Support Vector Regression: Risk Quadrangle Framework

Anton Malandii* Stan Uryasev*

January 4, 2023

Abstract

This paper investigates Support Vector Regression (SVR) in the context of the fundamental risk quadrangle paradigm. It is shown that both formulations of SVR, ε-SVR and ν-SVR, correspond to the minimization of equivalent regular error measures (Vapnik error and superquantile (CVaR) norm, respectively) with a regularization penalty. These error measures, in turn, give rise to corresponding risk quadrangles. Additionally, the technique used for the construction of quadrangles serves as a powerful tool in proving the equivalence between ε-SVR and ν-SVR.

In general, the Support Vector (SV) algorithm, which is firmly grounded in the framework of Vapnik-Chervonenkis (VC) theory, is proven to be a highly effective classification tool and a generalization of regression. In the context of VC theory, a problem of function estimation relies on the so-called principle of empirical risk minimization, which is essentially a minimization of the expected value of the loss function of the regression residual.

The fundamental risk quadrangle theory takes another approach. The concept of generalized regression in this paradigm is developed through the regular measure of error, which is defined axiomatically and, therefore, takes it beyond the expected loss types of error. However, it turns out that errors of expectation type play a significant role in statistical estimation. In particular, regression with respect to a regular error measure of expectation type estimates the corresponding conditional statistic from the same quadrangle (e.g., least-squares regression estimates conditional mean, quantile regression estimates conditional quantile).

By constructing the fundamental risk quadrangle, which corresponds to SVR, we show that SVR is the asymptotically unbiased estimator of the average of two symmetric conditional quantiles. Additionally, SVR is formulated as a regular deviation minimization problem with a regularization penalty by invoking Error Shaping Decomposition of Regression. Finally, the dual formulation of SVR in the risk quadrangle framework is derived.

Keywords: Support Vector Regression, Vapnik-Chervonenkis Theory, Fundamental Risk Quadrangle, Superquantile, Superquantile Norm, Stochastic Optimization, Estimation.

*Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794, USA. Email: anton.malandii@stonybrook.edu, stanislav.uryasev@stonybrook.edu

1 Introduction

Approximating a complex or partially observed random variable Y by a function of an observed random vector $\boldsymbol{X} = (X_1, \dots, X_n)^T$ is a classical problem in stochastic modeling, which is called *regression*. To solve this problem, it is required to find a such function $\hat{f}(\boldsymbol{X})$ from a given class of functions $\mathcal{F}$ that minimizes the difference (regression residual) $Z_f = Y - \hat{f}(\boldsymbol{X})$ with respect to the choice of a distance measure between Y and $\hat{f}$. Usually, a norm serves as a distance measure (e.g., $\mathcal{L}^1$-regression, $\mathcal{L}^2$-regression), however, more generally, cf. Rockafellar et al. [2008], axiomatically defined error measures, frequently turn out to be preferable. But the selection of a distance measure is not the only concern when dealing with regression. It is also of significant importance to choose an appropriate class of approximation functions $\mathcal{F}$. It is obvious that a chosen class of functions should have necessary approximation properties. Typically universal approximators such as polynomials, splines, wavelets, or neural networks serve this purpose.

From the statistical perspective, the task of regression is to *estimate* a conditional *statistic* $\mathcal{S}(Y)$ of a random variable Y given $\boldsymbol{X}$ by finding a function $\hat{f}$, which is called the *best estimator*. For example, $\mathcal{L}^1$-regression estimates the conditional median[1], i.e.,

$$\hat{f}_{\mathcal{L}^1}(\boldsymbol{X}) \in \operatorname{med}[Y|\boldsymbol{X}],$$

and $\mathcal{L}^2$-regression estimates the conditional mean

$$\hat{f}_{\mathcal{L}^2}(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}].$$

There exist two general frameworks in which regression can be carried out: VC theory, popular in the machine learning community, cf. Vapnik [1995] and the risk quadrangle, well-known in the field of stochastic optimization, cf. Rockafellar and Uryasev [2013]. The former framework considers error measures only of the expectation type, i.e., $\mathbb{E}[\mathcal{L}(Z_f)]$, where $\mathcal{L}$ is a so-called loss function hence the choice of error here boils down to the choice of a loss function. Additionally, this framework is a powerful tool for analyzing the class of approximation functions $\mathcal{F}$, in particular its VC-dimension. Conversely, the latter framework considers axiomatically defined error measures $\mathcal{E}(Z_f)$ that are not necessarily of expectation type. Moreover, it links a selected error with other uncertainty measures, such as risk, deviation, and regret. However, it turns out that errors of expectation type play a significant role when it comes to estimation. Indeed, the Regression Theorem in the risk quadrangle theory states that in this case, the best estimator is a conditional statistic from the corresponding quadrangle, i.e.,

$$\hat{f}_{\mathcal{E}}(\boldsymbol{X}) \in \mathcal{S}(Y|\boldsymbol{X}).$$

Hence the choice of a loss function corresponds to a particular statistic.

In the context of machine learning, regression is understood as a procedure for the optimal fitting of a given dataset $(\boldsymbol{x}_i, Y_i)_{i=1}^l$, $\boldsymbol{x}_i \in \mathbb{R}^n$, $Y_i \in \mathbb{R}$, which is called a *training sample*. The goal here is constructing a prediction model $\hat{f}(\boldsymbol{x})$, which gives a reasonable forecast $\hat{f}(\boldsymbol{x}_i)$ for the future outcome Y_{l+1}. Combining the risk quadrangle and VC theory, we obtain

$$\hat{f}_{\mathcal{E}}(\boldsymbol{x}) \in \mathcal{S}(Y|\boldsymbol{X} = \boldsymbol{x}),$$

[1] We use "$\in$" here to emphasize that the optimal solution may be not unique.

i.e., the optimal prediction model is a conditional statistic.

This paper investigates SVR in the fundamental risk quadrangle framework. We show that both formulations of SVR, ε-SVR and ν-SVR, correspond to the minimization of equivalent regular error measures with a regularization penalty. These error measures, in turn, give rise to corresponding risk quadrangles. The equivalence is established through the dual superquantile optimization formula, cf. Guan, Giao, and Wang [2022], which additionally reveals a duality between ε and ν. Furthermore, we show that the error measure in the ν-SVR is actually a superquantile (CVaR) norm. Hence based on the equivalence result, we treat SVR as regularized superquantile norm minimization, which admits an elegant dual formulation that is equivalent to the classical one, cf. Schölkopf et al. [2000]. Finally, having the risk quadrangle that corresponds to SVR constructed, we obtain that SVR is the asymptotically unbiased estimator of the average of two symmetric conditional quantiles, i.e.,

$$\hat{f}(\boldsymbol{X}) \in \frac{1}{2}\left(q_{(1-\alpha)/2}\left(Y \,\middle|\, \boldsymbol{X}\right) + q_{(1+\alpha)/2}\left(Y \,\middle|\, \boldsymbol{X}\right)\right).$$

2 SVR Formulations

This section, provides a comprehensive review of two existing SVR formulations, i.e., ε-SVR and ν-SVR. In a nutshell, the optimization problem statements for both SVRs are introduced in two equivalent formats, and the equivalence of the solution sets of the aforementioned SVR formulations is reviewed, along with ε and ν parameters interpretation.

2.1 The ε-SVR

Consider the linear regression problem with a training sample

$$X^l = (\boldsymbol{x}_i, y_i)_{i=1}^l, \tag{2.1}$$

where each $\boldsymbol{x}_i \in \mathbb{R}^n$ is the *feature* vector and each $y_i \in \mathbb{R}$ is the *target output* and one needs to find a hyperplane

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b, \quad (\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$$

that optimally fits the given training data.

This problem can be efficiently solved via ε-SVR introduced by Vapnik et al. [1996]. For $C > 0$, $\varepsilon > 0$, and a general norm[2] $\|\cdot\|$, the standard form of ε-SVR is as follows

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ \text{s.t.} \quad & \boldsymbol{w}^T\boldsymbol{x}_i + b - y_i \le \varepsilon + \xi_i \\ & y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0, \quad i = 1, \ldots, l. \end{aligned} \tag{2.2}$$

The optimal fit in ε-SVR means that the method searches for $f_{\boldsymbol{w},b}(\boldsymbol{x})$ that has at most ε deviation from the actually obtained targets y_i for all the training data. The constant $C > 0$

[2]Typically ℓ^2, $\|\cdot\|_2$ or ℓ^1, $\|\cdot\|_1$ norms are chosen.

determines the trade-off between the flatness of $f_{\boldsymbol{w},b}(\boldsymbol{x})$ and the amount up to which deviations larger than ε are tolerated.

Problem (2.2) can be equivalently reformulated in terms of ε-insensitive loss function

$$\mathcal{L}_\varepsilon(\xi) = [|\xi| - \varepsilon]_+, \quad \xi \in \mathbb{R} \tag{2.3}$$

where $[\cdot]_+ = \max[0, \cdot]$. Indeed, (2.2) is equivalent to

$$\min_{\boldsymbol{w},b} \quad \frac{1}{l}\sum_{i=1}^{l} \mathcal{L}_\varepsilon(y_i - f_{\boldsymbol{w},b}(\boldsymbol{x}_i)) + \frac{1}{2C}\|\boldsymbol{w}\|^2. \tag{2.4}$$

For purposes that will be evident later, we rewrite (2.4) using slightly different notations, usually adopted in machine learning literature:

$$\min_{\boldsymbol{w},b} \quad \frac{1}{l}\mathbf{1}_l^T \boldsymbol{L}_\varepsilon(\boldsymbol{y} - \hat{\mathbf{X}}\boldsymbol{w} - \mathbf{1}_l b) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2, \tag{2.5}$$

where $\mathbf{1}_l = (1, \ldots, 1)^T \in \mathbb{R}^l$, $\boldsymbol{y} = (y_1, \ldots, y_l)^T$, $\hat{\mathbf{X}} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_l^T) \in \mathbb{R}^{l\times n}$, $\lambda = 1/C$, and finally the loss vector $\boldsymbol{L}_\varepsilon(\boldsymbol{y} - \hat{\mathbf{X}}\boldsymbol{w} - \mathbf{1}_l b) = (\mathcal{L}_\varepsilon(y_1 - f_{\boldsymbol{w},b}(\boldsymbol{x}_1)), \ldots, \mathcal{L}_\varepsilon(y_l - f_{\boldsymbol{w},b}(\boldsymbol{x}_l)))^T$.

In the probabilistic framework, we consider that

$$\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{w}, b) = \boldsymbol{y} - \hat{\mathbf{X}}\boldsymbol{w} - \mathbf{1}_l b$$

is a random variable on a finite sample space $\Omega := \{\omega_1, \ldots, \omega_l\}$ (equipped with a sigma-algebra $\mathcal{A}$), $l < \infty$. Hence we can identify a random variable $\boldsymbol{z}$ as a vector in $\mathbb{R}^l$, i.e., $\boldsymbol{z} \in \mathbb{R}^l$. Also, as it is usually in the machine learning context, we define $\boldsymbol{p} = \mathbf{1}_l/l$ as a uniform probability measure on $\mathcal{A}$. Thus, we can define the mathematical expectation with respect to the probability measure $\boldsymbol{p}$ as follows

$$\mathbb{E}_{\boldsymbol{p}} = \boldsymbol{p}^T \boldsymbol{z},$$

and, therefore, equivalently reformulate (2.9) as

$$\min_{\boldsymbol{w},b} \quad \mathbb{E}_{\boldsymbol{p}}[\boldsymbol{L}_\varepsilon(\boldsymbol{z})] + \frac{\lambda}{2}\|\boldsymbol{w}\|^2. \tag{2.6}$$

2.2 The ν-SVR

The ν-SVR was introduced in Schölkopf et al. [2000], where parameter $\nu \in (0, 1]$ is used to control the number of support vectors. For (C, ν) and $\|\cdot\|$, ν-SVR solves

$$\begin{aligned} \min_{\boldsymbol{w},b,\varepsilon,\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad & \|\boldsymbol{w}\|^2 + C\left(\nu\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right) \\ \text{s.t.} \quad & \boldsymbol{w}^T\boldsymbol{x}_i + b - y_i \le \varepsilon + \xi_i \\ & y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0, \quad i = 1, \ldots, l, \quad \varepsilon \ge 0. \end{aligned} \tag{2.7}$$

The equivalence of ε-SVR and ν-SVR was established in Chang and Lin [2001] in the sense that ε-SVR with parameters (C, ε) has the same optimal solution as ν-SVR with parameters (C, ν).

Analogously to (2.4), ν-SVR can be equivalently reformulated as follows

$$\min_{\boldsymbol{w},b,\varepsilon} \quad \frac{1}{l}\sum_{i=1}^{l}\mathcal{L}_{\varepsilon}(y_i - f_{\boldsymbol{w},b}(\boldsymbol{x}_i) + \varepsilon\nu + \frac{1}{2C}\left\|\boldsymbol{w}\right\|^2, \tag{2.8}$$

or

$$\min_{\boldsymbol{w},b,\varepsilon} \quad \frac{1}{l}\mathbf{1}_l^T\boldsymbol{L}_{\varepsilon}(\boldsymbol{y} - \hat{\mathbf{X}}\boldsymbol{w} - \mathbf{1}_l b) + \varepsilon\nu + \frac{\lambda}{2}\left\|\boldsymbol{w}\right\|^2. \tag{2.9}$$

In the probabilistic framework, (2.9) is equivalent to the following minimization problem

$$\min_{\boldsymbol{w},b,\varepsilon} \quad \mathbb{E}_{\boldsymbol{p}}[\boldsymbol{L}_{\varepsilon}(\boldsymbol{z})] + \varepsilon\nu + \frac{\lambda}{2}\left\|\boldsymbol{w}\right\|^2. \tag{2.10}$$

Noting that

$$\min_{\varepsilon} \quad \mathbb{E}_{\boldsymbol{p}}[\boldsymbol{L}_{\varepsilon}(\boldsymbol{z})] + \varepsilon\nu = \nu\bar{q}_{1-\nu}(|\boldsymbol{z}|) = \left\langle\!\left\langle \boldsymbol{z}\right\rangle\!\right\rangle_{1-\nu},$$

we get

$$\min_{\boldsymbol{w},b} \quad \left\langle\!\left\langle \boldsymbol{z}\right\rangle\!\right\rangle_{1-\nu} + \frac{\lambda}{2}\left\|\boldsymbol{w}\right\|^2. \tag{2.11}$$

A function $\bar{q}_{1-\nu}(\cdot)$ is called the superquantile (or conditional value-at-risk), cf. Rockafellar and Uryasev [2000], and function $\left\langle\!\left\langle\cdot\right\rangle\!\right\rangle_{1-\nu}$ is called the superquantile (or CVaR) norm, cf. Pavlikov and Uryasev [2014]. These concepts are generalized to the infinite-dimensional probability spaces and discussed in greater detail in the subsequent section.

3 Risk Quadrangle Framework

The following subsections formally introduce the notions of quantile (VaR) and superquantile (CVaR) in the context of risk measures. Following the same logic, the superquantile norm is introduced in the context of error measures along with related quadrangles.

3.1 Superquantiles and Dual Superquantile Optimization Formula

When dealing with random variables, especially when it comes to optimization, one has to be able to rank them, i.e., assign a numerical value to each random variable that is of interest to the researcher. It is particularly significant when random variables represent a level of loss or hazard (e.g., pollution, contamination, etc.).

The concept of a risk measure resolves the above issue. Indeed, a risk measure is a mapping that assigns a real number to a random value, thus providing a quantification for it. This quantification is reasonable if a risk measure satisfies various combinations of the following properties: *constant neutrality, convexity, aversity, closedness, monotonicity, and homogeneity.* Measures that possess the first four properties are called *regular* risk measures.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X \in \mathcal{L}^2(\Omega)$ be a real-valued random variable, and the cumulative distribution function be denoted by $F_X(x) = \mathbb{P}(X \leq x),\ x \in \mathbb{R}$.

Definition 3.1 (Regular Risk Measure)**.** A functional $\mathcal{R} : \mathcal{L}^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ is called a *regular measure of risk* if it satisfies the following axioms

(R1) **constant neutrality:** $\mathcal{R}(C) = C, \quad \forall\, C = const.$;

(R2) **convexity:** $\mathcal{R}\left(\lambda X + (1-\lambda)Y\right) \leq \lambda\mathcal{R}(X) + (1-\lambda)\mathcal{R}(Y), \quad \forall\, X, Y$ and $\lambda \in [0,1]$;

(R3) **closedness:** $\{X \in \mathcal{L}^2(\Omega) | \mathcal{R}(X) \leq c\}$ is closed $\forall\, c < \infty$;

(R4) **aversity:** $\mathcal{R}(X) > \mathbb{E}X, \quad \forall\, X \neq const.$

Certainly the most famous regular risk measure is the superquantile (i.e. conditional value-at-risk, tail value-at-risk, average value-at-risk, expected shortfall). Superquantile is not only a risk measure with favorable mathematical properties, but it also can be efficiently optimized, as shown in Rockafellar and Uryasev [2000, 2002].

Definition 3.2 (Quantile)**.** The *quantile* (value-at-risk, VaR) of random variable X at confidence level $\alpha \in [0,1]$ is defined by the following equality

$$q_\alpha(X) = [q^-_\alpha(X), q^+_\alpha(X)], \tag{3.12}$$

where

$$q^-_\alpha(X) = \begin{cases} \sup\left\{x \,\middle|\, F_X(x) < \alpha\right\}, & \alpha \in (0,1] \\ \inf\left\{x \,\middle|\, F_X(x) \geq \alpha\right\}, & \alpha = 0 \end{cases} \tag{3.13}$$

$$q^+_\alpha(X) = \begin{cases} \inf\left\{x \,\middle|\, F_X(x) > \alpha\right\}, & \alpha \in [0,1) \\ \sup\left\{x \,\middle|\, F_X(x) \leq \alpha\right\}, & \alpha = 1 \end{cases} \tag{3.14}$$

If $q^-_\alpha(X) = q^+_\alpha(X)$ then

$$q_\alpha(X) = q^-_\alpha(X) = q^+_\alpha(X).$$

Remark 3.1 (Sum and scaling of quantiles)**.** Since quantile is a set, the sum of two quantiles is defined as a sum of sets, i.e., for $\alpha_1, \alpha_2 \in [0,1]$

$$q_{\alpha_1}(X) + q_{\alpha_2}(X) = \left\{\, v + w \,\middle|\, v \in q_{\alpha_1}(X), w \in q_{\alpha_2}(X) \,\right\}. \tag{3.15}$$

The scaling of quantile by an arbitrary constant $\lambda \in \mathbb{R}$ is defined as follows

$$\lambda q_{\alpha_1}(X) = \left\{\, \lambda w \,\middle|\, w \in q_{\alpha_1}(X) \,\right\}. \tag{3.16}$$

Definition 3.3 (Superquantile)**.** The *superquantile* (CVaR, TVaR, AVaR, ES) of random variable X at confidence level $\alpha \in [0,1]$ is defined by the following equality

$$\bar{q}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 q^-_\beta(X)\, d\beta, \quad \alpha \in (0,1). \tag{3.17}$$

For $\alpha = 0$:

$$\bar{q}_0(X) = \lim_{\varepsilon \to 0} \bar{q}_\varepsilon(X) = \mathbb{E}X.$$

For $\alpha = 1$:

$$\bar{q}_1(X) = q^-_1(X), \text{ if } q^-_1(X) < \infty.$$

The following Theorem 3.1 is a celebrated result by Rockafellar and Uryasev [2002].

Theorem 3.1 (Superquantile Optimization Formula)**.** *For random variable X and $\alpha \in (0,1)$, it holds*

$$\bar{q}_\alpha(X) = \min_C \left\{ C + \frac{1}{1-\alpha}\mathbb{E}[X-C]_+ \right\}, \tag{3.18}$$

and the set of minimizers for (3.18) is $q_\alpha(X)$.

It appears, cf. Guan, Giao, and Wang [2022], that there is a much deeper relation between the mean excess function

$$\mathbb{E}[X-x]_+, \quad x \in \mathbb{R}$$

and superquantile. Theorem 3.2 below reveals this profound connection.

Theorem 3.2 (Dual Superquantile Optimization Formula)**.** *For random variable X and $x \in \mathbb{R}$, it holds*

$$\mathbb{E}[X-x]_+ = \max_{\alpha\in[0,1]} \{(1-\alpha)(\bar{q}_\alpha(X)-x)\}, \tag{3.19}$$

and the set of maximizers for (3.19) is $[\mathbb{P}(X<x), \mathbb{P}(X\le x)]$.

Proof. cf. Appendix B. Proof for Theorem 3.2. □

Remark 3.2. Note that $(1-\alpha)\bar{q}_\alpha(X)$ is a concave function of α, cf. Rockafellar and Uryasev [2002]. Therefore, (3.19) is a concave optimization problem.

3.2 Superquantile Norm and Related Quadrangles

The notion of a norm definitely lies in the essence of analysis. Moreover, it not only plays a key role in theory but is also crucial to various mathematical applications. In particular, when it comes to the approximation of functions, a specific norm is usually used as a performance metric. Though, in stochastic optimization, when a random variable X represents an approximation error, it is common to use a general measure of error $\mathcal{E}(X)$ introduced and developed in Rockafellar et al. [2008].

Superquantile norm, introduced in Pavlikov and Uryasev [2014] in $\mathbb{R}^n$ and extended by Mafusalov and Uryasev [2016] to $\mathcal{L}^1(\Omega)$, is a particular case of a regular measure of error.

Definition 3.4 (Regular Error Measure)**.** A functional $\mathcal{E} : \mathcal{L}^2(\Omega) \to \mathbb{R}^+ \cup \{+\infty\}$ is called a *regular measure of error* if it satisfies the following axioms:

(E1) **zero neutrality:** $\mathcal{E}(0) = 0$;

(E2) **convexity:** $\mathcal{E}(\lambda X + (1-\lambda)Y) \le \lambda\mathcal{E}(X) + (1-\lambda)\mathcal{E}(Y), \quad \forall\, X, Y$ and $\lambda \in [0,1]$;

(E3) **closedness:** $\{X \in \mathcal{L}^2(\Omega) | \mathcal{E}(X) \le c\}$ is closed $\forall\, c < \infty$;

(E4) **robustness:** $\mathcal{E}(X) > 0, \quad \forall\, X \neq 0$.

Definition 3.5 (Scaled Superquantile Norm)**.** Let $X \in \mathcal{L}^1(\Omega)$ be a real-valued random variable. Then *scaled superquantile norm* of X with parameter $\alpha \in [0,1]$ is defined by

$$\langle\!\langle X \rangle\!\rangle_\alpha^S = \bar{q}_\alpha(|X|). \tag{3.20}$$

Formally, when referring to the superquantile norm, we by default assume its *scaled* version. However, following the Pavlikov and Uryasev [2014] below we define an equivalent *non-scaled* superquantile norm.

Definition 3.6 (Non-scaled Superquantile Norm)**.** Let $X \in \mathcal{L}^1(\Omega)$ be a real-valued random variable. Then *non-scaled superquantile norm* of X with parameter $\alpha \in [0, 1)$ is defined by

$$\langle\!\langle X \rangle\!\rangle_\alpha = (1-\alpha)\bar{q}_\alpha\left(|X|\right). \tag{3.21}$$

Mafusalov and Uryasev [2016] proved the following Proposition 3.3, which introduces the Superquantile Norm Quadrangle (for the definition of a quadrangle and related information, cf. Appendix A. Theoretical Background).

Proposition 3.3 (Superquantile Norm Quadrangle)**.** *For $\alpha \in [0, 1)$ the error measure $\mathcal{E}(X) = \langle\!\langle X \rangle\!\rangle_\alpha$ generates the following regular quadrangle:*

$$\begin{aligned}
\mathcal{S}(X) &= \frac{1}{2}\left(q_{(1-\alpha)/2}(X) + q_{(1+\alpha)/2}(X)\right),\\
\mathcal{R}(X) &= \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X)\right),\\
\mathcal{D}(X) &= \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X-\mathbb{E}X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X-\mathbb{E}X)\right),\\
\mathcal{V}(X) &= \langle\!\langle X \rangle\!\rangle_\alpha + \mathbb{E}X,\\
\mathcal{E}(X) &= \langle\!\langle X \rangle\!\rangle_\alpha.
\end{aligned}$$

Additionally of interest to us will be the following Proposition 3.4, which presents a quadrangle that is closely related to one in Proposition 3.3.

Proposition 3.4 (Quantile Symmetric Average Quadrangle)**.** *Let $X \in \mathcal{L}^2(\Omega)$, $x \geq 0$ and*

$$\mathcal{A}_x = \left\{ \alpha \in [0,1) \,\middle|\, \frac{1}{2}(q^-_{(1+\alpha)/2}(X) - q^-_{(1-\alpha)/2}(X)) \leq x \leq \frac{1}{2}(q^+_{(1+\alpha)/2}(X) - q^+_{(1-\alpha)/2}(X)) \right\}.$$

Then for $x \geq 0$, $\mathcal{E}(X) = \mathbb{E}[|X| - x]_+$ is an error measure[3], which generates the following quadrangle:

$$\begin{aligned}
\mathcal{S}(X) &= \bigcup_{\alpha \in \mathcal{A}_x} \left\{ \frac{1}{2}\left(q_{(1-\alpha)/2}(X) + q_{(1+\alpha)/2}(X)\right) \right\},\\
\mathcal{R}(X) &= \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X)\right) - (1-\alpha)x, \quad \forall\, \alpha \in \mathcal{A}_x\\
\mathcal{D}(X) &= \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X-\mathbb{E}X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X-\mathbb{E}X)\right) - (1-\alpha)x, \quad \forall\, \alpha \in \mathcal{A}_x\\
\mathcal{V}(X) &= \mathbb{E}[|X| - x]_+ + \mathbb{E}X,\\
\mathcal{E}(X) &= \mathbb{E}[|X| - x]_+ = \textit{Vapnik error.}
\end{aligned}$$

Proof. Relying on the Theorem 3.2 consider the following equality

$$\min_C \left\{ \mathbb{E}[|X - C| - x]_+ \right\} = \min_C \max_{\alpha \in [0,1)} \left\{ \langle\!\langle X - C \rangle\!\rangle_\alpha - (1-\alpha)x \right\}. \tag{3.22}$$

[3]The name "Vapnik error" introduced in this paper is inspired by Vapnik's ε-insensitive loss function (2.3).

Then by Remark 3.2 and Sion's minimax theorem, equality (3.22) can be equivalently rewritten as follows

$$\min_C \left\{ \mathbb{E}[|X-C|-x]_+ \right\} = \max_{\alpha\in[0,1)} \min_C \left\{ \langle\!\langle X-C \rangle\!\rangle_\alpha - (1-\alpha)x \right\}. \tag{3.23}$$

Furthermore, Proposition 3.3 implies that for each $\alpha \in [0,1)$

$$\begin{aligned} &\min_C \left\{ \langle\!\langle X-C \rangle\!\rangle_\alpha - (1-\alpha)x \right\} \\ &\quad = \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X) \right) - (1-\alpha)x - \mathbb{E}X, \end{aligned} \tag{3.24}$$

where optimal $C^* = \frac{1}{2}\left(q_{(1-\alpha)/2}(X) + q_{(1+\alpha)/2}(X) \right)$. By plugging (3.24) in (3.22), we get

$$\begin{aligned} &\min_C \left\{ \mathbb{E}[|X-C|-x]_+ \right\} = \\ &= \max_{\alpha\in[0,1)} \left\{ \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X) \right) - (1-\alpha)x - \mathbb{E}X \right\} \\ &= \max_{\alpha\in[0,1)} \left\{ \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X) \right) - \alpha x \right\} - x - \mathbb{E}X. \end{aligned} \tag{3.25}$$

Denote

$$\theta_x(\alpha) = \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X) \right) - \alpha x. \tag{3.26}$$

Then Remark 3.2 implies that $\theta_x(\alpha)$ is a concave function of α. Thus α belongs to $\operatorname{argmax}_\alpha \theta_x(\alpha)$ if and only if

$$\frac{\partial^+ \theta_x(\alpha)}{\partial \alpha} \le 0 \le \frac{\partial^- \theta_x(\alpha)}{\partial \alpha}, \quad x \ge 0$$

which implies that cf. Rockafellar and Uryasev [2002]

$$\frac{1}{2}(q^-_{(1+\alpha)/2}(X) - q^-_{(1-\alpha)/2}(X)) \le x \le \frac{1}{2}(q^+_{(1+\alpha)/2}(X) - q^+_{(1-\alpha)/2}(X)), \quad x \ge 0 \tag{3.27}$$

For $x \ge 0$ let

$$\mathcal{A}_x = \left\{ \alpha \in [0,1) \,\middle|\, \frac{1}{2}(q^-_{(1+\alpha)/2}(X) - q^-_{(1-\alpha)/2}(X)) \le x \le \frac{1}{2}(q^+_{(1+\alpha)/2}(X) - q^+_{(1-\alpha)/2}(X)) \right\}$$

be a set of points that satisfy (3.27). Then (3.25) implies that

$$\begin{aligned} \mathcal{D}(X) &= \min_C \left\{ \mathcal{E}(X-C) \right\} \\ &= \min_C \left\{ \mathbb{E}[|X-C|-x]_+ \right\} \\ &= \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(X) + (1-\alpha)\bar{q}_{(1+\alpha)/2}(X) \right) - (1-\alpha)x - \mathbb{E}X, \quad \alpha \in \mathcal{A}_x, \end{aligned} \tag{3.28}$$

where

$$\mathcal{S}(X) = \bigcup_{\alpha\in\mathcal{A}_x} \left\{ \frac{1}{2}\left(q_{(1-\alpha)/2}(X) + q_{(1+\alpha)/2}(X) \right) \right\}$$

is the minimizer for (3.28).

Finally, $\mathcal{R}(X) = \mathcal{D}(X) + \mathbb{E}X$ and $\mathcal{V}(X) = \mathcal{E}(X) + \mathbb{E}X$ imply a complete quadrangle quartet. □

Remark 3.3 (Existence of α)**.** Double inequality (3.27) implies that $x = 0$ if and only if $\alpha = 0$ and $x \to \infty$ if and only if $\alpha \to 1$. Therefore, for any $x \geq 0$ there always exists $\alpha \in [0, 1)$ such that $\alpha \in \mathcal{A}_x$.

Remark 3.4 (Regularity of the Quantile Symmetric Average Quadrangle)**.** Note that in general, the Vapnik error $\mathcal{E}(X) = \mathbb{E}[|X| - x]_+$ is not regular for each $x \geq 0$, since it fails to satisfy the robustness axiom, i.e., there exists $X \in \mathcal{L}^2(\Omega),\ X \not\equiv 0$ such that $\mathcal{E}(X) = 0$. Precisely, $\mathcal{E}(X) = 0$ for all $X \in \mathcal{L}^2(\Omega)$ such that $|X| \leq x$ almost surely for $x \geq 0$. This issue may be resolved by providing a reasonable upper bound for $x \geq 0$.

On the other hand, Superquantile Norm Quadrangle is regular, and it will play a crucial role in the further development of SVR.

Remark 3.5 (Mixed Quantile Quadrangle, cf. Rockafellar and Uryasev [2013])**.** A statistic is not unique with respect to the choice of an error. Here is an example of a quadrangle which also has an average of the two median range quantiles as its statistic. Let $\alpha_1 = (1-\alpha)/2,\ \alpha_2 = 1-\alpha_1$

$$
\begin{aligned}
\mathcal{S}(X) &= \frac{1}{2}\left(q_{\alpha_1}(X) + q_{\alpha_2}(X)\right), \\
\mathcal{R}(X) &= \frac{1}{2}\left(\bar{q}_{\alpha_1}(X) + \bar{q}_{\alpha_2}(X)\right), \\
\mathcal{D}(X) &= \frac{1}{2}\left(\bar{q}_{\alpha_1}(X - \mathbb{E}X) + \bar{q}_{\alpha_2}(X - \mathbb{E}X)\right), \\
\mathcal{V}(X) &= \min_{B_1,B_2}\Big\{ \frac{1}{2}\sum_{k=1}^{2} \mathcal{V}_{\alpha_k}(X - B_k) \,\Big|\, \frac{1}{2}\sum_{k=1}^{2} B_k = 0 \Big\}, \\
\mathcal{E}(X) &= \min_{B_1,B_2}\Big\{ \frac{1}{2}\sum_{k=1}^{2} \mathcal{E}_{\alpha_k}(X - B_k) \,\Big|\, \frac{1}{2}\sum_{k=1}^{2} B_k = 0 \Big\},
\end{aligned}
$$

where $\mathcal{V}_{\alpha_k}(X) = \dfrac{1}{1-\alpha_k} EX_+,\ \mathcal{E}_{\alpha_k}(X) = E\left[\dfrac{\alpha_k}{1-\alpha_k} X_+ + X_-\right].$

4 SVR as Generalized Regression

This section considers SVR as a regularized regression corresponding to the respective quadrangles. Firstly, we extend both SVR formulations to the stochastic case and establish their equivalence through the Dual Superquantile Optimization Formula. Then we discuss the statistical side of SVR, namely its estimation properties. Finally, we derive a dual formulation of the ν-SVR and discuss its nonlinear extension using the well-known kernel trick.

To take SVR formulations to the infinite-dimensional stochastic case, we consider the respective probabilistic formulation of both ν-SVR and ε-SVR and, instead of having an empirical dataset (2.1), we deal with the generalized regression problem.

Problem 4.1 (Generalized Regression)**.** *Given the random vector of independent identically distributed random variables $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ find a function from a given class of functions $f \in \mathcal{F}$, which solves the following optimization problem for approximating a given random variable Y*

$$
\min_{f} \quad \mathcal{E}(Z_f), \quad \textit{where } Z_f = Y - f(\boldsymbol{X}),\ f \in \mathcal{F}. \tag{4.29}
$$

Remark 4.1 (Parametric case and regularization)**.** Usually, in practice, one specifies a parametric class of functions $f(\boldsymbol{w}, \boldsymbol{X})$, where $\boldsymbol{w} \in \mathcal{W}$ is a given space of parameters (e.g., $\mathbb{R}^d$, $d \in \mathbb{N}$). In this setting, it is common to include a regularization penalty into the objective of the generalized regression problem (4.29), i.e.,

$$\min_{\boldsymbol{w}} \quad \mathcal{E}(Z_f) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2, \quad \text{where } Z_f = Y - f(\boldsymbol{w}, \boldsymbol{X}),\ f \in \mathcal{F},\ \boldsymbol{w} \in \mathcal{W},\ \lambda \in \mathbb{R}^+. \tag{4.30}$$

SVR definitely fits into this picture, since it is precisely a minimization of a regular error measure with a regularization penalty, where $\mathcal{F}$ is a class of affine functions, i.e., $f_{\boldsymbol{w},b}(\boldsymbol{X}) = \boldsymbol{w}^T\boldsymbol{X} + b$ and $\mathcal{W} = \mathbb{R}^{n+1}$.

4.1 The ε-SVR and ν-SVR as Error Minimization Problems

Below we formulate the ε-SVR and ν-SVR as stochastic optimization problems by employing the notion of error measures.

Property 4.1. Consider optimization problem (2.6) and let $Z(\boldsymbol{w}, b) = Y - (\boldsymbol{w}^T\boldsymbol{X} + b)$ be a real-valued random variable, where $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ is a random vector of independent identically distributed random variables (regressors) and Y be the target random variable (regressant). With this notation (2.6) is equivalent to the following stochastic optimization problem

$$\min_{\boldsymbol{w},b} \quad \mathbb{E}\left[|Z(\boldsymbol{w}, b)| - \varepsilon\right]_+ + \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \tag{4.31}$$

Analogously to the result of Takeda and Sugiyama [2008] for $E\nu$-SVM, we formulate the following property of the ν-SVR.

Property 4.2. Consider optimization problem (2.11) with $\nu = 1 - \alpha$ and let $Z(\boldsymbol{w}, b) = Y - (\boldsymbol{w}^T\boldsymbol{X} + b)$ be a real-valued random variable, where $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ is a random vector of independent identically distributed random variables (regressors) and Y be the target random variable (regressant). With this notation, (2.11) is equivalent to the following stochastic optimization problem

$$\min_{\boldsymbol{w},b} \quad \langle\!\langle Z(\boldsymbol{w}, b) \rangle\!\rangle_\alpha + \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \tag{4.32}$$

Given that the optimal solution of (2.6) is also the optimal solution (2.11), the natural question is whether this is the case for (4.31) and (4.32). The following Proposition 4.2 answers this question in the most complete and elegant way.

Proposition 4.2 (The ε-SVR & ν-SVR Equivalence)**.** *Let $Y - (\boldsymbol{w}^T\boldsymbol{X} + b) = Z(\boldsymbol{w}, b) \in \mathcal{L}^2(\Omega)$ be a real-valued random variable, and $\lambda > 0$ be a regularization constant. Then*

(i) for each $\alpha \in [0, 1)$ and for each $\varepsilon \in q_\alpha(|Z(\boldsymbol{w}^*, b^*)|)$ an optimal solution vector $(\boldsymbol{w}^*, b^*)$ of (4.32) is also an optimal solution vector of (4.31);

(ii) for each $\varepsilon > 0$ and for each $\alpha \in [\mathbb{P}(|Z(\boldsymbol{w}^*, b^*)| < \varepsilon), \mathbb{P}(|Z(\boldsymbol{w}^*, b^*)| \leq \varepsilon)) = \mathcal{I}_\varepsilon$ an optimal solution vector $(\boldsymbol{w}^*, b^*)$ of (4.31) is also an optimal solution vector of (4.32).

Proof. The proof of the proposition relies on the Theorem 3.2. By letting $X = |Z(\boldsymbol{w}, b)|$, $x = \varepsilon$, and adding $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ to the both sides of equation (3.19), we obtain

$$\mathbb{E}\left[|Z(\boldsymbol{w}, b)| - \varepsilon\right]_+ + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 = \max_{\alpha \in [0,1)} \left\{ \langle\!\langle Z(\boldsymbol{w}, b) \rangle\!\rangle_\alpha - (1-\alpha)\varepsilon + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \right\}, \tag{4.33}$$

where $[\mathbb{P}(|Z(\boldsymbol{w}, b)| < \varepsilon), \mathbb{P}(|Z(\boldsymbol{w}, b)| \leq \varepsilon))$ is the set of maximizers for (4.33).

Now, fix $\alpha \in [0, 1)$ and consider the optimization problem (4.32). Let $(\boldsymbol{w}^*, b^*)$ be an optimal solution of (4.32). Then (4.33) implies that for each $\varepsilon \in q_\alpha(|Z(\boldsymbol{w}^*, b^*)|)$

$$\mathbb{E}\left[|Z(\boldsymbol{w}^*, b^*)| - \varepsilon\right]_+ + \frac{\lambda}{2}\|\boldsymbol{w}^*\|^2 = \langle\!\langle Z(\boldsymbol{w}^*, b^*) \rangle\!\rangle_\alpha - (1-\alpha)\varepsilon + \frac{\lambda}{2}\|\boldsymbol{w}^*\|^2 .$$

Since equality (4.33) holds for any $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, then $(\boldsymbol{w}^*, b^*)$ is also an optimal solution of (4.31) for all $\varepsilon \in q_\alpha(|Z(\boldsymbol{w}^*, b^*)|)$, which proves *(i)*.

To prove *(ii)*, fix $\varepsilon > 0$ and let $(\boldsymbol{w}^*, b^*)$ be an optimal solution of (4.31). Then (4.33) implies that for each $\alpha \in \mathcal{I}_\varepsilon$

$$\mathbb{E}\left[|Z(\boldsymbol{w}^*, b^*)| - \varepsilon\right]_+ + \frac{\lambda}{2}\|\boldsymbol{w}^*\|^2 = \langle\!\langle Z(\boldsymbol{w}^*, b^*) \rangle\!\rangle_\alpha - (1-\alpha)\varepsilon + \frac{\lambda}{2}\|\boldsymbol{w}^*\|^2 .$$

Therefore, $(\boldsymbol{w}^*, b^*)$ is also an optimal solution of (4.32) for all $\alpha \in \mathcal{I}_\varepsilon$. □

Remark 4.2. Note that the regression residual $Z(\boldsymbol{w}, b) = Y - (\boldsymbol{w}^T\boldsymbol{X} + b)$ can have a more general form, i.e., $Z(\boldsymbol{w}, b) = Y - f(\boldsymbol{w}, \boldsymbol{X})$, where $f \in \mathcal{F}$ is a class of function that is wider than the class of affine functions. In other words, we do not have to restrict ourselves to linear regression. However, in the case of ℓ^2 regularization, the class of affine functions is sufficient since one may apply a kernel trick.

4.2 The ε-SVR and ν-SVR as Deviation Minimization Problems

This section studies ε-SVR and ν-SVR as stochastic optimization problems through the concept of deviation measures. Rockafellar et al. [2008] proved a theorem, which connects the generalized regression problem with minimization of deviation. Below a generalized version of the theorem, which has been proved in Rockafellar and Uryasev [2013], is presented.

Theorem 4.3 (Error Shaping Decomposition of Regression)**.** *Consider problem (4.29) for a random vector $\boldsymbol{X}$ and Y in the case of $\mathcal{E}$ being a regular measure of error and $\mathcal{F}$ being a class of functions $f : \mathbb{R}^n \to \mathbb{R}$ such that*

$$f \in \mathcal{F} \implies f + C \in \mathcal{F} \text{ for all } C \in \mathbb{R}.$$

Let $\mathcal{D}$ and $\mathcal{S}$ correspond to $\mathcal{E}$. Problem (4.29) is equivalent then to:

$$\text{minimize } \mathcal{D}(Z_f) \text{ over all } f \in \mathcal{F} \text{ such that } 0 \in \mathcal{S}(Z_f). \tag{4.34}$$

Corollary 4.4 (Linear Regression)**.** *Consider the generalized regression problem (4.29), where $\mathcal{F}$ is a class of affine functions, hence $Z_f = Y - \boldsymbol{w}^T\boldsymbol{X} - b$ and denote by $\bar{Z}(\boldsymbol{w}) = Y - \boldsymbol{w}^T\boldsymbol{X}$. Then the problem (4.29) is equivalent to* [4]

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \mathcal{D}(\bar{Z}(\boldsymbol{w})) \\ s.t. \quad & b \in S(\bar{Z}(\boldsymbol{w})). \end{aligned} \tag{4.35}$$

[4] Although the expression $b \in S(\bar{Z}(\boldsymbol{w}))$ is formally a constraint, the problem (4.35) should be understood as a two-stage optimization problem, where one first solves the minimization problem and then uses the optimal solution vector $\boldsymbol{w}$ to calculate the intercept b.

Proof. The corollary is a direct consequence of the Theorem 4.3. □

Corollary 4.5 (Error Shaping Decomposition of SVR)**.** *Let* $Y-(\boldsymbol{w}^T\boldsymbol{X}+b)=Z(\boldsymbol{w},b)\in\mathcal{L}^2(\Omega)$ *be a real-valued random variable, and* $\lambda>0$ *be a regularization constant. Then*

(i) $\forall\,\alpha\in\mathcal{A}_\varepsilon$[5] *problem* (4.31) *is equivalent to*

$$\begin{aligned}\min_{\boldsymbol{w}}\quad & \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(\bar{Z}(\boldsymbol{w}))+(1-\alpha)\bar{q}_{(1+\alpha)/2}(\bar{Z}(\boldsymbol{w}))\right)-\mathbb{E}[\bar{Z}(\boldsymbol{w})]-(1-\alpha)\varepsilon+\frac{\lambda}{2}\|\boldsymbol{w}\|^2\\ s.t.\quad & b\in\bigcup_{\alpha\in\mathcal{A}_x}\left\{\frac{1}{2}\left(q_{(1-\alpha)/2}(\bar{Z}(\boldsymbol{w}))+q_{(1+\alpha)/2}(\bar{Z}(\boldsymbol{w}))\right)\right\}\end{aligned}\tag{4.36}$$

(ii) $\forall\,\alpha\in[0,1)$ *problem* (4.32) *is equivalent to*

$$\begin{aligned}\min_{\boldsymbol{w}}\quad & \frac{1}{2}\left((1+\alpha)\bar{q}_{(1-\alpha)/2}(\bar{Z}(\boldsymbol{w}))+(1-\alpha)\bar{q}_{(1+\alpha)/2}(\bar{Z}(\boldsymbol{w}))\right)-\mathbb{E}[\bar{Z}(\boldsymbol{w})]+\frac{\lambda}{2}\|\boldsymbol{w}\|^2\\ s.t.\quad & b\in\frac{1}{2}\left(q_{(1-\alpha)/2}(\bar{Z}(\boldsymbol{w}))+q_{(1+\alpha)/2}(\bar{Z}(\boldsymbol{w}))\right)\end{aligned}\tag{4.37}$$

(iii) $\forall\,\alpha\in[0,1)\ \exists\,\varepsilon\geq 0$: *an optimal solution vector* $(\boldsymbol{w}^*,b^*)$ *of* (4.37) *is an optimal solution vector of* (4.36) *and moreover* $\alpha\in\mathcal{A}_\varepsilon$.

(iv) $\forall\,\varepsilon\geq 0\ \exists\,\alpha\in\mathcal{A}_\varepsilon$: *an optimal solution vector* $(\boldsymbol{w}^*,b^*)$ *of* (4.36) *is an optimal solution vector of* (4.37).

Proof. Items *(i),(ii)* follow from the Theorem 4.3, since $\|\boldsymbol{w}\|$ obviously does not depend on b.
Items *(iii),(iv)* are direct implication of Theorem 4.3 and Proposition 4.2. □

4.3 Estimation

This subsection discusses the topic of generalized regression in the risk quadrangle framework. As described in Problem 4.1, the regression task is to approximate a random variable Y by a function $f(\boldsymbol{X})\in\mathcal{F}$ of the random vector $\boldsymbol{X}$. By the regression being "generalized" we mean that the approximation error (residual) $Z_f=Y-f(\boldsymbol{X})$ may be assessed for its nonzeroness by a regular error measure $\mathcal{E}$. The function f is called an *estimator*, and the class of functions $\mathcal{F}$ is called a *class of estimators*. The function $\hat{f}\in\mathcal{F}$ is called *the best estimator* in the class $\mathcal{F}$ with respect to a particular choice of an error measure $\mathcal{E}$ if it solves the problem (4.29). The following Regression Theorem from Rockafellar and Uryasev [2013] provides a link between the function $\hat{f}$ and a statistic $\mathcal{S}$ that corresponds to an error measure $\mathcal{E}$.

Theorem 4.6 (Regression)**.** *Consider problem (4.29) for a random vector* $\boldsymbol{X}$ *and* Y *in the case of* $\mathcal{E}$ *being a regular measure of error and* $\mathcal{F}$ *being a class of functions* $f:\mathbb{R}^n\to\mathbb{R}$ *such that*

$$f\in\mathcal{F}\quad\Longrightarrow\quad f+C\in\mathcal{F}\ \text{ for all } C\in\mathbb{R}.$$

[5] $\mathcal{A}_\varepsilon=\mathcal{A}_x$ for $\varepsilon=x$, where $\mathcal{A}_x$ is defined in Proposition 3.4

Let $\mathcal{D}$ and $\mathcal{S}$ correspond to $\mathcal{E}$. Moreover let $\mathcal{E}$ be of ***expectation type*** *and let $\mathcal{F}$ include a function f satisfying*

$$\begin{array}{l} f(\boldsymbol{x}) \in \mathcal{S}\left(Y \,\middle|\, \boldsymbol{x}\right) \;\; \textit{almost surely for } \boldsymbol{x} \in D, \\ \textit{where } Y \,\middle|\, \boldsymbol{x} = Y_{\boldsymbol{X}=\boldsymbol{x}} \;\; \textit{(conditional distribution)}, \end{array} \tag{4.38}$$

with D being the support of the distribution in $\mathbb{R}^n$ induced by $\boldsymbol{X}$.[6]

Then $\hat{f}$ solves the regression problem and ***estimates*** *this conditional statistic*[7] *in the sense that*

$$\hat{f}(\boldsymbol{X}) \in \mathcal{S}\left(Y \,\middle|\, \boldsymbol{X}\right) \;\; \textit{almost surely.} \tag{4.39}$$

Remark 4.3 (Equivalence of statistics)**.** Consider the Superquantile Norm and Quantile Symmetric Average quadrangles. Then Corollary 4.5 (without the regularization term) implies that[8]

$$\forall\, \alpha \in [0,1)\; \exists\, x : x \in q_\alpha(|Z_{\hat{f}}|) \text{ and } \alpha \in \mathcal{A}_x. \tag{4.40}$$

In other words, the regression with respect to each error estimates the same statistic.

Before going straight to SVR, let's consider two classical examples[9], to illustrate the main statement of the theorem.

Example 4.1 (Least Squares)**.** In the "least squares" approach $\mathcal{E}(X) = \mathbb{E}[X]^2$, thus the regression problem is as follows

$$\min_{f \in \mathcal{F}} \quad \mathbb{E}[Z_f]^2. \tag{4.41}$$

Solving this problem, we obtain that the best estimator is $\hat{f}(\boldsymbol{X}) = \mathbb{E}\left[Y \,\middle|\, \boldsymbol{X}\right]$.

Example 4.2 (Quantile Regression)**.** In the quantile regression, we have

$$\min_{f \in \mathcal{F}} \quad \mathbb{E}\left[\frac{\alpha}{1-\alpha}(Z_f)_+ + (Z_f)_-\right]. \tag{4.42}$$

The best estimator is $\hat{f}(\boldsymbol{X}) \in q_\alpha\left(Y \,\middle|\, \boldsymbol{X}\right)$ (conditional quantile).

Now, consider the Vapnik error, i.e., $\mathcal{E}(X) = \mathbb{E}[|X| - x]_+,\; x \geq 0$. According to Proposition 3.4, the best estimator is

$$\hat{f}(\boldsymbol{X}) \in \bigcup_{\alpha \in \mathcal{A}_x} \left\{ \frac{1}{2}\left(q_{(1-\alpha)/2}\left(Y \,\middle|\, \boldsymbol{X}\right) + q_{(1+\alpha)/2}\left(Y \,\middle|\, \boldsymbol{X}\right)\right) \right\}. \tag{4.43}$$

On the other hand, consider the superquantile norm, i.e., $\mathcal{E}(X) = \langle\!\langle X \rangle\!\rangle_\alpha$. Clearly, this error is not of expectation type, however, the Remark 4.3 implies that

$$\hat{f}(\boldsymbol{X}) \in \frac{1}{2}\left(q_{(1-\alpha)/2}\left(Y \,\middle|\, \boldsymbol{X}\right) + q_{(1+\alpha)/2}\left(Y \,\middle|\, \boldsymbol{X}\right)\right), \tag{4.44}$$

[6] Almost surely, in (4.38), refers to this distribution.

[7] It is assumed, for this part, that the distribution of $Y(\boldsymbol{x})$ for $\boldsymbol{x} \in D$ belongs to $\mathcal{L}^2(\Omega)$, and the same then for the random variable $Y(\boldsymbol{X})$ obtained from it.

[8] In the set $\mathcal{A}_x$, X should be substituted with $Z_{\hat{f}}$.

[9] For more examples, cf. Rockafellar and Uryasev [2013]

where the choice of $\alpha \in [0,1)$ generates an interval $q_\alpha(|Z_{\hat{f}}|)$ such that any point from this interval corresponds to the choice of $x \geq 0$ from the Vapnik error, which is certainly of expectation type.

Finally, note that SVR is simply a regularized version of generalized regression with respect to the regular error measures. Moreover, Proposition 4.2 and Remark 4.3 assure the equivalence of both formulations, thus from now on, we only consider the ν-SVR[10] in our further analysis.

Regarding the estimation properties of SVR, we note that adding a regularization penalty biases the estimator hence the best estimator, in this case, is a biased version of (4.44). However, by letting the regularization parameter tend to zero as the sample size increases, we obtain that the best estimator is the asymptotically unbiased estimator of the average of two symmetric conditional quantiles (4.44). For instance, in the most commonly used SVR solver LIBSVM, cf. Chang and Lin [2011], the regularization parameter $\lambda = \dfrac{1}{Cl}$, where $l \in \mathbb{N}$ is a sample size, and thus $\lambda \to 0$ as $l \to \infty$. The following property summarizes the aforementioned.

Property 4.3 (SVR estimation)**.** Given the dataset (2.1), consider the following generalized regression problem

$$\min_{\boldsymbol{w}} \quad \langle\!\langle Z_f \rangle\!\rangle_\alpha + \frac{\lambda}{2} \|\boldsymbol{w}\|^2, \quad f \in \mathcal{F},\ \boldsymbol{w} \in \mathcal{W} \tag{4.45}$$

where $\lambda \to 0$ as $l \to \infty$. Then the best estimator $\hat{f} \in \mathcal{F}$ is the asymptotically unbiased estimator of the average of two symmetric conditional quantiles, i.e., $\hat{f}$ is given by the (4.44).

4.4 Dual Formulation and Kernalization

This subsection deals with the dual formulation of SVR and implements a well-known kernel trick, which provides an elegant generalization of SVR to the nonlinear case. Before directly jumping to the derivation of the dual formulation, let's introduce a couple of definitions and notations to avoid further confusion.

Let $\mathbb{X}$ be a normed space over $\mathbb{R}$ with norm $\|\cdot\|$ (i.e., $\|x\| \in \mathbb{R}$ for $x \in \mathbb{X}$). Then, the dual normed space denoted by $\mathbb{X}^*$ is defined as the set of all continuous linear functionals from $\mathbb{X}$ into $\mathbb{R}$.

Definition 4.1 (Dual norm)**.** For $f \in \mathbb{X}^*$, the *dual norm*, denoted by $\|\cdot\|_*$ of f is defined by

$$\|f\|_* = \sup\left\{\, |f(x)| : x \in \mathbb{X},\ \|x\| \leq 1 \,\right\} = \sup\left\{ \frac{|f(x)|}{\|x\|} : x \in \mathbb{X},\ x \neq 0 \right\}. \tag{4.46}$$

Definition 4.2 (Conjugate function)**.** Let $f : \mathbb{X} \to \mathbb{R} \cup \{\infty\}$. Then a function on $\mathbb{X}^*$, defined by the following equality

$$f^*(y) = \sup_{x \in \mathbb{X}} \{\langle y, x \rangle - f(x)\}, \tag{4.47}$$

is called a *conjugate* of f or the *Legendre-Young-Fenchel transformation*.

[10]The choice of the ν-SVR is preferable for two reasons. First, the error measure, in this case, is a norm and this simplifies duality. Second, the Vapnik error has regularity issues, cf. Remark 3.4, which makes ε-SVR less attractive.

Now, we will derive the dual formulation of the problem (2.11). Let $\alpha = 1-\nu$ and $\lambda = 1/C$ and note that $\langle\!\langle\cdot\rangle\!\rangle_\alpha = (1-\alpha)\langle\!\langle\cdot\rangle\!\rangle_\alpha^S$. Then (2.11) is equivalent to the following primal SVR problem

$$\begin{aligned} \mathbf{p}^\star := \min_{\boldsymbol{w},b,z} \quad & C(1-\alpha)\langle\!\langle \boldsymbol{z}\rangle\!\rangle_\alpha^S + \frac{1}{2}\left\|\boldsymbol{w}\right\|^2 \\ \text{s.t.} \quad & \boldsymbol{z} = \boldsymbol{y} - \hat{\mathbf{X}}\boldsymbol{w} - \mathbf{1}_l b. \end{aligned} \tag{4.48}$$

Define the Lagrangian[11]

$$\begin{aligned} \mathcal{L}(\boldsymbol{w},b,\boldsymbol{z},\boldsymbol{\mu}) &= C(1-\alpha)\langle\!\langle \boldsymbol{z}\rangle\!\rangle_\alpha^S + \frac{1}{2}\left\|\boldsymbol{w}\right\|^2 + \boldsymbol{\mu}^T(\boldsymbol{y} - \hat{\mathbf{X}}\boldsymbol{w} - \mathbf{1}_l b - \boldsymbol{z}) \\ &= -\left(\boldsymbol{\mu}^T\boldsymbol{z} - C(1-\alpha)\langle\!\langle \boldsymbol{z}\rangle\!\rangle_\alpha^S\right) - \left((\hat{\mathbf{X}}^T\boldsymbol{\mu})^T\boldsymbol{w} - \frac{1}{2}\left\|\boldsymbol{w}\right\|^2\right) + \boldsymbol{\mu}^T\boldsymbol{y} - \boldsymbol{\mu}^T\mathbf{1}_l b. \end{aligned}$$

Then

$$\min_{\boldsymbol{w},b,\boldsymbol{z}} \quad \mathcal{L}(\boldsymbol{w},b,\boldsymbol{z},\boldsymbol{\mu}) = \min_b \quad -C(1-\alpha)\left(\langle\!\langle \boldsymbol{\mu}/C(1-\alpha)\rangle\!\rangle_\alpha^S\right)^* - \left(\frac{1}{2}\left\|\hat{\mathbf{X}}^T\boldsymbol{\mu}\right\|^2\right)^* + \boldsymbol{\mu}^T\boldsymbol{y} - \boldsymbol{\mu}^T\mathbf{1}_l b,$$

which leads to the implicit constraint $\boldsymbol{\mu}^T\mathbf{1}_l = 0$. Note that in general, cf. Boyd and Vandenberghe [2004], for $\boldsymbol{x} \in \mathbb{R}^n$

$$\left\|\boldsymbol{x}\right\|^* = \begin{cases} 0 & \left\|\boldsymbol{x}\right\|_* \le 1 \\ \infty & \text{otherwise,} \end{cases}$$

and

$$\left(\frac{1}{2}\left\|\boldsymbol{x}\right\|^2\right)^* = \frac{1}{2}\left\|\boldsymbol{x}\right\|_*^2.$$

Hence

$$-C(1-\alpha)\left(\langle\!\langle \boldsymbol{\mu}/C(1-\alpha)\rangle\!\rangle_\alpha^S\right)^* = \begin{cases} 0 & \langle\!\langle \boldsymbol{\mu}\rangle\!\rangle_{\alpha*}^S \le C(1-\alpha) \\ -\infty & \text{otherwise,} \end{cases}$$

and

$$\left(\frac{1}{2}\left\|\hat{\mathbf{X}}^T\boldsymbol{\mu}\right\|^2\right)^* = \frac{1}{2}\left\|\hat{\mathbf{X}}^T\boldsymbol{\mu}\right\|_*^2.$$

Finally, we note that cf. Mafusalov and Uryasev [2016],

$$\langle\!\langle \boldsymbol{x}\rangle\!\rangle_{\alpha*}^S = \max\left\{\ \left\|\boldsymbol{x}\right\|_1, n(1-\alpha)\left\|\boldsymbol{x}\right\|_\infty\ \right\}.$$

Therefore, the dual formulation of SVR is as follows

$$\begin{aligned} \mathbf{d}^\star := \max_{\boldsymbol{\mu}} \quad & \boldsymbol{\mu}^T\boldsymbol{y} - \frac{1}{2}\left\|\hat{\mathbf{X}}^T\boldsymbol{\mu}\right\|_*^2 \\ \text{s.t.} \quad & \left\|\boldsymbol{\mu}\right\|_1 \le C(1-\alpha),\ \left\|\boldsymbol{\mu}\right\|_\infty \le \frac{C}{l} \\ & \boldsymbol{\mu}^T\mathbf{1}_l = 0. \end{aligned} \tag{4.49}$$

[11] The Lagrange multiplier $\boldsymbol{\mu} = (\mu^1, \ldots, \mu^l)^T$.

Let's consider the classical case, which naturally leads to kernalization, namely the case, where we have SVR with ℓ^2 regularization penalty. The dual formulation (4.49) becomes

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \boldsymbol{\mu}^T \boldsymbol{y} - \frac{1}{2}\boldsymbol{\mu}^T \hat{\mathbf{X}}\hat{\mathbf{X}}^T \boldsymbol{\mu} \\ \text{s.t.} \quad & \|\boldsymbol{\mu}\|_1 \leq C(1-\alpha),\ \|\boldsymbol{\mu}\|_\infty \leq \frac{C}{l} \\ & \boldsymbol{\mu}^T \mathbf{1}_l = 0. \end{aligned} \tag{4.50}$$

Introducing the kernel function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and noting that the objective function in (4.50) depends on feature vectors only through their inner product, we define the kernel matrix

$$\mathbf{K} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)_{i,j=1}^l,$$

and substitute $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ in (4.50) with $\mathbf{K}$, thus obtaining the nonlinear extension of SVR

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \boldsymbol{\mu}^T \boldsymbol{y} - \frac{1}{2}\boldsymbol{\mu}^T \mathbf{K}\boldsymbol{\mu} \\ \text{s.t.} \quad & \|\boldsymbol{\mu}\|_1 \leq C(1-\alpha),\ \|\boldsymbol{\mu}\|_\infty \leq \frac{C}{l} \\ & \boldsymbol{\mu}^T \mathbf{1}_l = 0. \end{aligned} \tag{4.51}$$

Let us compare (4.51) with the classical SVR dual formulation from Schölkopf et al. [2000]

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \boldsymbol{y} - \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \mathbf{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \\ \text{s.t.} \quad & (\boldsymbol{\alpha} + \boldsymbol{\alpha}^*)^T \mathbf{1}_l \leq C(1-\alpha), \\ & (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \mathbf{1}_l = 0, \\ & 0 \leq \alpha_i, \alpha_i^* \leq \frac{C}{l}, \quad i = 1, \ldots, l. \end{aligned} \tag{4.52}$$

Notice that problems (4.51) and (4.52) have equivalent primal formulations, (2.11) and (2.7) respectively, in the sense that from an optimal solution of one problem an optimal solution for the other can be constructed. Therefore, dual problems are also equivalent. However, one may also notice that problem (4.52) has twice more unknown variables as (4.51). Solvers such as Portfolio Safeguard[12] (PSG) that work directly with convex functions can benefit from problem statement (4.51). On the other hand, the most popular SVR solver such as LIBSVM, cf. Chang and Lin [2011], works with (4.52).

5 Numerical Experiments

The following case study implements SVR for simulated data and numerically confirms the

(a) equivalence between ε-SVR and ν-SVR based on Proposition 4.2;

(b) error shaping decomposition of SVR based on Corollary 4.5;

[12]Download from `http://www.aorda.com/`

(c) equivalence between the primal problem statement (4.48) and the dual problem statement (4.50) ;

The case study results, data, and codes are posted at this link[13].

As a true law $f(x) = x,\ x \in [0,1]$ is chosen with $[0,1]$ interval being uniformly partitioned by points $x_i,\ i = 1, \ldots, l$. Then depended variable is simulated as follows

$$y_i = x_i + \epsilon_i, \quad i = 1, \ldots, l,$$

where error terms $\epsilon_i \sim \text{Laplace}(0,1)$ are distributed according to the Laplace distribution with density

$$\rho(x; a, d) = \frac{1}{2d} \exp\left(-\frac{|x-a|}{2d}\right), \quad a = 0,\ d = 1.$$

For the numerical implementation of SVR, the PSG package is utilized. Mathematical programming problems in PSG are formulated in a concise format, which makes problem structures transparent and easy to understand. This is achieved by representing objectives and constraints with a set of standardized functions with clear engineering interpretations.

5.1 Primal Problem Formulations.

This section deals with the primal problem formulations and their equivalence. First, we solve the SVR problem as an error minimization problem. Second, we solve it as a deviation minimization problem (with the same parameters setting) and numerically confirm this approach. Finally, the SVR problem is solved in its dual formulation with the linear kernel.

Regularized Error Minimization. To numerically establish the equivalence between the ν-SVR and ε-SVR, we fix $\alpha = 0.6,\ l = 1000,\ C = 1,\ \lambda = \dfrac{1}{2Cl}$, and solve the optimization problem (4.32) with ℓ^2 penalty. Then we set $\varepsilon = q_\alpha(|Z(\boldsymbol{w}^*, b^*)|)$, which can be done in PSG by using the function `var_risk(`α`, matrix)`, and solve (4.31). Having the solution of (4.31), we calculate the midpoint of the interval $\mathcal{I}_\varepsilon$ from Proposition 4.2, which can be done in PSG by using the function `pr_pen(`ε`, matrix)` and then setting $\alpha_{new} = 1 -$`pr_pen(`ε`, matrix)`. The equivalence then follows from $\alpha \approx \alpha_{new}$.

Regularized Deviation Minimization. To numerically confirm the error shaping decomposition of SVR, we solve (4.37) using the same parameters as for error minimization. This is done by first minimizing the deviation and then calculating `var_risk(`$\frac{1-\alpha}{2}$`, matrix)` and `var_risk(`$\frac{1+\alpha}{2}$`, matrix)` separately in PSG. Finally, we set

$$b^* = \frac{1}{2}\left(\texttt{var\_risk((}1-\alpha\texttt{)/2, matrix)} + \texttt{var\_risk((}1+\alpha\texttt{)/2, matrix)}\right). \tag{5.53}$$

5.2 Dual Problem Formulations

Consider the dual problem (4.50). To make this problem equivalent to the primal with $\lambda = \dfrac{1}{2Cl}$, we set $C \to Cl$ in (4.50). Moreover, to speed up the calculation we replace the constraint

[13] `http://uryasev.ams.stonybrook.edu/index.php/research/testproblems/advanced-statistics/support-vecto`

$\|\boldsymbol{\mu}\|_\infty \leq C$ with the so-called box constraint $-C \leq \mu^i \leq C,\ i = 1, \ldots, l$. With these adjustments (4.50) can be rewritten as

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \boldsymbol{\mu}^T \boldsymbol{y} - \frac{1}{2} \boldsymbol{\mu}^T \hat{\mathbf{X}} \hat{\mathbf{X}}^T \boldsymbol{\mu} \\ \text{s.t.} \quad & \|\boldsymbol{\mu}\|_1 \leq Cl(1-\alpha),\ \boldsymbol{\mu}^T \mathbf{1}_l = 0, \\ & -C \leq \mu^i \leq C,\ i = 1, \ldots, l. \end{aligned} \tag{5.54}$$

To numerically confirm the equivalence between the primal and dual problem, we first solve the dual problem (5.54) in PSG and then set $\boldsymbol{w}^* = \hat{\boldsymbol{X}} \boldsymbol{\mu}^*$ and calculate the intercept b^* using (5.53).

5.3 Summary

This section summarizes the results of all numerical experiments that have been conducted.

Method	Uncertainty Measure	$b^*, \boldsymbol{w}^*$	α	ε	Solving Time (s)
ν-SVR (primal)	error	0.020089, 0.932221	0.6	0.914845	0.01
ε-SVR (primal)	error	0.020089, 0.932221	0.600380	0.914845	0.01
ν-SVR (primal)	deviation	0.019983, 0.932221	0.6	0.914739	0.01
ν-SVR (dual)	error	0.019974, 0.932232	0.6	0.914740	0.09

Table 1: Optimization outputs: SVR, primal and dual formulations.

Optimization outputs from Table 1 numerically confirm the equivalence between SVR formulations.

6 Conclusion

This paper investigated SVR in the fundamental risk quadrangle framework. It provided theoretical insights into the statistical nature of ν-SVR, and ε-SVR formulations. Much like the VC theory, the risk quadrangle perfectly serves as an extended framework for SVR and links this machine learning technique with classical statistics and risk management.

The contribution of this paper is threefold. Firstly, it has been shown that ε-SVR and ν-SVR correspond to a minimization of the Vapnik error and the superquantile norm respectively with a regularization penalty. More importantly, the quadrangle that corresponds to the Vapnik error has been constructed. Secondly, the equivalence of these two error measures has been established theoretically through the dual superquantile optimization formula and numerically by conducting the case study. Finally, based on the equivalence result it has been concluded that SVR is the asymptotically unbiased estimator of the average of two conditional median range quantiles. Additionally, a general dual formulation of the SVR has been derived, which potentially may lead to more efficient numerical algorithms.

Moreover, all theoretical results were confirmed by the numerical experiments on simulated data.

Acknowledgement

Authors would like to thank **TBD**

Appendix A. Theoretical Background

The fundamental risk quadrangle paradigm was developed in Rockafellar and Uryasev [2013]. This framework established a profound connection between risk management, reliability, statistics, and stochastic optimization theories. In particular, the risk quadrangle theory provides a unified framework for generalized regression.

$$\begin{array}{ccc} \text{risk } \mathcal{R} & \longleftrightarrow & \mathcal{D} \text{ deviation} \\ \uparrow & \mathcal{S} & \uparrow \\ \text{regret } \mathcal{V} & \longleftrightarrow & \mathcal{E} \text{ error} \end{array}$$

Diagram 1: The Fundamental Risk Quadrangle

The risk quadrangle methodology united risk functions for a random value X in groups (quadrangles) consisting of five elements:

- Risk $\mathcal{R}(X)$, which provides a numerical surrogate for the overall hazard in X.
- Deviation $\mathcal{D}(X)$, which measures the "nonconstancy" in X as its uncertainty.
- Error $\mathcal{E}(X)$, which measures the "nonzeroness" in X.
- Regret $\mathcal{V}(X)$, which measures the "regret" in facing the mix of outcomes of X.
- Statistic $\mathcal{S}(X)$ associated with X through $\mathcal{E}$ and $\mathcal{V}$.

The following diagram contains general relationships between elements of the quadrangle:

$$\mathcal{D}(X) = \min_C \big\{ \mathcal{E}(X - C) \big\} = \mathcal{R}(X - \mathbb{E}X)$$

$$\mathcal{R}(X) = \min_C \big\{ C + \mathcal{V}(X - C) \big\} = \mathbb{E}X + \mathcal{D}(X)$$

$$\mathcal{S}(X) = \operatorname*{argmin}_C \big\{ \mathcal{E}(X - C) \big\} = \operatorname*{argmin}_C \big\{ C + \mathcal{V}(X - C) \big\}$$

$$\mathcal{E}(X) = \mathcal{V}(X) - \mathbb{E}X, \qquad \mathcal{V}(X) = \mathbb{E}X + \mathcal{E}(X)$$

Diagram 2: The Relationship Formulae

where $\mathbb{E}X$ denotes the mathematical expectation of X and the statistic, $\mathcal{S}(X)$, can be a set if the minimum is achieved for multiple points.

Probably the most famous quadrangle is the quantile quadrangle, cf. Rockafellar and Uryasev (2013), named after the quantile statistic. This quadrangle establishes relations between the superquantile (CVaR) optimization technique described in Rockafellar and Uryasev [2000, 2002] and quantile regression, cf. Koenker and Bassett [1978], Koenker [2005]. In particular, it was shown that superquantile minimization and the quantile regression are similar procedures based on the quantile statistic in the regret and error representation of risk and deviation.

Theorem 6.1 (Quadrangle Theorem)**.** *Let $X \in \mathcal{L}^2$. Then*

(a) *The relations $\mathcal{D}(X) = \mathcal{R}(X) - EX$ and $\mathcal{R}(X) = EX + \mathcal{D}(X)$ give a one-to-one correspondence between regular measures of risk $\mathcal{R}$ and regular measures of deviation $\mathcal{D}$. In this correspondence, $\mathcal{R}$ is positively homogeneous if and only if $\mathcal{D}$ is positively homogeneous. On the other hand,*

$$\mathcal{R} \text{ is monotonic if and only if } \mathcal{D}(X) \le \sup X - EX \text{ for all } X. \tag{6.55}$$

(b) *The relations $\mathcal{E}(X) = \mathcal{V}(X) - EX$ and $\mathcal{V}(X) = EX + \mathcal{E}(X)$ give a one-to-one correspondence between regular measures of regret $\mathcal{V}$ and regular measures of error $\mathcal{E}$. In this correspondence, $\mathcal{V}$ is positively homogeneous if and only if $\mathcal{E}$ is positively homogeneous. On the other hand,*

$$\mathcal{V} \text{ is monotonic if and only if } \mathcal{E}(X) \le |EX| \text{ for } X \le 0. \tag{6.56}$$

(c) *For any regular measure of regret $\mathcal{V}$, a regular measure of risk $\mathcal{R}$ is obtained by*

$$\mathcal{R}(X) = \min_C \Big\{ C + \mathcal{V}(X - C) \Big\}. \tag{6.57}$$

If $\mathcal{V}$ is positively homogeneous, $\mathcal{R}$ is positively homogeneous. If $\mathcal{V}$ is monotonic, $\mathcal{R}$ is monotonic.

(d) *For any regular measure of error $\mathcal{E}$, a regular measure of deviation $\mathcal{D}$ is obtained by*

$$\mathcal{D}(X) = \min_C \Big\{ \mathcal{E}(X - C) \Big\}. \tag{6.58}$$

If $\mathcal{E}$ is positively homogeneous, $\mathcal{D}$ is positively homogeneous. If $\mathcal{E}$ satisfies the condition in (6.56)*, then $\mathcal{D}$ satisfies the condition in* (6.55)*.*

(e) *In both (c) and (d), as long as the expression being minimized is finite for some C, the set of C values for which the minimum is attained is a nonempty, closed, bounded interval.*[14] *Moreover, when $\mathcal{V}$ and $\mathcal{E}$ are paired as in (b), the interval comes out the same and gives the associated statistic:*

$$\operatorname*{argmin}_C \big\{ C + \mathcal{V}(X - C) \big\} = \mathcal{S}(X) = \operatorname*{argmin}_C \big\{ \mathcal{E}(X - C) \big\}, \quad \text{with } \mathcal{S}(X + C) = \mathcal{S}(X) + C. \tag{6.59}$$

Appendix B. Proof for Theorem 3.2

Guan, Giao, and Wang [2022] proved this theorem under the assumption that $(\Omega, \mathcal{F}, \mathbb{P})$ is an atomless probability space. In the following proof, this assumption is dropped.

Proof. Denote by

$$f(\alpha) = -(1 - \alpha)\bar{q}_\alpha(X),$$

and consider the conjugate function of $f(\alpha)$, i.e.,

$$f^*(x) = \max_{\alpha \in [0,1]} \{x\alpha + (1 - \alpha)\bar{q}_\alpha(X)\}, \quad x \in \mathbb{R}. \tag{6.60}$$

[14] Typically this interval reduces to a single point.

Since $f(\alpha)$ is a concave function, cf. Rockafellar and Uryasev [2002], then α belongs to $\operatorname{argmax}_{\alpha} \{x\alpha + (1-\alpha)\bar{q}_\alpha(X)\}$ if and only if

$$\frac{\partial^+}{\partial x}(x\alpha + (1-\alpha)\bar{q}_\alpha(X)) \leq 0 \leq \frac{\partial^-}{\partial x}(x\alpha + (1-\alpha)\bar{q}_\alpha(X)), \quad x \in \mathbb{R}.$$

Thus, cf. Rockafellar and Uryasev [2002]

$$q_\alpha^-(X) \leq x \leq q_\alpha^+(X) \iff \alpha^* \in [\mathbb{P}(X < x), \mathbb{P}(X \leq x)], \quad x \in \mathbb{R}. \tag{6.61}$$

By the Theorem 3.1

$$\bar{q}_\alpha(X) = q_\alpha(X) + \frac{1}{1-\alpha}\mathbb{E}(X - q_\alpha(X))_+$$

which together with (6.61) implies that

$$\begin{aligned} f^*(x) &= x\alpha^* + (1-\alpha^*)\bar{q}_{\alpha^*}(X) \\ &= x\alpha^* + (1-\alpha^*)\left(q_{\alpha^*}(X) + \frac{1}{1-\alpha^*}\mathbb{E}(X - q_{\alpha^*}(X))_+\right) \\ &= x\alpha^* + (1-\alpha^*)x + \mathbb{E}(X-x)_+ \\ &= x + \mathbb{E}(X-x)_+, \quad x \in \mathbb{R}. \end{aligned} \tag{6.62}$$

Therefore,

$$\begin{aligned} \mathbb{E}(X-x)_+ + x &= \max_{\alpha\in[0,1]} \{x\alpha + (1-\alpha)\bar{q}_\alpha(X)\} \\ &\iff \\ \mathbb{E}(X-x)_+ &= \max_{\alpha\in[0,1]} \{(1-\alpha)(\bar{q}_\alpha(X) - x)\}, \end{aligned} \tag{6.63}$$

where $[\mathbb{P}(X < x), \mathbb{P}(X \leq x)]$ are the set of maximizers for (6.63). □

References

Rockafellar, R. T. and Uryasev, S. (2000) "Optimization of conditional value-at-risk," *Journal of Risk* **2**, 21–42.

Rockafellar, R. T. and Uryasev, S. (2002) "Conditional value-at-risk for general loss distributions," *Journal of Banking and Finance* **26**, 1443–1471.

Rockafellar, R. T., Uryasev, S., and Zabarankin, M. (2008) "Risk tuning with generalized linear regression," *Math. of Operations Research* **33** (3), 712–729.

Pavlikov, K., Uryasev, S. (2014). "CVaR norm and applications in optimization," *Optimization Letters* **8** (7), 1999–2020.

Rockafellar, R. T., Uryasev, S. (2013). "The fundamental risk quadrangle in risk management, optimization and statistical estimation," *Surveys in Operations Research and Management Science* **18** (1), 33–53.

Boyd, S., Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V. (1996). "Support vector regression machines," *Advances in Neural Information Processing Systems* **8**, 155-161.

Vapnik, V., Golowich, S. E., Smola, A. (1996). "Support vector method for function approximation, regression estimation and signal processing," *Advances in Neural Information Processing Systems* **9**, 281-287.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Data Mining and Knowledge Discovery, Springer.

Koenker, R. and Bassett, G. (1978) "Regression quantiles," *Econometrica* **46**, 33–50.

Koenker, R. (2005) *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press.

Schölkopf, B., Smola, A. J., Williamson, R. C., Bartlett, P. L. (2000) "New support vector algorithms," *Neural Computation* **12**(5), 1207-1245.

Smola, A. J., Schölkopf, B. (2004) "A tutorial on support vector regression," *Statistics and Computing* **14** (3), 199-222.

Mafusalov, A. and Uryasev, S. (2016) "CVaR (superquantile) norm: stochastic case," *European Journal of Operational Research* **249**, 200–208.

Chang, C.-C., Lin, C.-J. (2001) "Training ν-support vector regression: theory and algorithms," *Neural Computation* **14** (8), 1959–1977.

Chang, C.-C., Lin, C.-J. (2011) "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* **2** (3), 1-27.

Takeda, A. and Sugiyama.M. (2008) "ν-support vector machine as conditional value-at-risk minimization," *In Proceedings of the 25th International Conference on Machine Learning* 1056–1063.

Norton, M., Mafusalov, A., and Uryasev, S. (2017) "Soft margin support vector classification as buffered probability minimization," *Journal of Machine Learning Research* **18**, 1-43.

Guan, Y., Jiao, Z., Wang, R. (2022) "A reverse expected shortfall optimization formula," `https://arxiv.org/abs/2203.02599`