Convex Tube Regression

Xiaolin Huang, Lei Shi, Kristiaan Pelckmans, and Johan A. K. Suykens,

***Abstract*—For a given set of data, tube regression tries to find a function f and a small positive constant t such that at least a certain percentage of the training data are contained in the tube $\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : y \in [f(x) - t, f(x) + t]\}$. The data falling outside the tube are regarded as outliers and they have no effect on the result of tube regression. Tube regression is a distribution-free and robust method but is non-convex and hard to compute. With the help of the pinball loss, in this paper, a convex approach for tube regression is proposed and called convex tube regression (CTR). The dual problem of CTR is derived and then a nonparametric tube regression method based on positive definite kernels is established. CTR can be regarded as an extension to ν-support vector machine (ν-SVM) and support/tolerance vector tubes (SVT). CTR enjoys a similar sparsity property as ν-SVM and has a better performance for noise following a skewed distribution. Numerical experiments illustrate the effectiveness of the proposed method.**

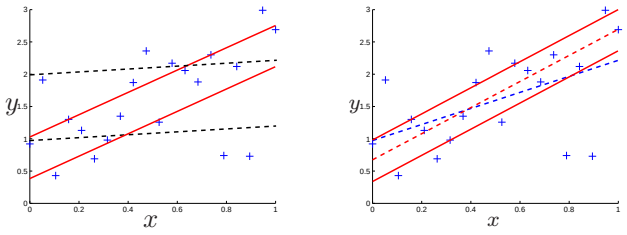
***Index Terms*—robust regression, convex approach, nonparametric model.**

I. Introduction

In the regression field, we attempt to establish a function $f(x)$ mapping $x \in \mathbb{R}^d$ to $y \in \mathbb{R}$, according to a given set of observed data $\{(x_i, y_i)\}_{i=1}^n$. Typically, $f(x)$ is obtained from minimizing an error criterion. For example, the minimizer of sum of the squared errors, i.e., $\arg\min_f \sum_{i=1}^n (y_i - f(x_i))^2$, is called the least squares method. Due to the existence of noise, in many applications, we are not only interested in the predicted value $\hat{f}(x)$ but also in the corresponding prediction interval, that means that the probability of the real value belonging to $[\hat{f}(x) - t_\rho, \hat{f}(x) + t_\rho]$ is not less than ρ, where t_ρ is the width of the prediction interval for probability ρ. The analysis for the prediction interval depends on the distribution of the noise, which is usually not known in advance. In this paper, we consider an alterative regression method, which pursues a tube $[f - t, f + t]$ covering a certain percentage of the observed data. To illustrate this idea, we consider linear regression for the data shown by blue crosses in Fig. 1. In this example, we are looking for a linear tube covering 50% of the data. There are many tubes satisfying this requirement. Two of such tubes are shown by the red solid lines and black dashed lines in Fig. 1(a). Intuitively, a tube with a small width gives a precise prediction. Hence, we assume that the tube of the red solid lines is better for prediction than that of the black dashed lines. Therefore, we should try to minimize t with the constraint that a certain percentage of the training data are covered by $[f - t, f + t]$. From its geometrical interpretation, we call this method *tube regression*. Unlike prediction interval analysis, tube regression is distribution-free and its statistical property has been discussed in [1]. When f is chosen as an affine function, i.e., $f(x) = w^T x + b$, we can formulate the (linear) tube regression as below,

$$\begin{aligned} \min_{w,b,t} \quad & t \\ \text{s.t.} \quad & \sum_{i=1}^{n} I\left(y_i \in [w^T x_i + b - t, w^T x_i + b + t]\right) \geq \rho n, \end{aligned} \tag{1}$$

where $I(a)$ stands for an indicator function, which equals one when a is true and equals zero otherwise, and ρ is a user defined constant. One can choose its value according to the application or knowledge about the noise distribution. Generally, a small ρ is suitable for heavy noise and one prefers a large ρ when the sampling density is low.



(a) (b)

Fig. 1. Illustration of tube regression. (a) The data shown by blue crosses are covered by two tubes, shown by red solid and black dashed lines. The tube of the red solid lines has a smaller width hence it is better for prediction than the other one; (b) Tube regression (red lines) is more robust than the least squares method (blue dashed line).

In the regression field, we prefer good generalization capability and low complexity, which can be measured by the norm of w. Thus, $\|w\|$ and (1) should be considered together in a bi-criterion optimization problem, as discussed in [2]. To solve the bi-criterion problem, we can use an additional constraint

Manuscript received 2013;

This work was supported in part by the scholarship of the Flemish Government; Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), G.0377.12 (Structured models), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); IBBT; EU: ERNSI; ERC AdG A-DATADRIVE-B, FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other: Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger. L. Shi is also supported by the National Natural Science Foundation of China (11201079). Johan Suykens is a professor at KU Leuven, Belgium.

X. Huang, L. Shi, and J.A.K. Suykens are with the Department of Electrical Engineering, ESAT-SCD-SISTA, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. (e-mails: huangxl06@mails.tsinghua.edu.cn, leishi@fudan.edu.cn, johan.suykens@esat.kuleuven.be). K. Pelckmans is with the Department of Information Technology, the Uppsala University, SE-751 05 Uppsala, Sweden. (e-mail: kp@it.uu.se).

$\|w\| \leq D$ for (1) and consider different $\|\cdot\|$, e.g., l_1 norm leads to a sparse result. Another and more popular regularization method is the weighted sum of $\|w\|$ and t. Following this discussion, we use l_2 norm with trade-off parameter γ and establish the Tikhonov regularization tube regression as below,

$$\begin{aligned} \min_{w,b,t} \quad & \frac{1}{2\gamma} w^T w + t \\ \text{s.t.} \quad & \sum_{i=1}^{n} I\left(y_i \in [w^T x_i + b - t, w^T x_i + b + t]\right) \geq \rho n. \end{aligned} \tag{2}$$

The data falling outside the tube, which may contain outliers, has no effect on the tube width and the result of tube regression is robust. In Fig. 1, there are two points in the right-bottom part, which may be away from their real values and can be regarded as outliers. The boundaries and the middle line of a tube with a small length are shown in Fig. 1(b) by red solid and dashed lines, respectively. One can see that though this tube is not a unique optimal solution to (2), it gives a good result. Tube regression (2) is not sensitive to outliers. In contrast, the result of the least squares method, shown by a blue dashed line, is significantly affected by the outliers. In fact, the concept of tube regression has appeared in the robust regression field as least medium squares (LMS) regression [3] [4]. Denote the k-th maximum of $\{u_i\}_{i=1}^n$ by $\max_{1\leq i\leq n}^k\{u_i\}$:

$$\max_{1\leq i\leq n}^k\{u_i\} = u_{\Gamma(k)} \quad \text{with} \quad u_{\Gamma(1)} \geq u_{\Gamma(2)} \geq \cdots \geq u_{\Gamma(n)}.$$

Then the linear LMS estimator can be written as

$$\min_{w,b} \quad \max_{1\leq i\leq n}^k \left\{\left(w^T x_i + b - y_i\right)^2\right\}, \tag{3}$$

which is equivalent to (1) for $k = \min\{u : \frac{n-u+1}{n} \geq \rho\}$. A regularization term $w^T w$ can be added for the LMS regression as well. When the medium squared error is minimized, (3) is regarded as the most robust estimator in view of the breakdown point [5]. Tube regression leads to a tube with a small width containing a required percentage of the data and the result is robust to outliers. However, tube regression or LMS estimation is not convex unless $\rho = 1$ for (2) or $k = 1$ for (3). When $\rho = 1$, we want the tube to cover all the data, resulting in l_∞ regression or minimax approximation [6]. One can see that this estimation is very sensitive to outliers. Then tube regression can be seen as a robust extension to minimax approximation and is applicable in many fields involving l_∞ norm.

Another topic closely related to tube regression is quantile regression, see [7] for parametric methods and [8] [9] for nonparametric methods. The q-quantile regression gives a function f^q satisfying $\text{Prob}\{y \leq f^q(x)\} = q$. By quantile regression, we can get a q_1-quantile function f^{q_1} and a q_2-quantile function f^{q_2}. If the two functions are parallel, they can be used to construct a tube $[f^{q_1}, f^{q_2}]$. When $q_1 - q_2 = \rho$, this tube covers the required ρ percentage of the data. Motivated by the relationship between quantile regression and tube regression, we will construct a convex formulation with the help of the pinball loss to approximately solve tube regression. Then a dual formulation is given and a nonparametric tube regression method is established.

As mentioned previously, tube regression (2) and LMS estimation (3) are non-convex. There have been some approximation algorithms available in [10] [11] [12] [13] and [14]. The most popular method for computing LMS estimator is PROGRESS suggested by [4] and modified by [15]. The method recently proposed in [1] also can be seen as a convex approach for tube regression. Besides of these heuristics, there have been several algorithms for finding the global optimum, see [16] [17] and [18]. Though we cannot theoretically show the superiority of the proposed method over the existing algorithms, we evaluate its performance on numerical experiments. Moreover, via the dual problem of the proposed convex formulation, tube regression using positive definite kernels is developed. To the best of our knowledge, nonparametric models corresponding to other tube regression methods have not been available until now.

The remainder of the paper is organized as follows: a convex formulation for tube regression is given in Section II. Section III discusses its dual problem and constructs a nonparametric tube regression model. Then the proposed methods are evaluated by numerical experiments in Section IV. Section V ends the paper with concluding remarks.

II. Tube Regression with the Pinball Loss

A. Quantile regression and the pinball loss

To construct a convex formulation for tube regression, we first investigate quantile regression and the related pinball loss. The pinball loss $L_\tau(u)$ can be defined as

$$L_\tau(u) = \begin{cases} \frac{1}{\tau}u, & u \geq 0, \\ -u, & u < 0, \end{cases}$$

where τ is a positive constant. Then a $\frac{1}{1+\tau}$-quantile linear function can be constructed by minimizing the sum of the pinball losses on the training data, i.e.,

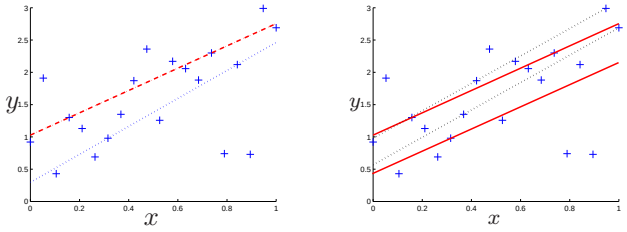
$$\begin{aligned} \min_{w,t,\xi^+} \quad & \sum_{i=1}^{n} L_\tau(\xi_i^+) \\ \text{s.t.} \quad & \xi_i^+ = y_i - (w^T x_i + b), \forall i = 1, \cdots, n. \end{aligned} \tag{4}$$

In order to explain intuitively the quantile property of the sum of the pinball losses, let us consider a unit increment of $f(x_i)$, denoted by Δf, which does not change the sign of ξ_i^+ for $i : \xi_i^+ \neq 0$. Then, such increment will make the penalty of the data with $\xi_i^+ > 0$ decrease $\frac{1}{\tau}\Delta f$ and the penalty on data with $\xi_i^+ < 0$ increase Δf. Therefore, for the optimal solution to (4), τ controls the ratio of the numbers of elements in $\{i : \xi_i^+ > 0\}$ and $\{i : \xi_i^+ < 0\}$. Hence (4) results in a $\frac{1}{1+\tau}$-quantile function. Similarly, the solution to the following optimization problem leads to a $\frac{\tau}{1+\tau}$-quantile linear function,

$$\begin{aligned} \min_{w,b,\xi^-} \quad & \sum_{i=1}^{n} L_\tau(\xi_i^-) \\ \text{s.t.} \quad & \xi_i^- = w^T x_i + b - y_i, \forall i = 1, \cdots, n. \end{aligned} \tag{5}$$

Consider again the data shown in Fig. 1. We can use (4) to find a line, above which there are 5 data points. This line is shown in Fig. 2(a) and it can be used as the upper boundary of the desired tube. In the same way, the lower boundary can

be constructed via (5), of which the result is illustrated by a blue dotted line. If we can use the two lines as the boundaries, a tube covering 50% of the training data can be constructed.



(a) (b)

Fig. 2. Quantile regression and tube regression for data shown by crosses. (a) The red dashed line represents a quantile function, above which there are 5 data points; the blue dotted line represents another quantile function, below which there are 5 data points; (b) The results of (6) are illustrated by red solid lines ($\tau_+ = \tau_- = 0.25$) and black dotted lines ($\tau_+ = 0.1, \tau_- = 0.4$), the latter of which is the optimal solution to (2). These two tubes cover the same number of data, but the amounts of the data above or below them are different.

B. Convex tube regression

Via (4) and (5), we get the desired quantile functions. However, they are not parallel and cannot be used for tube regression directly. When two lines are parallel, they can be represented by their middle value and the distance between them as $w^T x + b \pm t$, where t should be nonnegative. Function $w^T x + b + t$ and $w^T x + b - t$ correspond to the boundaries of a tube, that means $w^T x + b + t$ and $w^T x + b - t$ provide good solutions to (4) and (5), respectively. Motivated by this observation, we consider (4) and (5) together for the requirement of covering a certain percentage of data, i.e., the constraint in (2), then we obtain the following convex approach for tube regression,

$$\begin{aligned} \min_{w,b,t,\xi^+,\xi^-} \quad & \frac{1}{2\gamma} w^T w + Ct + \sum_{i=1}^{n} L_{\tau_+}(\xi_i^+) + \sum_{i=1}^{n} L_{\tau_-}(\xi_i^-) \\ \text{s.t.} \quad & \xi_i^+ = y_i - (w^T x_i + b + t), \forall i = 1, \cdots, n, \\ & \xi_i^- = (w^T x_i + b - t) - y_i, \forall i = 1, \cdots, n, \\ & t \geq 0. \end{aligned} \tag{6}$$

To distinguish with the original tube regression, we call (6) *convex tube regression* (CTR). Certainly, CTR is not equivalent to (2), which is a non-convex problem. Though (6) and (2) are not the same, CTR still enjoys robustness to outliers, which results from the fact the loss $L_{\tau_+}(\xi_i^+) + L_{\tau_-}(\xi_i^-)$ has bounded derivatives. When t is given, $L_{\tau_+}(\xi_i^+) + L_{\tau_-}(\xi_i^-)$ is a function of $w^T x_i + b$ and illustrated in Fig. 3 by two lines corresponding to $\tau_+ = \tau_- = 0.25, t = 1$ (blue solid line) and $\tau_+ = 0.1, \tau_- = 0.2, t = 0.5$ (red dashed line). The loss function consists of two non-zero slope parts and one flat part, which performs like an insensitive zone. t gives the width and the value of the flat part, and τ_+, τ_- determine the slopes of the other parts.

One noticeable point is that though minimizing the sum of the pinball losses results in a quantile function, considering t, (4) and (5) together does not necessarily lead to the quantile functions defined by (4) or (5). The basic character of the

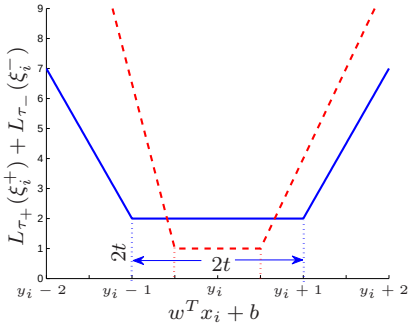


Fig. 3. Loss value $L_{\tau_+}(\xi_i^+) + L_{\tau_-}(\xi_i^-)$: $\tau_+ = \tau_- = 0.25, t = 1$ (blue solid line) and $\tau_+ = 0.1, \tau_- = 0.2, t = 0.5$ (red dashed line).

pinball loss is that the ratio of its slopes on the positive interval and negative interval is a constant, which results in the quantile property for (4) as mentioned in subsection II-A. Based on this character, in (6), we pursue a small width, require parallelism of the boundary lines and still have the quantile-type property: the tube obtained by (6) can cover the required percentage of the training data, and the locations of the boundaries are controlled by τ_+ and τ_-, as shown in the following proposition.

Proposition 1: For the fixed $\tau_+, \tau_- \geq 0$, when $C = 2\min\{\tau_+, \tau_-\}n$, the optimal solution to (6) satisfies

$$\begin{aligned} \sum_{i=1}^{n} I\left(y_i \leq w^T x_i + b + t\right) &\geq (1 - \tau_+)n, \\ \sum_{i=1}^{n} I\left(y_i \geq w^T x_i + b - t\right) &\geq (1 - \tau_-)n, \end{aligned}$$

and therefore,

$$\sum_{i=1}^{n} I\left(y_i \in \left[w^T x_i + b - t, w^T x_i + b + t\right]\right) \geq (1 - \tau_+ - \tau_-)n.$$

Proof: First, (6) can be equivalently transformed into

$$\begin{aligned} \min_{w,b,t,\xi^+,\xi^-} \quad & \frac{1}{2\gamma} w^T w + Ct + \sum_{i=1}^{n} \xi_i^+ + \sum_{i=1}^{n} \xi_i^- \\ \text{s.t.} \quad & w^T x_i + b + t - y_i \leq \xi_i^+, \forall i = 1, \cdots, n, \\ & -(w^T x_i + b + t - y_i) \leq \tau_+ \xi_i^+, \forall i = 1, \cdots, n, \\ & w^T x_i + b - t - y_i \leq \tau_- \xi_i^-, \forall i = 1, \cdots, n, \\ & -(w^T x_i + b - t - y_i) \leq \xi_i^-, \forall i = 1, \cdots, n, \\ & t \geq 0. \end{aligned} \tag{7}$$

By introducing Lagrange multipliers $\alpha_i^+, \beta_i^+, \alpha_i^-, \beta_i^-, i = 1, \cdots, n$ and ζ, which correspond to the above constraints and are nonnegative, we get the following Lagrangian

$$\begin{aligned} & \mathcal{L}(w, b, t, \xi^+, \xi^-; \alpha^+, \beta^+, \alpha^-, \beta^-, \zeta) \\ = \quad & \frac{1}{2\gamma} w^T w + Ct + \sum_{i=1}^{n} \xi_i^+ + \sum_{i=1}^{n} \xi_i^- - \zeta t \\ & + \sum_{i=1}^{n} \alpha_i^+ (w^T x_i + b + t - y_i - \xi_i^+) \end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=1}^{n} \beta_i^+ (w^T x_i + b + t - y_i + \tau_+ \xi_i^+) \\
&+ \sum_{i=1}^{n} \alpha_i^- (w^T x_i + b - t - y_i - \tau_- \xi_i^-) \\
&- \sum_{i=1}^{n} \beta_i^- (w^T x_i + b - t - y_i + \xi_i^-).
\end{aligned}
$$

According to the saddle point condition, we have that

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial t} &= C - \zeta + \sum_{i=1}^{n} (\alpha_i^+ - \beta_i^+ - \alpha_i^- + \beta_i^-) = 0, \quad (8) \\
\frac{\partial \mathcal{L}}{\partial b} &= \sum_{i=1}^{n} (\alpha_i^+ - \beta_i^+ + \alpha_i^- - \beta_i^-) = 0, \quad (9) \\
\frac{\partial \mathcal{L}}{\partial \xi_i^+} &= 1 - \alpha_i^+ - \tau_+ \beta_i^+ = 0, i = 1, 2, \cdots, N, \quad (10) \\
\frac{\partial \mathcal{L}}{\partial \xi_i^-} &= 1 - \tau_- \alpha_i^- - \beta_i^- = 0, i = 1, 2, \cdots, N. \quad (11)
\end{aligned}
$$

Substituting (10) and (11) into (8) and (9) results in the following two conditions,

$$
\begin{aligned}
C + 2n - \sum_{i=1}^{n} (1 + \tau_+) \beta_i^+ - \sum_{i=1}^{n} (1 + \tau_-) \alpha_i^- \geq 0, \\
\sum_{i=1}^{n} (1 + \tau_+) \beta_i^+ - \sum_{i=1}^{n} (1 + \tau_-) \alpha_i^- = 0.
\end{aligned}
$$

Therefore,

$$
\sum_{i=1}^{n} (1 + \tau_+) \beta_i^+ = \sum_{i=1}^{n} (1 + \tau_-) \alpha_i^- \leq n + \frac{C}{2}. \quad (12)
$$

When $C = 2 \min\{\tau_+, \tau_-\} n$, we have

$$
\sum_{i=1}^{n} \beta_i^+ \leq n \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i^- \leq n. \quad (13)
$$

For a point above the tube, the solution to (7) satisfies

$$
w^T x_i + b + t - y_i < \xi_i^+, \ \ -(w^T x_i + b + t - y_i) = \tau_+ \xi_i^+.
$$

According to the complementary slackness condition and condition (10), for such data, the corresponding dual variables satisfy $\alpha_i^+ = 0, \beta_i^+ = \frac{1}{\tau_+}$. From (13), we can conclude that the number of data falling above the tube is not larger than $\tau_+ n$. A similar analysis tells us that the number of data below the tube is not larger than $\tau_- n$. Then the number of data falling outside the tube is not larger than $(\tau_+ + \tau_-) n$. ■

From the above proposition, to pursue a linear tube covering ρ percentage of the data, we can set τ_+, τ_- such that $\tau_+ + \tau_- = 1 - \rho$ and solve CTR (6). And the location of the tube can be controlled by τ_+, τ_-. Now we illustrate the performance of CTR on the data shown in Fig. 2. To obtain a tube covering $\rho = 50\%$ data, we set $\tau_+ = \tau_- = 0.25$ and solve (6). Above the obtained tube, there are 5 data points and below it there are 5 data points. For a given pair of τ_+, τ_-, CTR considers only one case of the localization of the tube. To solve tube regression (2), we should consider different values of τ_+, τ_- with $\tau_+ + \tau_- = 1 - \rho$. For example, $\tau_+ = 0.1, \tau_- = 0.4$ results in the optimal solution to (2), shown by a black dotted line, for which there are 2 data points above the tube and 8 data points below it.

In (6), the percentage of the data covered by the tube is mainly controlled by τ_+ and τ_-. Hence, though we prove Proposition 1 for $C = 2 \min\{\tau_+, \tau_-\} n$, we can use a larger C to pursue a smaller width. When $\sum_{i=1}^{n} (1 + \tau_+) \beta_i^+ < n + \frac{C}{2}$, we have $\zeta > 0$ and $t = 0$. A similar relationship holds for α_i^-. Hence, the maximal reasonable value of C is $\min\{\frac{2n}{\tau_+}, \frac{2n}{\tau_-}\}$. For $0 < C < \min\{\frac{2n}{\tau_+}, \frac{2n}{\tau_-}\}$, the optimal t is strictly larger than zero, i.e., $t \geq 0$ is redundant. But in CTR (6), we keep this constraint, since when C larger than $\min\{\frac{2n}{\tau_+}, \frac{2n}{\tau_-}\}$, $t \geq 0$ will be active. In this case, $t = 0$ is optimal to (6) and then (6) reduces to quantile regression. In practice, we always set $C = n$, for which we can not give the upper bound for data outside the tube. But in most cases, the obtained tube has a small width and contains the required percentage of the data.

C. Relation to ν-SVM and SVT

In this subsection, we discuss the relationship between the proposed CTR and two other existing methods, ν support vector machine (ν-SVM, [19]) and support/tolerance vector tubes (SVT, [1]). We consider the case $\tau_+ = \tau_-$ and set $\gamma' = \tau_+ \gamma$, $C' = \frac{C}{\tau_+}$. Then (7) can be equivalently transformed into

$$
\begin{aligned}
\min_{w,b,t,\xi^+,\xi^-} \quad & \frac{1}{2\gamma'} w^T w + C' t + \sum_{i=1}^{n} \xi_i^+ + \sum_{i=1}^{n} \xi_i^- \\
\text{s.t.} \quad & w^T x_i + b + t - y_i \leq \frac{1}{\tau_+} \xi_i^+, \forall i = 1, \cdots, n, \\
& -(w^T x_i + b + t - y_i) \leq \xi_i^+, \forall i = 1, \cdots, n, \\
& w^T x_i + b - t - y_i \leq \xi_i^-, \forall i = 1, \cdots, n, \\
& -(w^T x_i + b - t - y_i) \leq \frac{1}{\tau_-} \xi_i^-, \forall i = 1, \cdots, n, \\
& t \geq 0.
\end{aligned}
$$

When τ_+ tends to zero, the first constraint in the above becomes to $\xi_i^+ \geq 0$. Similarly, the fourth constraint becomes $\xi_i^- \geq 0$ for $\tau_- = 0$. Accordingly, when $\tau_+ = \tau_- = 0$, CTR reduces to

$$
\begin{aligned}
\min_{w,b,t,\xi^+,\xi^-} \quad & \frac{1}{2\gamma'} w^T w + C' t + \sum_{i=1}^{n} \xi_i^+ + \sum_{i=1}^{n} \xi_i^- \\
\text{s.t.} \quad & y_i - (w^T x_i + b) \leq t + \xi_i^+, \forall i = 1, \cdots, n, \\
& w^T x_i + b - y_i \leq t + \xi_i^-, \forall i = 1, \cdots, n, \quad (14) \\
& t \geq 0, \ \xi_i^+ \geq 0, \ \xi_i^- \geq 0, \forall i = 1, \cdots, n.
\end{aligned}
$$

This is the well-known ν-SVM, which is an extension to support vector machine regression (SVR, [20]). In SVR, the ε-insensitive zone loss function is used. There is no penalty on the points in the ε-insensitive zone, which can also be regarded as a tube. Then ν-SVM pursues a small ε-insensitive zone by adding the term ε in the objective.

Furthermore, we find that the first and the second constraint of (14) are equal to

$$\begin{array}{rcl} y_i - (w^T x_i + b) & \leq & t + \max\{\xi_i^+, \xi_i^-\}, \\ w^T x_i + b - y_i & \leq & t + \max\{\xi_i^+, \xi_i^-\}. \end{array} \quad (15)$$

Obviously, the feasible set of (15) covers that of the first and the second constraint of (14). Hence, to verify the equivalence between (15) and the first and the second constraint in (14), we need to show that any feasible solution to (15) is related to a feasible solution to (14). Consider a given pair of w, b and suppose $y_i - (w^T x_i + b) \leq 0$. If $\tilde{\xi}_i^+, \tilde{\xi}_i^-$ are feasible for (15), then $\xi_i^+ = \max\{\tilde{\xi}_i^+, \tilde{\xi}_i^-\}$, $\xi_i^- = 0$ are feasible for (14) and correspond to a smaller or equivalent objective value of $\tilde{\xi}_i^+, \tilde{\xi}_i^-$. A similar discussion holds for the case of $y_i - (w^T x_i + b) > 0$. Thus, the first and the second constraint of (14) can be replaced by (15).

The left side of the first constraint of (14), i.e., $y_i - (w^T x_i + b)$, and the left side of the second one, i.e., $w^T x_i + b - y_i$, are of opposite signs. Thus, for the optimal solution to (14), we have $\xi_i^+ \xi_i^- = 0$. Therefore, by defining $\xi_i = \max\{\xi_i^+, \xi_i^-\}$, we have $\xi_i = \xi_i^+ + \xi_i^-$ and hence (14) can be transformed into

$$\begin{array}{ll} \min\limits_{w,b,t,\xi} & \frac{1}{2\gamma'} w^T w + C' t + \sum\limits_{i=1}^n \xi_i \\ \text{s.t.} & y_i - (w^T x_i + b) \leq t + \xi_i, \forall i = 1, \cdots, n, \\ & w^T x_i + b - y_i \leq t + \xi_i, \forall i = 1, \cdots, n, \\ & t \geq 0,\ \xi_i \geq 0, \forall i = 1, \cdots, n. \end{array} \quad (16)$$

From the above discussion, we can see the relationship between CTR and another convex formulation for covering a certain percentage of the training data. This method, called support/tolerance vector tube (SVT), was proposed by [1] and has the following formulation,

$$\begin{array}{ll} \min\limits_{w,b,t,\xi} & C' t + \sum\limits_{i=1}^n \xi_i \\ \text{s.t.} & -t - \xi_i \leq w^T x_i + b - y_i \leq t + \xi_i, \\ & \xi_i \geq 0, \forall i = 1, \cdots, n. \end{array} \quad (17)$$

For SVT, a property similar to Proposition 1 has been proved and the percentage of the data falling outside can be upper bounded. By comparison of (17) and (16), it can be seen that the difference is that $t \geq 0$ and $w^T w$ are included in (16). As discussed in subsection II-B, when a moderate C' is used, $t \geq 0$ can be excluded but when C' is too large, $t \geq 0$ should be considered in (16). Generally, $t \geq 0$ is not a crucial difference and we can conclude that CTR with $\tau_+ = \tau_- = 0$ is the l_2 regularized form of SVT. Besides the regularization term, the extension of CTR from SVT is mainly that various locations of the tube can be considered by different values of τ_+, τ_-. This extension may lead to a better solution for tube regression (2) as indicated in Fig. 2.

III. Nonparametric Convex Tube Regression

A. Dual problem

The proposed convex tube regression (6) is a convex approach for tube regression (2). In the pure view of solving a non-convex problem, we cannot claim its superiority over some heuristics and we will evaluate its performance by a numerical study. One interesting aspect is that via the convex formulation, tube regression based on positive definite kernels can be established.

We first introduce a nonlinear feature map $\phi(x_i)$ and replace x_i in (6) with $\phi(x_i)$, resulting in the following formulation,

$$\begin{array}{ll} \min\limits_{w,b,t,\xi^+,\xi^-} & \frac{1}{2\gamma} w^T w + Ct + \sum\limits_{i=1}^n L_{\tau_+}(\xi_i^+) + \sum\limits_{i=1}^n L_{\tau_-}(\xi_i^-) \\ \text{s.t.} & \xi_i^+ = y_i - (w^T \phi(x_i) + b + t), \forall i, \\ & \xi_i^- = (w^T \phi(x_i) + b - t) - y_i, \forall i, \\ & t \geq 0, \end{array}$$

which can be transformed into

$$\begin{array}{ll} \min\limits_{w,b,t,\xi^+,\xi^-} & \frac{1}{2\gamma} w^T w + Ct + \sum\limits_{i=1}^n \xi_i^+ + \sum\limits_{i=1}^n \xi_i^- \\ \text{s.t.} & w^T \phi(x_i) + b + t - y_i \leq \xi_i^+, \forall i, \\ & -(w^T \phi(x_i) + b + t - y_i) \leq \tau_+ \xi_i^+, \forall i, \\ & w^T \phi(x_i) + b - t - y_i \leq \tau_- \xi_i^-, \forall i, \\ & -(w^T \phi(x_i) + b - t - y_i) \leq \xi_i^-, \forall i, \\ & t \geq 0. \end{array}$$

Introduce Lagrange multipliers $\alpha_i^+, \beta_i^+, \alpha_i^-, \beta_i^-, i = 1, \cdots, n$ and ζ, which are all nonnegative. The Lagrangian is

$$\begin{array}{rl} & \mathcal{L}(w, b, t, \xi^+, \xi^-; \alpha^+, \beta^+, \alpha^-, \beta^-, \zeta) \\ = & \frac{1}{2\gamma} w^T w + Ct + \sum\limits_{i=1}^n \xi_i^+ + \sum\limits_{i=1}^n \xi_i^- - \zeta t \\ & + \sum\limits_{i=1}^n \alpha_i^+ (w^T \phi(x_i) + b + t - y_i - \xi_i^+) \\ & - \sum\limits_{i=1}^n \beta_i^+ (w^T \phi(x_i) + b + t - y_i + \tau_+ \xi_i^+) \\ & + \sum\limits_{i=1}^n \alpha_i^- (w^T \phi(x_i) + b - t - y_i - \tau_- \xi_i^-) \\ & - \sum\limits_{i=1}^n \beta_i^- (w^T \phi(x_i) + b - t - y_i + \xi_i^-). \end{array}$$

According to the saddle point condition, we have

$$\begin{array}{rcl} \frac{\partial \mathcal{L}}{\partial t} & = & C - \zeta + \sum\limits_{i=1}^n (\alpha_i^+ - \beta_i^+ - \alpha_i^- + \beta_i^-) = 0, \\ \frac{\partial \mathcal{L}}{\partial w} & = & \frac{1}{\gamma} w + \sum\limits_{i=1}^n (\alpha_i^+ - \beta_i^+ + \alpha_i^- - \beta_i^-) \phi(x_i) = 0, \\ \frac{\partial \mathcal{L}}{\partial b} & = & \sum\limits_{i=1}^n (\alpha_i^+ - \beta_i^+ + \alpha_i^- - \beta_i^-) = 0, \\ \frac{\partial \mathcal{L}}{\partial \xi_i^+} & = & 1 - \alpha_i^+ - \tau_+ \beta_i^+ = 0, i = 1, 2, \cdots, N, \\ \frac{\partial \mathcal{L}}{\partial \xi_i^-} & = & 1 - \tau_- \alpha_i^- - \beta_i^- = 0, i = 1, 2, \cdots, N. \end{array}$$

Besides, there are $\alpha_i^+, \beta_i^+, \alpha_i^-, \beta_i^- \geq 0, \forall i$ and $\zeta \geq 0$. Therefore, by introducing new dual variables $\lambda_i = \gamma(1 + \tau_+)\beta_i^+$,

$\lambda_i^* = \gamma(1+\tau_-)\alpha_i^-$ and a positive definite kernel $\mathcal{K}$ with $\mathcal{K}_{ij} = \phi(x_i)^T\phi(x_j)$, we obtain the following *nonparametric CTR* formulation,

$$\begin{aligned}\max_{\lambda,\lambda^*} \quad & -\frac{1}{2}(\lambda-\lambda^*)^T\mathcal{K}(\lambda-\lambda^*) + \sum_{i=1}^n y_i(\lambda_i-\lambda_i^*) \\ \text{s.t.} \quad & \sum_{i=1}^n(\lambda_i-\lambda_i^*) = 0, \\ & \sum_{i=1}^n(\lambda_i+\lambda_i^*) \le \gamma(2n+C), \\ & 0 \le \lambda_i \le \left(1+\frac{1}{\tau^+}\right)\gamma, \forall i = 1,\cdots,n, \\ & 0 \le \lambda_i^* \le \left(1+\frac{1}{\tau^-}\right)\gamma, \forall i = 1,\cdots,n.\end{aligned} \tag{18}$$

In (18), we can use any positive-definite kernel, such as radial basis function (RBF) kernel or polynomial kernel, to obtain a nonparametric tube. After solving (18), we get the optimal dual variables λ, λ^*, and calculate $w^T\phi(x)$ by

$$w^T\phi(x) = \sum_{i=1}^n(\lambda_i-\lambda_i^*)\mathcal{K}(x,x_i).$$

To compute the bias term b and the width of the tube t, we consider data x_i with $0 < \beta_i < 1+\frac{1}{\tau^+}$, denoted by $\mathcal{S}_0^+ = \{i : 0 < \beta_i < 1+\frac{1}{\tau^+}\}$. These data are located on the upper boundary of the tube, i.e.,

$$\sum_{j=1}^n(\lambda_j-\lambda_j^*)\mathcal{K}(x_j,x_i) + b + t - y_i = 0, \forall i \in \mathcal{S}_0^+.$$

Similarly, for data in $\mathcal{S}_0^- = \{i : 0 < \alpha_i < 1+\frac{1}{\tau^-}\}$, there is

$$\sum_{j=1}^n(\lambda_j-\lambda_j^*)\mathcal{K}(x_j,x_i) + b - t - y_i = 0, \forall i \in \mathcal{S}_0^-.$$

Using one element in $\mathcal{S}_0^+$ and one element in $\mathcal{S}_0^-$, the bias term b and the tube width t can be calculated. After that, we get the nonparametric formulation of a tube covering a certain percentage of the data. The middle of the tube can be expressed by the dual variables as

$$f(x) = \sum_{i=1}^n(\lambda_i-\lambda_i^*)\mathcal{K}(x,x_i) + b, \tag{19}$$

the upper and the lower boundary of the tube are $f(x)+t$ and $f(x)-t$, respectively.

B. Sparseness of nonparametric tube regression

Similarly to Proposition 1, the percentage of the data covered by the nonparametric convex tube (18) can be guaranteed by the following proposition.

Proposition 2: When $C = 2\min\{\tau_+,\tau_-\}n$, the optimal solution to (18) satisfies

$$\begin{aligned}&\sum_{i=1}^n I\left(y_i \le \textstyle\sum_{j=1}^n(\lambda_j-\lambda_j^*)\mathcal{K}(x,x_j)+b+t\right) \ge (1-\tau_+)n, \\ &\sum_{i=1}^n I\left(y_i \ge \textstyle\sum_{j=1}^n(\lambda_j-\lambda_j^*)\mathcal{K}(x,x_j)+b-t\right) \ge (1-\tau_-)n,\end{aligned}$$

and

$$\sum_{i=1}^n I\left(y_i \in \left[\sum_{j=1}^n(\lambda_j-\lambda_j^*)\mathcal{K}(x,x_j)+b-t, \sum_{j=1}^n(\lambda_j-\lambda_j^*)\mathcal{K}(x,x_j)+b+t\right]\right) \ge (1-\tau_+-\tau_-)n.$$

This proposition comes from the constraints of (18) and the complementary slackness condition. The proof is similar to that of Proposition 1 and is omitted here. The percentage of the data covered by the tube and the location of the tube are mainly controlled by τ_+ and τ_-. As discussed for the linear formulation (6), we always set $C = n$ and adjust γ only.

We are also interested in the sparseness of nonparametric CTR. The upper boundary of the tube divides the training set into two parts. If a point (x_i, y_i) is above the tube, i.e., $y_i > w^T\phi(x_i)+b+t$, we know from the complementary slackness condition that $\alpha_i^+ = 0, \beta_i^+ = \frac{1}{\tau_+}$. Otherwise, when a point is below the upper boundary, it can be within or below the tube and we have $\alpha_i^+ = 1, \beta_i^+ = 0$. Similar analysis holds for α_i^- and β_i^-. Noticing $\lambda_i = \gamma(1+\tau_+)\beta_i^+$, $\lambda_i^* = \gamma(1+\tau_-)\alpha_i^-$, we know the following properties:

- Data above the tube: $\lambda_i - \lambda_i^* = (1+\frac{1}{\tau^+})\gamma$;
- Data within the tube: $\lambda_i - \lambda_i^* = 0$;
- Data below the tube: $\lambda_i - \lambda_i^* = -(1+\frac{1}{\tau^-})\gamma$.

Besides, for a point on the boundary, there is $-(1+\frac{1}{\tau^-})\gamma < \lambda_i - \lambda_i^* < (1+\frac{1}{\tau^+})\gamma$. From this discussion, one can see that the data within the tube have zero dual variables, which is similar to the sparseness property of ν-SVM. This phenomenon can also be observed from Fig. 3, which illustrates that the loss is constant for data falling into the tube. As shown in Proposition 2, the number of points covered by the tube is controlled by τ_+ and τ_-. Following this, the sparsity, i.e., the fraction of the zero dual variables, of the nonparametric CTR (18) can be controlled by τ_+ and τ_-.

The sparsity comes from the nature of tube regression. Suppose that the tube $\left[\tilde{f}-\tilde{t}, \tilde{f}+\tilde{t}\right]$ is the optimal solution to tube regression (2). The training part can be partitioned into the following subsets

$$\begin{aligned}\tilde{\mathcal{S}}_{\text{out}} &= \left\{i : y_i < \tilde{f}(x_i)-\tilde{t} \text{ or } y_i > \tilde{f}(x_i)+\tilde{t}\right\}, \\ \tilde{\mathcal{S}}_0 &= \left\{i : y_i = \tilde{f}(x_i)-\tilde{t} \text{ or } y_i = \tilde{f}(x_i)+\tilde{t}\right\}, \\ \tilde{\mathcal{S}}_{\text{in}} &= \left\{i : \tilde{f}(x_i)-\tilde{t} < y_i < \tilde{f}(x_i)+\tilde{t}\right\}.\end{aligned}$$

Let us consider a perturbation on the data in $\tilde{\mathcal{S}}_{\text{out}}$ or $\tilde{\mathcal{S}}_{\text{in}}$. As long as the perturbation is small enough, the perturbed point is also in $\tilde{\mathcal{S}}_{\text{out}}$ or $\tilde{\mathcal{S}}_{\text{in}}$ and $\left[\tilde{f}-\tilde{t}, \tilde{f}+\tilde{t}\right]$ is still the optimal tube. That means a small perturbation on $(x_i, y_i), i \in \tilde{\mathcal{S}}_{\text{out}} \bigcup \tilde{\mathcal{S}}_{\text{in}}$ will not affect the tube regression result. This property is related to the robustness of tube regression, since the data with big noise are usually in $\tilde{\mathcal{S}}_{\text{out}}$ and have no effect on tube regression. Meanwhile, this property also implies the sparsity of tube regression: the derivatives of the optimal tube to data falling outside and within the tube are zero, hence in the ideal situation, the optimal dual variables corresponding

to $(x_i, y_i), i \in \tilde{\mathcal{S}}_{\text{out}} \bigcup \tilde{\mathcal{S}}_{\text{in}}$ are zero. However, the convexity condition makes the loss for $(x_i, y_i), i \in \tilde{\mathcal{S}}_{\text{out}}$ non-constant, which follows that in the proposed convex tube regression, only the data in $\tilde{\mathcal{S}}_{\text{in}}$ are related to zero dual variables.

Next we illustrate the performance of nonparametric convex tube regression by a simple nonlinear example shown in Fig. 4. The underlying function is displayed by a blue solid line. The observed data (blue crosses) are corrupted by noise and contain outliers. The existence of the outliers makes the result of a least squares method using RBF kernel

$$\mathcal{K}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$$

with $\sigma = 0.5$, illustrated by a green dash-dotted line, deviate significantly from the underlying function in some region. Then nonparametric CTR (18) is applied to address this problem. We want the tube to cover 80% data, i.e., $\tau_+ + \tau_- = 0.2$, which also means that almost 80% dual variables are zero. When the sum of τ_+ and τ_- is given, different ratios between them are related to different locations of the tube. This flexibility is suitable to deal with asymmetric noise distribution, for which the probability of the noise larger than and smaller than the mean value are different. By the least squares method, an approximation for the mean value is calculated. We count the numbers of positive and negative residuals. Then, the ratio of them reflects the asymmetry of the noise distribution and can be used to determine $\frac{\tau_+}{\tau_-}$. In this example, there are 42 data points above the least square regressor (green dash-dotted line) and there are 58 data below it. This observation implies that there should be less points above the tube than below it. Therefore, we choose $\frac{\tau_+}{\tau_-} = \frac{42}{58}$ and solve nonparametric CTR (18) with $\gamma = 100$ and RBF kernel ($\sigma = 0.5$). The resulted tube is illustrated by its middle (red solid line) and boundaries (red dotted lines). There are 8 points above the obtained nonparametric tube and 12 below it, illustrating the necessity of using $\tau_+ \neq \tau_-$ for asymmetric noise distribution.

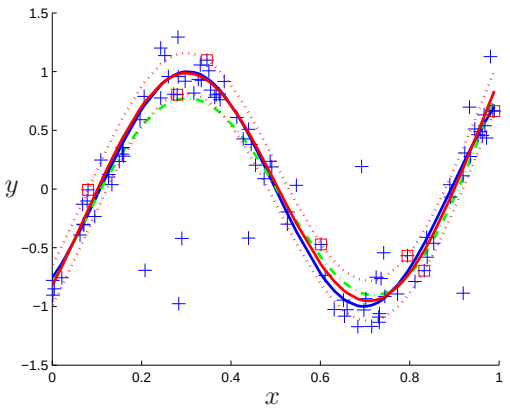


Fig. 4. An example of a nonparametric convex tube regression. The data containing outliers are shown as blue crosses. The existence of outliers makes the result of the least squares method (green dash-dotted line) deviate from the underlying function (blue solid line). The tube obtained by (18) is illustrated by its middle (red solid line) and boundaries (red dotted lines). The points marked by red squares are located on top of the boundaries. Between the boundaries, the points result in zero dual variables.

We summarize the above discussion in the following convex tube regression algorithm, where τ_+, τ_- are determined with the help of least squares support vector machine (LS-SVM, [21] [22]). Since the least square method is easy to be corrupted by noise, a trivial choice $\tau_+ = \tau_-$ is considered as well. In CTR, users also need to provide the coefficient of the regularization term and kernel parameters, which are obtained by 10-fold cross validation based on LS-SVM (implemented by LS-SVMLab v1.8, [23]).

Algorithm 1: Convex Tube Regression Algorithm

- Input $\{(x_i, y_i)\}_{i=1}^n$ (the training data), ρ (the required percentage of the covered training data), and choose the positive definite kernel;
- Tune regularization constant γ and kernel parameters by 10-fold cross validation based on LS-SVM;
- Use the tuned parameters in LS-SVM and denote the approximation value as $\hat{y}_i$.
- Set τ_+, τ_- such that $\frac{\tau_+}{\tau_-} = \frac{\sum_i I(y_i > \hat{y}_i)}{\sum_i I(y_i < \hat{y}_i)}$, $\tau_+ + \tau_- = 1 - \rho$. Solve nonparametric CTR (18) and denote the result as $\lambda^{(1)}, \lambda^{*(1)}, b^{(1)}$ and $t^{(1)}$;
- Set $\tau_+ = \tau_- = \frac{1-\rho}{2}$ and solve nonparametric CTR (18), of which the results are $\lambda^{(2)}, \lambda^{*(2)}, b^{(2)}$ and $t^{(2)}$;

if $t^{(1)} \leq t^{(2)}$ **then**
- Set $\lambda = \lambda^{(1)}, \lambda^* = \lambda^{*(1)}, b = b^{(1)}, t = t^{(1)}$;

else
- Set $\lambda = \lambda^{(2)}, \lambda^* = \lambda^{*(2)}, b = b^{(2)}, t = t^{(2)}$;

end
- Algorithm ends and returns the tube $\big[f - t, f + t\big]$, where f is calculated by (19).

IV. Numerical Experiments

In the numerical experiments, three aspects of convex tube regression will be considered:

- The proposed method is a convex approach for tube regression. To show its effectiveness, we will compare CTR with some existing heuristics for (3);
- Tube regression is distribution-free and can deal with outliers. Hence, we will use CTR in one linear problem, indicating its robustness to outliers;
- To evaluate the nonparametric tube regression, we will compare it with two other nonparametric regression methods, ν-SVM and LS-SVM, on data sets with outliers and asymmetric noise distribution.

A. Optimization performance

In this paper, convex tube regression is proposed as a convex approach for tube regression. As mentioned in the Introduction, there have been some algorithms for the non-convex problem (3), and PROGRESS ([4] [15]) is the most popular. SVT proposed in [1] can be used to solve (3) as well. In the following, we apply them to real data sets downloaded from http://www.agoras.ua.ac.be/ and search a tube containing at least k data samples for each data set. All the experiments are done in Matlab R2011a in Core 2-2.83 GHz, 2.96G RAM.

In this experiment, we consider the performance of CTR for solving the non-convex LMS regression (3). The existing algorithms for LMS regression focus on linear model. Hence we use linear CTR (6) or (18) with a linear kernel. Meanwhile, we do not care about generalization error in this experiment. Thus, when run Algorithm 1, we skip the parameter tuning phase, choose a linear kernel and a large enough γ. We compare the result of CTR with SVT and PROGRESS. For SVT, there is one parameter for the upper bound of the number of points falling outside the tube. Accordingly, we set it to be k. PROGRESS is based on random sampling. To make a fair comparison, we run PROGRESS repetitively and let PROGRESS and Algorithm 1 have the same computation time. The widths of the tube containing the required number of the data are compared in Table I, where the best one is written in bold. For each data set, the data name, the dimension d, the number of data n, and the result of the descent algorithm using linear programming (DALP), which is a global algorithm recently proposed in [18], are given as well for reference.

TABLE I
WIDTH OF THE TUBE CONTAINING k DATA

data set	d	n	k	PROG.	SVT	CTR	DALP
AIRCRAFT	4	23	12	3.164	4.156	**2.587**	1.043
			16	5.450	5.544	**5.179**	2.904
			20	**7.400**	11.03	8.854	5.289
AIRMAY	3	31	16	11.45	14.07	**11.13**	7.231
			20	14.91	111.7	**11.92**	11.92
			24	21.99	306.8	**20.96**	14.66
COLEMAN	6	20	11	**0.283**	0.699	0.561	0.120
			15	1.029	0.927	**0.726**	0.509
			18	2.003	3.130	**1.928**	0.906
DELIVERY	2	25	13	1.005	0.920	**0.904**	0.758
			17	**1.911**	3.165	2.057	1.350
			21	**2.984**	3.825	4.016	2.984
EDUCAT	3	50	26	20.70	27.04	**18.56**	16.11
			36	37.28	**34.85**	35.68	32.28
			43	60.39	**58.65**	61.27	44.06
ENKNOCK	4	16	9	0.184	0.669	**0.110**	0.091
			11	0.932	0.981	**0.663**	0.228
			14	2.514	1.881	**1.297**	1.297
EXACT	3	25	13	**0.500**	0.700	**0.500**	0.250
			18	0.889	1.039	**0.500**	0.500
			23	2.000	**1.375**	**1.375**	1.300
HAWKINS	3	75	38	0.597	**0.577**	**0.577**	0.419
			51	**0.736**	0.851	0.799	0.716
			66	**1.337**	2.398	1.967	1.047
HEART	2	12	7	**0.766**	1.646	1.237	0.542
			9	2.565	2.680	**2.149**	1.974
			11	5.195	5.738	**4.715**	3.591
PHOSRHR	2	18	10	6.573	4.752	**4.084**	4.084
			13	11.27	13.38	**11.21**	9.678
			16	18.44	27.37	**16.09**	16.09
SALINITY	4	28	15	0.509	0.668	**0.468**	0.258
			20	1.028	1.030	**0.824**	0.808
			25	1.813	1.796	**1.711**	1.354
STACKLOSS	4	21	11	**1.000**	1.840	1.079	0.500
			15	2.244	2.555	**2.208**	1.816
			19	4.500	4.362	**4.333**	2.844
WOOD	6	20	11	**0.004**	0.014	0.009	0.002
			15	**0.012**	0.018	0.016	0.006
			18	0.035	0.030	**0.027**	0.021

From Table I, one can see that the proposed convex approximation method generally gives good results for tube regression. Compared to SVT, the advantage of CTR is that we can consider various locations of the tube, which makes CTR perform better. For some data sets, PROGRESS gives good results. However, the nonparametric formulation based on PROGRESS is hard to construct.

B. Robustness

The motivation of tube regression is to deal with outliers. Though CTR is an approximation to (2), it still enjoys some robustness. To evaluate robustness of CTR, a linear function $f(x) = w^T x + b$ with $w = [1, 0.5, -0.5, -1, 2]^T$ and $b = -3$ is considered. We generate $N = 100$ data following a uniform distribution in $[0, 1]^5$, i.e., $\{x \in \mathbb{R}^5, 0 \le x(i) \le 1, 1 \le i \le 5\}$. The least squares method is the best for Gaussian noise. In this experiment, a chi-square distribution noise is considered, that means $y_i = f(x_i) + \delta_{\chi^2}$, where δ_{χ^2} follows a chi-square distribution with zero mean and 4 degrees of freedom. Then we let the ratio of the variance of noise δ_{χ^2} to that of $f(x)$, denoted as r_{noise}, equals to 0.05 and 0.1. Fig. 5 illustrates the probability density function of a chi-square distribution with zero mean and variance 8.00. From this figure, one can see the asymmetry of the noise distribution: the mean of δ_{χ^2} is zero and Prob$\{\delta_{\chi^2} < 0\} = 59.4\%$, Prob$\{\delta_{\chi^2} < 0\} = 40.6\%$. Besides the noise, some outliers are added into the training set. We randomly select r_o percentage of samplings and replace their observed values by random values following a uniform distribution in the range of $f(x), x \in [0, 1]^5$.

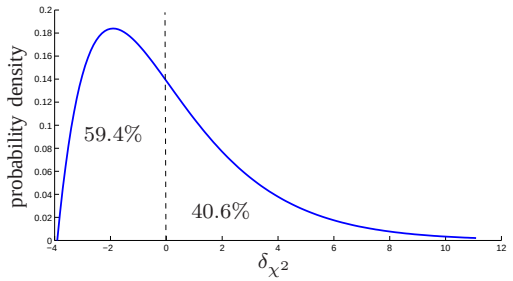


Fig. 5. The probability density function of δ_{χ^2}. The mean of δ_{χ^2} is zero and the variance of δ_{χ^2} is 8.00. This distribution is asymmetric and the probability of δ_{χ^2} being positive is 40.6%.

For the above data, we pursue a linear tube covering 80% data by solving (6), i.e., Algorithm 1 with a linear kernel and $\rho = 0.8$ is applied. To handle outliers, we also consider robust regression using a Huber loss, which is implemented by the statistics toolbox of Matlab. For each group of r_{noise} and r_o, the above experiments are repeated 10 times. The average and the standard deviation of the results are shown in Table II.

The least squares method is suitable to deal with Gaussian noise. When the noise follows another distribution or there exist outliers, CTR performs better than the least squares method, as shown in Table II. Robust regression using a Huber loss is also a good method for handling outliers and it has similar regression performance as CTR. Compared with a Huber loss method, the tube regression gives not only the predictive value, but also a width covering a required percentage of the data. Unlike prediction interval analysis, the obtained tube is distribution-free. Its potential applications can be found in the field of l_∞ regression or minimax approximation, an extreme case of tube regression with $\rho = 1$. l_∞ regression pursues a tube, which has a small width and covers all the training data. It originates from the worst-case analysis and has been applied widely in e.g., circuit design [24] [25], signal processing [26]

TABLE II
THE AVERAGE AND STANDARD DEVIATION OF THE REGRESSION RESULTS

	Real value	1	0.5	−0.5	−1	2	−3
$r_{\text{noise}} = 0.05$	Least Squares	0.947 ± 0.125	0.487 ± 0.060	-0.434 ± 0.107	-0.928 ± 0.136	1.835 ± 0.075	-2.942 ± 0.066
$r_o = 5\%$	Huber Loss	0.977 ± 0.085	0.519 ± 0.048	-0.500 ± 0.051	-0.952 ± 0.060	1.943 ± 0.050	-3.009 ± 0.062
	CTR	0.992 ± 0.036	0.507 ± 0.022	-0.502 ± 0.029	-1.001 ± 0.055	1.994 ± 0.042	-2.997 ± 0.055
$r_{\text{noise}} = 0.05$	Least Squares	0.877 ± 0.103	0.477 ± 0.146	-0.514 ± 0.147	-0.904 ± 0.143	1.868 ± 0.120	-2.880 ± 0.163
$r_o = 10\%$	Huber Loss	0.972 ± 0.039	0.508 ± 0.064	-0.506 ± 0.067	-0.974 ± 0.060	1.987 ± 0.087	-3.024 ± 0.073
	CTR	1.000 ± 0.035	0.500 ± 0.031	-0.518 ± 0.028	-0.998 ± 0.035	2.013 ± 0.028	-2.987 ± 0.040
$r_{\text{noise}} = 0.10$	Least Squares	1.027 ± 0.111	0.514 ± 0.121	-0.489 ± 0.116	-1.014 ± 0.111	1.926 ± 0.076	-2.983 ± 0.077
$r_o = 5\%$	Huber Loss	0.989 ± 0.075	0.525 ± 0.047	-0.515 ± 0.057	-0.971 ± 0.068	2.002 ± 0.041	-3.026 ± 0.049
	CTR	1.007 ± 0.054	0.497 ± 0.103	-0.488 ± 0.083	-0.992 ± 0.085	1.997 ± 0.056	-3.024 ± 0.081
$r_{\text{noise}} = 0.10$	Least Squares	0.874 ± 0.196	0.491 ± 0.158	-0.426 ± 0.136	-0.960 ± 0.094	1.798 ± 0.103	-2.867 ± 0.118
$r_o = 10\%$	Huber Loss	0.936 ± 0.130	0.507 ± 0.087	-0.450 ± 0.035	-1.041 ± 0.070	1.952 ± 0.066	-2.908 ± 0.081
	CTR	0.992 ± 0.052	0.499 ± 0.070	-0.492 ± 0.061	-0.980 ± 0.060	1.997 ± 0.082	-2.996 ± 0.101

[27], portfolio optimization [28] [29], and so on. Its result is sensitive to outliers. In contrast, tube regression is more robust. Therefore, in the field involving l_∞ regression, convex tube regression is a potential extension for data containing outliers.

In tube regression, we want the tube $[f - t, f + t]$ cover ρ percentage of the training data. Then if the new data $(\tilde{x}, \tilde{y})$ comes from the same distribution of the training set, the probability of the data falling into the tube, i.e., Prob$\{\tilde{y} \in [f(\tilde{x}) - t, f(\tilde{x}) + t]\}$, is related to ρ. To evaluate this point, we use the same statistics (noise distribution, fraction of outliers, values of outliers) on the testing data. Then we count the number of data falling into the tube. The average of the tube width and the percentage of the testing data covered by the tube are listed in Table III. In this experiment, we pursue a tube covering 80% training data. It can be expected that by increasing the training data size, the percentage of the testing data covered by the obtained tube will tend to 80%. Here we indicate this using $N = 200$ training data.

TABLE III
THE AVERAGE TUBE WIDTH AND THE PERCENTAGE OF THE TESTING DATA COVERED BY THE TUBE

		$N = 100$		$N = 200$	
r_{noise}	r_o	Width	Percentage	Width	Percentage
0.05	0.05	0.1694	73.9%	0.1824	75.0%
0.05	0.10	0.1671	72.5%	0.1831	77.7%
0.10	0.05	0.2605	76.3%	0.2676	77.6%
0.10	0.10	0.2495	71.6%	0.2671	75.5%

C. Nonlinear regression with outliers

One interesting aspect of the proposed CTR is that a nonparametric formulation can be developed. The output of (18) includes the middle and the width of the tube, both of which are useful in applications. In this experiment, we first evaluate the regression performance and then report the tube width and the percentage of testing data falling into the nonparametric tube. We consider testing functions provided in [30]. These functions are listed below and have been used in many papers to examine the performance of different regression methods, see, e.g., [31] [32] [33] and [34].

Fun. 1: $f_1(x) = \sin(x(1)x(2))$, $D = [-2, 2]^2$.

Fun. 2: $f_2(x) = \exp\big(x(1)\sin(\pi x(2))\big)$, $D = [-1, 1]^2$.

Fun. 3: $f_3(x) = \dfrac{40 f_3^1(x)}{f_3^2(x) + f_3^3(x)}$, $D = [0, 1]^2$, where

$$\begin{aligned} f_3^1(x) &= \exp\Big(8\big((x(1) - 0.5)^2 + (x(2) - 0.5)^2\big)\Big) \\ f_3^2(x) &= \exp\Big(8\big((x(1) - 0.2)^2 + (x(2) - 0.7)^2\big)\Big) \\ f_3^3(x) &= \exp\Big(8\big((x(1) - 0.7)^2 + (x(2) - 0.2)^2\big)\Big). \end{aligned}$$

Fun. 4: $f_4(x) = 42.659\big(0.1 + (x(1) - 0.5) f_4^1(x)\big)$, $D = [-0.5, 0.5]^2$, where

$$\begin{aligned} f_4^1(x) &= 0.05 + (x(1) - 0.5)^4 - 10(x(1) - 0.5)^2 \\ &\quad (x(2) - 0.5)^2 + 5(x(2) - 0.5)^4. \end{aligned}$$

Fun. 5: $f_5(x) = 1.3356\big(f_5^1(x(1)) + f_5^2(x(2))\big)$, $D = [0, 1]^2$, where

$$\begin{aligned} f_5^1(x(1)) &= 1.5(1 - x(1)) + \exp(2x(1) - 1) \\ &\quad \sin(3\pi(x(1) - 0.6)^2) \\ f_5^2(x(2)) &= \exp(3(x(2) - 0.5))\sin(4\pi(x(2) - 0.9)^2). \end{aligned}$$

Fun. 6: $f_6(x) = 10\sin(\pi x(1)x(2)) + 20(x(3) - 0.5)^2 + 5x(4) + 10x(5) + 0x(6)$, $D = [-1, 1]^6$.

Fun. 7: $f_7(x) = \exp\big(2x(1)\sin(\pi x(4))\big) + \sin(x(2)x(3))$, $D = [-0.25, 0.25]^4$.

For the two-dimensional functions above, we generate 400 deterministic data samples $x_i \in \mathbb{R}^2$, $1 \le i \le 400$, evenly spaced along its domain axes. Then we add noise and outliers following the same way as used in subsection IV-B. For the high-dimensional functions, the same procedure is conducted except that 400 data are randomly selected within the domain of interest and taken from a uniform probability distribution.

For each group of data, we use nonparametric CTR (18), ν-SVM, and LS-SVM to obtain the approximation $\hat{f}$. The RBF kernel is used in the three methods. Since the computation time of LS-SVM is significantly smaller than that of the other two methods, we use 10-fold cross validation based on LS-SVM to find the suitable kernel parameter and the coefficient

TABLE IV
RELATIVE SUM OF THE SQUARE ERRORS ON TESTING DATA

	r_{noise}	r_o	LS-SVM	ν-SVM	CTR
f_1	0.05	0.05	0.0156 ± 0.0051	0.0143 ± 0.0049	0.0115 ± 0.0050
	0.05	0.10	0.0263 ± 0.0067	0.0226 ± 0.0064	0.0177 ± 0.0081
	0.10	0.05	0.0180 ± 0.0074	0.0222 ± 0.0048	0.0190 ± 0.0072
	0.10	0.10	0.0292 ± 0.0119	0.0277 ± 0.0074	0.0239 ± 0.0078
f_2	0.05	0.05	0.0231 ± 0.0083	0.0156 ± 0.0038	0.0111 ± 0.0056
	0.05	0.10	0.0484 ± 0.0100	0.0302 ± 0.0089	0.0161 ± 0.0062
	0.10	0.05	0.0259 ± 0.0098	0.0252 ± 0.0044	0.0157 ± 0.0079
	0.10	0.10	0.0585 ± 0.0246	0.0410 ± 0.0187	0.0300 ± 0.0160
f_3	0.05	0.05	0.0278 ± 0.0127	0.0152 ± 0.0084	0.0062 ± 0.0012
	0.05	0.10	0.0495 ± 0.0160	0.0302 ± 0.0222	0.0110 ± 0.0038
	0.10	0.05	0.0281 ± 0.0056	0.0217 ± 0.0105	0.0090 ± 0.0028
	0.10	0.10	0.0511 ± 0.0108	0.0355 ± 0.0128	0.0197 ± 0.0118
f_4	0.05	0.05	0.0243 ± 0.0149	0.0136 ± 0.0078	0.0030 ± 0.0015
	0.05	0.10	0.0632 ± 0.0212	0.0278 ± 0.0151	0.0105 ± 0.0076
	0.10	0.05	0.0282 ± 0.0142	0.0192 ± 0.0121	0.0069 ± 0.0056
	0.10	0.10	0.0647 ± 0.0226	0.0330 ± 0.0095	0.0189 ± 0.0093
f_5	0.05	0.05	0.0241 ± 0.0026	0.0191 ± 0.0048	0.0147 ± 0.0070
	0.05	0.10	0.0442 ± 0.0183	0.0317 ± 0.0127	0.0161 ± 0.0051
	0.10	0.05	0.0302 ± 0.0070	0.0248 ± 0.0076	0.0184 ± 0.0093
	0.10	0.10	0.0490 ± 0.0118	0.0346 ± 0.0168	0.0226 ± 0.0078
f_6	0.05	0.05	0.0616 ± 0.0158	0.0512 ± 0.0210	0.0410 ± 0.0089
	0.05	0.10	0.0812 ± 0.0348	0.0869 ± 0.0416	0.0580 ± 0.0277
	0.10	0.05	0.0635 ± 0.0109	0.0770 ± 0.0391	0.0493 ± 0.0111
	0.10	0.10	0.0736 ± 0.0166	0.0819 ± 0.0190	0.0550 ± 0.0176
f_7	0.05	0.05	0.0241 ± 0.0186	0.0140 ± 0.0064	0.0167 ± 0.0083
	0.05	0.10	0.0385 ± 0.0126	0.0258 ± 0.0059	0.0225 ± 0.0197
	0.10	0.05	0.0261 ± 0.0088	0.0262 ± 0.0133	0.0256 ± 0.0159
	0.10	0.10	0.0430 ± 0.0213	0.0329 ± 0.0122	0.0264 ± 0.0040

of the regularization term, then we use them in all the three methods. For ν-SVM, the lower bound on the fraction of the data covered by the tube are set to be 80%. For CTR (18), we also set $\rho = 0.8$ and run Algorithm 1. After obtaining the approximation function $\hat{f}$, 100 testing data uniformly distributed in the domain are generated and the relative sum of the squared errors (RSSE), defined below, are calculated,

$$\text{RSSE} = \frac{\sum_{x \in V} \big(f(x) - \hat{f}(x)\big)^2}{\sum_{x \in V} \big(f(x) - \text{E}(f)\big)^2},$$

where V is the set of testing data, $\hat{f}$ is the approximation of f, and $\text{E}(f)$ is the average value of $f(x)$, $x \in V$. For each case, we repeat the process above 10 times and report the average and the standard deviation of RSSEs in Table IV.

LS-SVM is suitable to deal with Gaussian noise and enjoys low computational complexity. If the data contain outliers, one should consider the reweighted strategy for LS-SVM, see [35]. In this experiment, to show the effect of the outliers, we use the original LS-SVM and hence it loses some accuracy. ν-SVM is robust since its loss function increases linearly when the residual tends to infinity. As discussed previously, ν-SVM can be seen as a special case of CTR with $\tau_+ = \tau_- = 0$. To handle asymmetric noise distribution, various locations of the tube should be considered. As indicated in Table IV, nonparametric CTR (18) gives good results in this experiment.

We are also interested in the probability of a new point falling into the nonparametric tube, if the new data and the training data come from the same distribution. For this, we let the testing data be corrupted by outliers and noise following the same distributions as in the training data. Then Table V reports the average of the tube width and the percentage of the testing data being covered by the obtained tube. To evaluate the width of the tube, we also give the range of the function values for each function for reference.

TABLE V
THE AVERAGE WIDTH OF THE NONPARAMETRIC TUBE AND THE AVERAGE PERCENTAGE OF THE TESTING DATA COVERED BY THE TUBE

Function	r_{noise}	r_o	Width	Percentage
f_1 $[-1.00, 1.00]$	0.05	0.05	0.1470	69.15%
	0.05	0.10	0.1666	69.83%
	0.10	0.05	0.2010	67.30%
	0.10	0.10	0.2298	69.71%
f_2 $[0.39, 2.56]$	0.05	0.05	0.0988	66.33%
	0.05	0.10	0.1158	70.53%
	0.10	0.05	0.1509	70.53%
	0.10	0.10	0.1669	70.87%
f_3 $[0.30, 15.1]$	0.05	0.05	0.7552	73.83%
	0.05	0.10	0.7350	70.82%
	0.10	0.05	1.0007	69.00%
	0.10	0.10	1.2642	77.18%
f_4 $[-48.0, 152]$	0.05	0.05	7.8654	79.44%
	0.05	0.10	9.1289	81.38%
	0.10	0.05	11.231	78.87%
	0.10	0.10	13.171	80.11%
f_5 $[0.08, 6.25]$	0.05	0.05	0.2832	65.20%
	0.05	0.10	0.3560	72.15%
	0.10	0.05	0.3102	68.95%
	0.10	0.10	0.4817	70.54%
f_6 $[-21.0, 61.0]$	0.05	0.05	4.3202	84.45%
	0.05	0.10	4.8514	68.54%
	0.10	0.05	5.6551	67.25%
	0.10	0.10	7.2499	76.90%
f_7 $[0.70, 1.40]$	0.05	0.05	0.0357	77.10%
	0.05	0.10	0.0393	70.85%
	0.10	0.05	0.0442	70.30%
	0.10	0.10	0.0586	68.80%

V. CONCLUSION

Tube regression tries to find a tube with a small width which covers a certain percentage of the training data. Tube regression is a robust and distribution-free method but the related optimization problem is non-convex. In this paper, a convex formulation is proposed using the methods of quantile regression. Based on convex tube regression, nonparametric convex tube regression has been given and then nonlinear kernels are applied. Numerical experiments indicate that CTR is an efficient convex approach for tube regression. The experiments on nonlinear regression also illustrate that nonparametric CTR is a potential nonlinear data processing method for outliers and asymmetric noise distributions.

REFERENCES

[1] K. Pelckmans, J. De Brabanter, J.A.K. Suykens, and B. De Moor, "Least conservative support and tolerance tubes," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3799–3806, 2009.

[2] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge university press, 2004.

[3] P. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[4] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. Wiley-IEEE, 1987.

[5] D. Donoho and P. Huber, "The notion of breakdown point," *A Festschrift for Erich L. Lehmann in Honor of His Sixty-fifth Birthday*, pp. 157–184, 1982.

[6] M. Powell, *Approximation Theory and Methods: Chapter 7, The theory of minimax approximation*. Cambridge University Press, 1981.

[7] R. Koenker, *Quantile Regression*. Cambridge Univ Press, 2005.

[8] I. Steinwart and A. Christmann, "How SVMs can estimate quantiles and the median," *Advances in Neural Information Processing Systems*, vol. 20, pp. 305–312, 2008.

[9] ——, "Estimating conditional quantiles with the help of the pinball loss," *Bernoulli*, vol. 17, no. 1, pp. 211–225, 2011.

[10] P. Tichavsky, "Algorithms for and geometrical characterization of solutions in the LMS and the LTS linear regression," *Computational Statistics Quarterly*, vol. 8, pp. 139–151, 1991.

[11] P. Boček and P. Lachout, "Linear programming approach to LMS-estimation," *Computational Statistics & Data Analysis*, vol. 19, no. 2, pp. 129–134, 1995.

[12] C. Olson, "An approximation algorithm for least median of squares regression," *Information Processing Letters*, vol. 63, no. 5, pp. 237–241, 1997.

[13] V. Verardi and C. Croux, "Robust regression in stata," *The Stata Journal*, vol. 9, no. 3, pp. 439–453, 2009.

[14] P. Winker, M. Lyra, and C. Sharpe, "Least median of squares estimation by optimization heuristics with an application to the capm and a multi-factor model," *Computational Management Science*, vol. 8, no. 1, pp. 103–123, 2011.

[15] P. Rousseeuw and M. Hubert, "Recent developments in PROGRESS," *Lecture Notes-Monograph Series*, vol. 31, pp. 201–214, 1997.

[16] J. Steele and W. Steiger, "Algorithms and complexity for least median of squares regression," *Discrete Applied Mathematics*, vol. 14, no. 1, pp. 93–100, 1986.

[17] A. Stromberg, "Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression," *SIAM Journal on Scientific Computing*, vol. 14, pp. 1289–1299, 1993.

[18] X. Huang, J. Xu, S. Wang, and C. Xu, "Minimization of the k-th maximum and its application on LMS regression and VaR optimization," *Journal of the Operational Research Society*, vol. 63, no. 11, pp. 1479–1491, 2012.

[19] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[20] V. Vapnik, *The Nature of Statistical Learning*, Springer-Verlag, 1995.

[21] J.A.K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[22] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[23] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J.A.K. Suykens, "LS-SVMlab Toolbox User's Guide version 1.8," Internal Report 10-146, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2010.

[24] K. Antreich, H. Graeb, and C. Wieser, "Circuit analysis and optimization driven by worst-case distances," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 13, no. 1, pp. 57–71, 1994.

[25] N. Papamarkos and C. Chamzas, "A new approach for the design of digital integrators," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 43, no. 9, pp. 785–791, 1996.

[26] I. Kollar, R. Pintelon, and J. Schoukens, "Optimal FIR and IIR hilbert transformer design via LS and minimax fitting," *IEEE Transactions on Instrumentation and Measurement*, vol. 39, no. 6, pp. 847–852, 1990.

[27] T. Dvorkind, H. Kirshner, Y. Eldar, and M. Porat, "Minimax approximation of representation coefficients from generalized samples," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4430–4443, 2007.

[28] M. Young, "A minimax portfolio selection rule with linear programming solution," *Management Science*, vol. 44, no. 5, pp. 673–683, 1998.

[29] X. Cai, K. Teo, X. Yang, and X. Zhou, "Portfolio optimization under a minimax rule," *Management Science*, vol. 46, no. 7, pp. 957–972, 2000.

[30] V. Cherkassky, D. Gehring, and F. Mulier, "Comparison of adaptive methods for function estimation from samples," *IEEE Transactions on Neural Networks*, vol. 7, no. 4, pp. 969–984, 1996.

[31] D. Hush and B. Horne, "Efficient algorithms for function approximation with piecewise linear sigmoidal networks," *IEEE Transactions on Neural Networks*, vol. 9, no. 6, pp. 1129–1141, 1998.

[32] J. González, I. Rojas, J. Ortega, H. Pomares, F. Fernández, and A. Díaz, "Multiobjective evolutionary optimization of the size, shape, and position parameters of radial basis function networks for function approximation," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1478–1495, 2003.

[33] A. Martínez-Estudillo, F. Martínez-Estudillo, C. Hervás-Martínez, and N. García-Pedrajas, "Evolutionary product unit based neural networks for regression," *Neural Networks*, vol. 19, no. 4, pp. 477–486, 2006.

[34] S. Wang, X. Huang, and Y. Yam, "A neural network of smooth hinge functions," *IEEE Transactions on Neural Networks*, vol. 21, no. 9, pp. 1381–1395, 2010.

[35] J.A.K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: robustness and sparse approximation," *Neurocomputing*, vol. 48, no. 1, pp. 85–105, 2002.