a finite set of data points as in Eq. (7). Since the function $\frac{1}{2}\|\mathbf{z}^+ - \mathbf{z}^-\|^2$ in problem (A.1) is a continuous function on the compact set $\mathscr{U} \times \mathscr{V}$, then by applying basic calculus, we know $\mathbf{c} \in \mathscr{U}$ and $\mathbf{d} \in \mathscr{V}$ exist such that the infimum of the objective function is attained at the two points. So formulation (A.1) is well defined. In our case, the $\mathscr{U}$ and $\mathscr{V}$ are strictly separable. It implies that $\mathbf{c} \notin \mathscr{V}$ and $\mathbf{d} \notin \mathscr{U}$, or equivalently $\|\mathbf{c}-\mathbf{d}\| > 0$. The following theorem characterizes the closest points of $\mathscr{U}$ and $\mathscr{V}$.

Theorem A.2 (Separation of $\mathscr{U}$ and $\mathscr{V}$). *Let $\mathscr{U}$ and $\mathscr{V}$ be defined as in Eq.* (7) *with $\mathscr{U} \cap \mathscr{V} = \emptyset$, then $(\mathbf{c} \in \mathscr{U}, \mathbf{d} \in \mathscr{V})$ is the solution to problem* (A.1) *if and only if* $\min\{\mathbf{a}'\mathbf{z} \mid \mathbf{z} \in \mathscr{U}\} \geqslant \mathbf{a}'\mathbf{c}$ *and* $\max\{\mathbf{a}'\mathbf{z} \mid \mathbf{z} \in \mathscr{V}\} \leqslant \mathbf{a}'\mathbf{d}$ *where* $\mathbf{a} = \mathbf{c} - \mathbf{d}$.

Proof. Without loss of generality, we only prove one of the conditions. Consider the convex set $\mathscr{U}$. From Eq. (7), $\mathscr{U}$ is a closed and convex set. $(\mathbf{c}, \mathbf{d})$ is the solution of problem (A.1) if and only if $\mathbf{c}$ is the point in the polyhedron $\mathscr{U}$ closest to $\mathbf{d}$. Since $\mathbf{d} \notin \mathscr{U}$, by the Lemma 16, $(\mathbf{d}-\mathbf{c})'(\mathbf{z}-\mathbf{c}) \leqslant 0, \forall \mathbf{z} \in \mathscr{U}$. Denoting $\mathbf{a} = \mathbf{c} - \mathbf{d}$, we have $\mathbf{a}'\mathbf{z} \geqslant \mathbf{a}'\mathbf{c}, \forall \mathbf{z} \in \mathscr{U}$, which means $\min\{\mathbf{a}'\mathbf{z} \mid \mathbf{z} \in \mathscr{U}\} \geqslant \mathbf{a}'\mathbf{c}$. Similar arguments hold true for the other linear program. □

References

[1] P.L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Trans. Inf. Theory 44 (2) (1998) 525–536.

[2] P. Bartlett, J. Shawe-Taylor, Generalization performance of support vector machines and other pattern classifiers, in: B.Schölkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, USA, 1998.

[3] M.S. Bazaraa, H.D. Sherali, C.M. Shetty, Nonlinear Programming—Theory and Algorithms, 2nd Edition, Wiley, New York, 1993.

[4] K.P. Bennett, E.J. Bredensteiner, Duality and geometry in SVM classifiers, in: Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000, pp. 57–64.

[5] J. Bi, K.P. Bennett, Duality, geometry, and support vector regression, in: T.G. Dietterich, S. Becker, Z, Ghahramani (Eds.), Advances in Neural Information Processing Systems, Vol. 14, MIT Press, Cambridge, MA, 2002, pp. 593–600.

[6] D. Crisp, C.J.C. Burges, A geometric interpretation of ν-SVM classifiers, in: S.A. Solla, M.S. Kearns, D.A. Cohn (Eds.), Advances in Neural Information Processing Systems, Vol. 11, MIT Press, Cambridge, MA, 1999, pp. 244–251.

[7] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK, 2000.

[8] A. Demiriz, K.P. Bennett, C.M. Breneman, M.J. Embrechts, Support vector machine regression in chemometrics, Computing Science and Statistics, Proceedings of the 33rd Symposium on the Interface, American Statistical Association for the Interface Foundation of North America, Washington, D.C., 2001.

[9] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, in: M.C. Mozer, M.I. Jordan, T.Petsche, (Eds.), Advances in Neural Information Processing Systems, Vol. 9, MIT Press, Cambridge, MA, 1997, pp. 155–161.

[10] E.G. Gilbert, Minimizing the quadratic form on a convex set, SIAM J. Control 4 (1966) 61–79.

[11] E.G. Gilbert, D.W. Johnson, S.S. Keerthi, A fast procedure for computing the distance between complex objects in three dimensional space, IEEE J. Robotics Automat. 4 (1988) 193–203.

[12] S.S. Keerthi, S.K. Shevade, C.Bhattacharyya, K.R.K. Murthy, A fast iterative nearest point algorithm for support vector machines, Technical Report TR-ISL-99-03, Intelligent Systems Lab, Indian Institute of Science, Bangalore 560 012, India, 1999. http://guppy.mpe.nus.edu.sg/∼mpessk/npa.shtml.

[13] S.S. Keerthi, S.K. Shevade, C.Bhattacharyya, K.R.K. Murthy, A fast iterative nearest point algorithm for support vector machine classifier design, IEEE Trans. Neural Networks 11 (1) (2000) 124–136.

[14] A. Kowalczyk, Maximal margin perceptron, in: A.J. Smola, P. Bartlett, B. Schöelkopf, C.Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 1999, pp. 75–113.

[15] O.L. Mangasarian, Nonlinear Programming, SIAM, Philadelphia, 1994.

[16] B.F. Mitchell, V.F. Dem'yanov, V.N. Malozemov, Finding the point of a polyhedron closest to the origin, SIAM J. Control 12 (1974) 19–26.

[17] J. Platt, Sequential minimal optimization: a fast algorithm for training support vector machines, in: B. Schölkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 185–208, http://www.research.microsoft.com/∼jplatt/smo.html.

[18] B. Schölkopf, B. Smola, R. Williamson, Shrinking the tube: a new support vector regression algorithm, in: S.A. Solla, M.S. Kearns, D.A. Cohn, (Eds.), Advances in Neural Information Processing Systems, Vol. 11, MIT Press, Cambridge, MA, 1999, pp. 330–336.

[19] B. Schölkopf, A. Smola, R.C. Williamson, P.L. Bartlett, New support vector algorithms, Neural Comput. 12 (2000) 1207–1245.

[20] J. Shawe-Taylor, N. Cristianini, Robust bounds on generalization from the margin distribution, NeuroCOLT2 NC2–TR–1998–029, Royal Holloway, University of London, London, UK, 1998.

[21] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[22] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[23] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, 1990.

[24] R.C. Williamson, A.J. Smola, B. Schölkopf, Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators, Technical Report NC-TR-98-019, Royal Holloway College, University of London, London, UK, 1998 http://www.neurocolt.com.

Jinbo Bi is a Ph.D. student in the Department of Mathematical Sciences at Renselar Polytechnic Institute, working on her thesis “Support vector regression with applications in automated drug discovery”. She received her M.Sc. in Automatic Control and her B.Sc. in Applied Mathematics from Beijing Institute of Technology in China. Her research interests include mathematical programming, statistical learning, machine learning, kernel methods and support vector machines.

Kristin P. Bennett is an Associate Professor in the Mathematical Sciences Department at Rensselaer Polytechnic Institute. She is an active member of the machine learning, data mining, and operations research communities serving an associate editor for SIAM Journal on Optimization, Naval Research Logistics, Machine Learning Journal, and Journal on Machine Learning Research, and as a program committee member for AAAI, Machine Learning, SIGKDD Knowledge Discovery and Data Mining, IEEE Data Mining, and SIAM Data Mining Conferences. She has a Ph.D. and M.S. in Computer Sciences from University of Wisconsin-Madison, and a B.S. in mathematics and computer science from University of Puget Sound. She has been working in support vector machines and closely related mathematical-programming approaches to learning since 1989.