# Learning Value-at-Risk and Expected Shortfall*

D. Barrera†, S. Crépey‡, E. Gobet§, Hoang-Dung Nguyen¶, B. Saadeddine‖

October 31, 2022

**Abstract**

We propose a non-asymptotic convergence analysis of a two-step approach to learn a conditional value-at-risk (VaR) and expected shortfall (ES) in a non-parametric setting using Rademacher and Vapnik-Chervonenkis bounds. Our approach for the VaR is extended to the problem of learning at once multiple VaRs corresponding to different quantile levels. This results in efficient learning schemes based on neural network quantile and least-squares regressions. An a posteriori Monte Carlo procedure is introduced to estimate distances to the ground-truth VaR and ES without access to the latter. This is illustrated using numerical experiments in a Gaussian toy-model and a financial case-study where the objective is to learn a dynamic initial margin.

**Keywords:** value-at-risk, expected shortfall, quantile regression, quantile crossings, neural networks.

**AMS Subject Classification:** 62G32,62L20,62M45,91G60,91G70.

# Contents

**1 Introduction** **2**

* Python notebooks reproducing the results of this paper are available on `https://github.com/BouazzaSE/Learning-VaR-and-ES`. HTML versions of the same notebooks are also available in order to view the experiments and the results on a browser without having to install Jupyter Notebook. Note that, due to GitHub size limitations, the HTML files must be downloaded locally (and then opened with a browser) to be displayed.

†Email: `j.barrerac@uniandes.edu.co`. Departamento de Matemáticas, Universidad de los Andes, Cra 1 # 18a-12, Edificio H. Bogotá, Colombia. Postal code: 111711.

‡Email: `stephane.crepey@lpsm.paris`. LPSM, Université Paris Cité, France. The research of S. Crépey benefited from the support of the Chair Stress Test, RISK Management and Financial Steering, led by the French Ecole polytechnique and its Foundation and sponsored by BNP Paribas

§Email: `emmanuel.gobet@polytechnique.edu`. Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France.

¶Email: `hoangdung.nguyen@natixis.com`. PhD student, Université Paris Cité, France. The research of H.-D. Nguyen is funded by a CIFRE grant from Natixis.

‖Email: `bouazza.saadeddine@ca-cib.com`. PhD student, University of Evry in Paris-Saclay and Quantitative research GMD/GMT Credit Agricole CIB, Paris.

**2 A Learning Algorithm for VaR and ES** **4**
2.1 VaR and ES as optimization problems . . . . . 5
2.2 The algorithm . . . . . 7

**3 Convergence Analysis of the Learning Algorithm** **8**
3.1 The approximation error of the estimator of VaR . . . . . 9
3.2 A confidence interval for the estimator of VaR . . . . . 13
3.3 A Rademacher confidence interval for the estimator of ES − VaR . . . . 17
3.4 A Vapnik-Chervonenkis confidence interval for the estimator of ES − VaR 19
3.5 A Posteriori Monte Carlo Validation of VaR and ES learners . . . . . 21

**4 Learning Using Neural Networks** **24**
4.1 Error bound of the single-$\alpha$ learning algorithm with one-layer neural networks . . . . . 24
4.2 Learning the VaR . . . . . 26
4.3 Learning the ES using a two-step approach . . . . . 27

**5 Multi-$\alpha$ learning for VaR** **28**
5.1 Related literature . . . . . 28
5.2 Extension of the bounds to multi-$\alpha$ learning . . . . . 29
5.3 Multi-$\alpha$ learning using neural networks . . . . . 30

**6 Conditionally Gaussian Toy Model** **34**
6.1 Numerical Results . . . . . 34

**7 Dynamic Initial Margin Case Study** **36**
7.1 Numerical Results . . . . . 38

**A Value-at-Risk and Expected Shortfall Representations** **41**

**B The Role of Data Transformations and Truncations** **44**

# 1 Introduction

Quantile regression is a classical statistical problem that has received attention since the 1750s. Koenker (2017) notes that the least absolute criterion (or pinball loss function) for the median even preceded the least squares for the mean (introduced by Legendre in 1805).

Quantile regression is commonly done in the context of linear models, where the ensuing minimization problem can be cast as a linear program and subsequently solved by the simplex method. When several quantile levels are jointly considered, a flaw inherent to linear quantile regression is the problem of crossing quantile curves. Alternative approaches include nonlinear quantile regression based on interior point methods (Koenker and Park, 1996) or nonparametric quantile regression often implemented by stochastic gradient descent methods (Rodrigues and Pereira, 2020).

In harmony with the numerous financial applications, we also refer to quantile as value-at-risk (VaR) and to superquantile, i.e. the expected loss given the loss exceeds the VaR (Rockafellar and Royset, 2013), as expected shortfall (ES). Dimitriadis and Bayer (2019) developed an asymptotic convergence analysis, establishing the consistency and asymptotic normality, under somewhat strong semiparametric assumptions and regularity conditions, of a joint linear regression estimator for the value-at-risk and expected shortfall based on their joint elicitability properties (Fissler and Ziegel, 2016; Fissler, Ziegel, and Gneiting, 2016), implemented numerically using the Nelder-Mead optimization algorithm. Closer to our proposals, Padilla, Tansey, and Chen (2020) consider quantile regression with ReLU networks, including a discussion on minimax rates for quantile functions with Hölder-related regularity conditions, and providing qualitative non-asymptotic estimates for such networks, of which our corresponding results can be considered quantitative versions. Shen, Jiao, Lin, Horowitz, and Huang (2021) consider a different approach to the non-asymptotic analysis, assuming that the target quantile function has a compositional structure in terms of Hölder-continuous functions. The authors derive Vapnik–Chervonenkis (VC)-based error bounds that only depend on the dimension of the composed functions, as opposed to the one of the inputs usually in the literature, and are therefore less impacted by the curse of dimensionality.

The contribution of the present paper is the non-asymptotic convergence analysis of a learning algorithm for VaR and ES using a two-step approach, possibly for multiple quantile levels at the same time, in a nonparametric setup. We also provide practical learning schemes to learn the conditional VaR and ES using neural networks as the function approximators. Our two-step methodology enables the reuse of the VaR neural network's hidden layers in the training of the neural network approximating the ES, allowing to learn the latter using a simple linear regression against a learned regression basis, hence quickly deducing a conditional ES predictor from the conditional VaR one. We also address the problem of learning multiple quantiles at the same time and propose methods to deal with the well-known quantile crossing issue (He, 1997; Koenker, 2004; Takeuchi, Le, Sears, and Smola, 2006). We provide an a posteriori error estimation method in order to compute errors against ground-truth values of the conditional VaR and ES, without the need to approximate the latter with a slow nested Monte Carlo procedure. For the purpose of assessing our proposed schemes, we provide numerical experiments in a Gaussian toy model, and a financial case-study where the goal is to learn a dynamic initial margin in a multi-factor model.

The paper is organized as follows. Section 2 presents our base learning algorithm. Relying on the general results of the companion paper (Barrera, 2022, Section 3), Section 3 performs the corresponding convergence analysis. Section 4 discusses specializations of this scheme and its errors to the case of inference via neural networks. We introduce multi-quantile extensions of the above in Section 5. Sections 6 and 7 discuss numerical experiments. Appendix A gathers classical elicitability results underlying different possible VaR and ES learning algorithms (including the one in Section 2, but also a joint representation à la Fissler and Ziegel (2016); Fissler, Ziegel, and Gneiting (2016), shown less efficient numerically in the paper's github). Appendix B discusses the role of data transformations in the scheme proposed and their consequences on the

respective error bounds.

We denote by $(\Omega, \mathcal{A}, \mathrm{P})$ a probability space, which admits all the random variables appearing below (the existence of $(\Omega, \mathcal{A}, \mathrm{P})$ can be verified a posteriori), with corresponding expectation operator denoted by $\mathrm{E}[\,\cdot\,]$, and we denote by $\mathcal{R}$ the Borel sigma algebra on $\mathbb{R}$.

## 2 A Learning Algorithm for VaR and ES

Let $S$ be a Polish space with Borel sigma algebra $\mathcal{S}$. From now on

$$(X, Y) : \Omega \to S \times \mathbb{R}$$

is a fixed random vector in $S \times \mathbb{R}$[1], with $Y \in L^1_{\mathrm{P}}$. We will utilize the usual notation $\mathrm{P}_X$, $\mathrm{P}_{(X,Y)}$ for the laws of $X$ and $(X, Y)$: for every Borel sets $A \subset S$, $A' \subset S \times \mathbb{R}$

$$\mathrm{P}_X(A) = \mathrm{P}[X \in A], \qquad \mathrm{P}_{(X,Y)}(A') = \mathrm{P}\left[(X, Y) \in A'\right].$$

We fix a conditional distribution function $\mu : S \times \mathcal{R} \to [0, 1]$ of $Y$ given $X$ (Kallenberg, 2006, Theorem 5.3 p.84), and we assume that the function $S \times \mathbb{R} \to \mathbb{R}$ defined by $(x, y) \mapsto \mu(x, (-\infty, y])$ is $(\mathcal{S} \otimes \mathcal{R})/\mathcal{R}$ (i.e. Borel)-measurable.[2] With these conventions, we will use implicitly the corresponding version

$$\mathrm{P}[Y \in \cdot | X] := \mu(X, \cdot)$$

of the conditional probability of $Y$ given $X$. In particular, we will use the conditional (cumulative) distribution (function) of $Y$ given $X$,

$$F_{Y|X}(y) := \mathrm{P}[Y \leq y \,|\, X] := \mu(X, (-\infty, y]). \tag{2.1}$$

We will finally assume, without loss of generality, that $F_{Y|X(\omega)}(\cdot)$ is integrable for every $\omega \in \Omega$.[3] In what follows, for a function $F : \mathbb{R} \to \mathbb{R}$ and $y_0 \in \mathbb{R}$

$$F(y_0-) := \lim_{y \uparrow y_0} F(y). \tag{2.2}$$

**Definition 2.1.** *The conditional value-at-risk (*VaR*) and expected shortfall (*ES*) of $Y$ given $X$ at the confidence level $\alpha \in (0, 1)$ are (cf.* (A.2)*)*

$$\begin{aligned} &\mathrm{VaR}(Y|X) := \mathrm{VaR}(F_{Y|X}) = \inf F^{-1}_{Y|X}([\alpha, 1]) = \inf\{y \in \mathbb{R};\, F_{Y|X}(y) \geq \alpha\}, \\ &\mathrm{ES}(Y|X) := \frac{1}{1 - F_{Y|X}(\mathrm{VaR}(Y|X)-)} \int_{[\mathrm{VaR}(Y|X), \infty)} y\, F_{Y|X}(dy). \end{aligned} \tag{2.3}$$

[1] i.e. an $\mathcal{A}/(\mathcal{S} \otimes \mathcal{R})$ measurable function.

[2] This the case if for instance $S = \mathbb{R}^d$ and $(X, Y)$ admits a density with respect to Lebesgue measure.

[3] Since $Y \in L^1_{\mathrm{P}}$, we have that

$$\infty > \mathrm{E}[|Y|] = \mathrm{E}[\mathrm{E}[|Y||X]] = \mathrm{E}\left[\int_{\mathbb{R}} |y| F_{Y|X}(dy)\right],$$

thus $F_{Y|X(\omega)}$ is integrable for P-a.e. $\omega$: it suffices to change the version of $X$ to guarantee integrability for every $\omega$.

**Lemma 2.1.** *The functions* $\omega \mapsto \mathrm{VaR}(Y|X(\omega))$ *and* $\omega \mapsto \mathrm{ES}(Y|X(\omega))$ *are* $\sigma(X)$*-measurable.*

***Proof.*** Given $t \in \mathbb{R}$,

$$\{\mathrm{VaR}(Y|X) < t\} = \cup_{n\in\mathbb{N}}\{F_{Y|X}(t-1/n) \geq \alpha\},$$

which is a countable union of $\sigma(X)$-measurable sets ($F_{Y|X}(y)$ is $\sigma(X)$-measurable for every fixed $y$). This shows the claim for $\mathrm{VaR}(Y|X)$. As for the $\sigma(X)$-measurability of $\mathrm{ES}(Y|X)$, notice that the function $es : S \times \mathbb{R} \to \mathbb{R}$ defined by

$$es(v,x) = \frac{1}{1-\mu(x,(-\infty,v))}\int y\mathbb{1}_{[v,\infty)}(y)\,\mu(x,dy)$$

is Borel-measurable (on the set where $\mu(x,(-\infty,v)) < 1$) and that

$$\mathrm{ES}(Y|X) = es(X, \mathrm{VaR}(Y|X)). \blacksquare$$

Then, from the Doob-Dynkin lemma:

**Corollary 2.2.** *There exist Borel measurable functions* $q : S \to \mathbb{R}$ *and* $s : S \to \mathbb{R}$ *such that*

$$q(X) = \mathrm{VaR}(Y|X), \qquad s(X) = \mathrm{ES}(Y|X), \qquad \mathrm{P}\text{-}a.s.. \tag{2.4}$$

Assuming $S = \mathbb{R}^d$, the goal of the article is to present and analyze algorithms for approximating ($\mathrm{P}_X$-versions of) the functions $q(\cdot)$ and/or $s(\cdot)$ in (2.4), efficient in high dimension $d$, based on i.i.d. samples of $(X,Y) \in \mathbb{R}^d \times \mathbb{R}$ and on suitable hypothesis spaces (including families of functions represented in terms of neural nets which are used in the experimental part of the paper)

$$\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}\}\,,\ \mathcal{G} = \{g : \mathbb{R}^d \to \mathbb{R}\}\,,\ \mathcal{H} = \{h = (f,g) : \mathbb{R}^d \to \mathbb{R}^2\}$$

for $q(\cdot)$, $s(\cdot)$, and $(q(\cdot), s(\cdot))$, respectively.

## 2.1 VaR and ES as optimization problems

Given $\alpha \in (0,1)$ and an increasing, continuously differentiable function $\iota : \mathbb{R} \to \mathbb{R}$, let $\rho_\iota : \mathbb{R}^2 \to \mathbb{R}$ be the loss function defined by

$$\rho_\iota(y,v) = (1-\alpha)^{-1}(\iota(y)-\iota(v))^+ + \iota(v). \tag{2.5}$$

Given a twice continuously differentiable function $\varsigma : [0,\infty) \to \mathbb{R}$ with $\varsigma''$ positive, let

$$\varrho_\varsigma(y,v,z) := \varsigma'(z)\left(z - (1-\alpha)^{-1}(y-v)^+\right) - \varsigma(z), \tag{2.6}$$

e.g.

$$\begin{aligned}\varrho_{.2}(y,v,z) =& z^2 - 2(1-\alpha)^{-1}(y-v)^+ z \\ =& (z-(1-\alpha)^{-1}(y-v)^+)^2 - ((1-\alpha)^{-1}(y-v)^+)^2.\end{aligned} \tag{2.7}$$

Given functions $\iota, \varsigma : \mathbb{R} \to \mathbb{R}$ with $\iota'$ nonnegative (possibly zero) and continuous, $\varsigma'$ negative and $\varsigma''$ non-vanishing, let $\rho_{\iota,\varsigma} : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by

$$\begin{aligned}\rho_{\iota,\varsigma}(y,v,z) =& (1-\alpha)^{-1}(\iota(y)-\iota(v))^+ + \iota(v) \\ &+\varsigma'(z)\left(z - v - (1-\alpha)^{-1}(y-v)^+\right) - \varsigma(z). \end{aligned} \tag{2.8}$$

To provide suitable representations of the functions $q(\cdot)$ and $s(\cdot)$ of (2.4) in the context of convex optimization, we will work under the following assumption.

**Assumption 2.2.** *$F_{Y|X}$ (defined by (2.1) for a given $\alpha$) satisfies Assumption A.4, P-a.s., and if $\rho_\iota$, $\varrho_\varsigma$, $\rho_{\iota,\varsigma}$, $q(\cdot)$ and $s(\cdot)$ are respectively as in (2.5), (2.6), (2.8), (2.4), then $\rho_\iota(Y, q(X))$, $\varrho_\varsigma(Y, q(X), s(X) - q(X))$ and $\rho_{\iota,\varsigma}(Y, q(X), s(X))$ are P-integrable.*

Our methods rely on the following elicitability (i.e. minimizing) properties of $\mathrm{VaR}(Y|X)$ and $\mathrm{ES}(Y|X)$ of the functions $q(\cdot)$ and $s(\cdot)$ . We use implicitly in the statement the convention $\mathrm{E}[h(X,Y)] = \infty$ whenever $h(X,Y)$ is not P-integrable. We also use the notation $\mathcal{L}(S)$ [resp. $\mathcal{L}_+(S)$] for the space of Borel measurable functions $S \to \mathbb{R}$ [resp. $S \to \mathbb{R}_+$].

**Theorem 2.3.** *Under Assumption 2.2:*

$$q(\cdot) \in \arg\min_{f \in \mathcal{L}(S)} \mathrm{E}\left[\rho_\iota(Y, f(X))\right], \tag{2.9}$$

$$s(\cdot) - q(\cdot) \in \arg\min_{g \in \mathcal{L}_+(S)} \mathrm{E}\left[\varrho_\varsigma(Y, q(X), g(X))\right], \tag{2.10}$$

$$(q(\cdot), s(\cdot)) \in \arg\min_{(f,g) \in \mathcal{L}(S)\times\mathcal{L}(S)} \mathrm{E}\left[\rho_{\iota,\varsigma}(Y, f(X), g(X))\right]. \tag{2.11}$$

*Even more*

$$s(X) = q(X) + (1-\alpha)^{-1}\mathrm{E}\left[(Y - q(X))^+ | X\right], \qquad \text{P-}a.s. \tag{2.12}$$

*(this does not depend on the assumptions on $\rho_\iota$, $\varrho_\varsigma$, $\rho_{\iota,\varsigma}$. For the P-integrability of $q(X)$ see the last paragraph in Appendix B).*

***Proof.*** All these statements are a straightforward consequence of the fact that, if $h(X,Y)$ is P-integrable, then

$$\mathrm{E}\left[h(X,Y)|X\right] = \int_{\mathbb{R}} h(X,y) F_{Y|X}(dy), \qquad \text{P-}a.s.,$$

together with the characterizations of VaR and ES in Lemmas A.1, A.2 and A.3.

To illustrate for $q(\cdot)$: using Lemma A.1 and the above identity, we obtain that

$$\mathrm{E}\left[\rho_\iota(Y, q(X))|X\right] \le \mathrm{E}\left[\rho_\iota(Y, f(X))|X\right], \qquad \text{P-}a.s.,$$

for every $f \in \mathcal{L}(S)$. This implies (2.9) by integrating with respect to P. The other statements can be proved in a similar fashion. ∎

**Remark 2.3.** *If* $(Y-q(X))^+ \in L^2_{\mathrm{P}}$*, then the representation* (2.12) *is also a consequence of the characterization* (2.10)*. To see this notice that, by the Pythagorean theorem and the nonnegativity of* $(Y-q(X))^+$*, any* $r \in \mathcal{L}_+(S)$ *satisfying*[4]

$$r(\cdot) \in \arg\min_{g \in \mathcal{L}_+(S)} \mathrm{E}\left[((1-\alpha)^{-1}(Y-q(X))^+ - g(X))^2\right] \tag{2.13}$$

*has the property that*

$$r(X) = \mathrm{E}\left[(1-\alpha)^{-1}(Y-q(X))^+ | X\right], \qquad \mathrm{P}\text{-}a.s..$$

*In view of* (2.7)*, it follows that the minimization criteria* (2.10) *and* (2.13) *are exactly the same, leading in particular to*

$$s(X) - q(X) = r(X) = (1-\alpha)^{-1}\mathrm{E}\left[(Y-q(X))^+ | X\right], \qquad \mathrm{P}\text{-}a.s.,$$

*as claimed by* (2.12)*.*

**Remark 2.4.** *The minimizers in* (2.9)-(2.11) *do not need to be unique: notice for instance that the proof of* (2.9) *(illustrated above) shows that any function* $q_1 : S \to \mathbb{R}$ *satisfying* $F_{Y|X}(q_1(X)) = \alpha$ *is a minimizer of* $f \mapsto \mathrm{E}\left[\rho_\iota(Y, f(X))\right]$*, and that there are infinitely many such functions if* $F^{-1}_{Y|X}(\alpha)$ *is an interval of positive length on a set with positive* $\mathrm{P}_X$*-measure.*

## 2.2 The algorithm

The functional representations in (2.9)-(2.13) give immediately rise to equally many approximation algorithms for conditional VaR and/or ES. In all cases, the numerical recipe is simply that of replacing the minimization problems in (2.9)-(2.13) by empirical versions: instead of $\mathcal{L}(S), \mathcal{L}_+(S)$ and $\mathcal{L}(S) \times \mathcal{L}(S)$ we use convenient hypotheses spaces $\mathcal{F} \subset \mathcal{L}(S)$, $\mathcal{G} \subset \mathcal{L}_+(S)$, and $\mathcal{H} \subset \mathcal{L}(S) \times \mathcal{L}(S)$; instead of integration with respect to P we use a Monte Carlo approximation based on (properly truncated) i.i.d. samples of $(X, Y)$.

After some preliminary empirical investigations reported in the paper's GitHub, the best turned out to be the simplest, i.e. the two-step algorithm that first uses (2.9) to obtain an approximation $\hat{q}(\cdot)$ of the (conditional) VaR, and then uses (2.12) together with the interpretation of the conditional expectation as a least-squares minimization problem, i.e. (2.13), to learn ES, using the approximation $\hat{q}(\cdot)$ obtained before. This two-step algorithm will be our main focus in what follows. Its pseudo-code is provided as Algorithm 1. The restrictions on $\mathcal{F}$ and $\iota$, the transformation $h_1, h_2$ and the truncations $T_B$ defined by $T_B y = \max\{\min\{y, B\}, -B\}, \qquad (y, B) \in \mathbb{R} \times [0, \infty)$, permit a fitting of the algorithm within the framework of the bounds developed in Barrera (2022). They may also have practical advantages, as discussed in Appendix B.

[4] The existence of such $r$ follows again from Kallenberg (2006, Lemma 1.13 p.7).

**Parameters:**

- The loss $\rho$ given by (2.5) with $\iota(z) = z$.
- Constants $(B_1, B_2, B_3) \in (0, \infty)^3$ with $B_1 \leq B_2$.
- A function $h_1 : S \times \mathbb{R} \to [-B_2, B_2]$ such that, for $\mathrm{P}_X$-a.e. $x \in S$, $h_{1,x}(\cdot) := h_1(x, \cdot)$ is increasing in a set $I_x$ with $\mathrm{P}[Y \in I_x | X = x] = 1$.
- A conditionally affine function $h_2(x, y) = \tau(x)y + \nu(x)$ with $\tau(x) > 0$ for $\mathrm{P}_X$ a.e. $x \in S$.
- A set $\mathcal{F}$ of Borel measurable functions $S \to [-B_1, B_1]$.
- A set $\mathcal{G}$ of Borel measurable functions $S \to [0, B_3]$.

**Input:** An i.i.d. sample

$$D = \{(X_k, Y_k)\}_{k=1}^n,$$

of $(X, Y)$.

Compute

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \sum_{k=1}^n \rho(h_1(X_k, Y_k), f(X_k)).$$
$$\hat{q}(x) = h_{1,x}^{-1} \circ \hat{f}(x)$$

Compute

$$\hat{g} \in \arg\min_{g \in \mathcal{G}} \sum_{k=1}^n \left(g(X_k) - T_{B_3}((1-\alpha)^{-1}(h_2(X_k, Y_k) - h_2(X_k, \hat{q}(X_k))^+)\right)^2$$
$$\hat{r}(x) := (\hat{g}(x) - \nu(x))/\tau(x)$$

**Return**$(\widehat{\mathrm{VaR}}(Y|\cdot), \widehat{\mathrm{ES}}(Y|\cdot)) = (\hat{q}(\cdot), \hat{q}(\cdot) + \hat{r}(\cdot))$

**Algorithm 1:** Estimates of conditional VaR and ES by regression in two steps with tilted loss (cf. (2.5)) for VaR and quadratic loss for ES.

# 3 Convergence Analysis of the Learning Algorithm

In what follows, we will be using the assumption $h_k(x, y) = y$ $(k = 1, 2)$ for the data transformations in Algorithm 1. Our results, therefore, leave open the error induced by the operations $(h_k(X, \cdot))^{-1}$ used for the final estimates.

We will use the notation

$$D = \{(X_j, Y_j)\}_{j=1}^n \tag{3.1}$$

for an i.i.d. sample of $(X, Y)$ (with $n$ given).

Using also the notation (2.5), (2.6), (2.8), we will denote, for $(f,g) \in \mathcal{L}(S) \times \mathcal{L}_+(S)$

$$\tilde{\rho}_\iota(f) := \mathrm{E}\left[\rho_\iota(Y, f(X))\right] \ , \ \hat{\rho}_\iota(f) := \frac{1}{n}\sum_{k=1}^{n} \rho_\iota(Y_k, f(X_k))$$

$$\tilde{\varrho}_\varsigma(f,g) := \mathrm{E}\left[\varrho_\varsigma(Y, f(X), g(X))\right] \ , \ \hat{\varrho}_\varsigma(f,g) := \frac{1}{n}\sum_{k=1}^{n} \varrho_\varsigma(Y_k, f(X_k), g(X_k))$$

$$\tilde{\rho}_{\iota,\varsigma}(f,g) = \mathrm{E}\left[\rho_{\iota,\varsigma}(Y, f(X), f(X)+g(X))\right],$$

$$\hat{\rho}_{\iota,\varsigma}(f,g) = \frac{1}{n}\sum_{k=1}^{n} \rho_{\iota,\varsigma}(Y_k, f(X_k), f(X_k)+g(X_k)). \tag{3.2}$$

Throughout this section,

$$\mathcal{F} \subset \mathcal{L}(S), \qquad \mathcal{G} \subset \mathcal{L}_+(S), \qquad\qquad \mathcal{H} \subset \mathcal{L}(S) \times \mathcal{L}_+(S)$$

will be fixed hypothesis spaces. Associated to these and to the loss functions in (3.2) there are the following quantities of interest,

$$\widetilde{q} \in \arg\min_{f\in\mathcal{F}} \tilde{\rho}_\iota(f), \qquad \hat{q} \in \arg\min_{f\in\mathcal{F}} \hat{\rho}_\iota(f) \tag{3.3}$$

and given $f \in \mathcal{L}(S)$,

$$\widetilde{r}_f \in \arg\min_{g\in\mathcal{G}} \tilde{\varrho}_\varsigma(f,g), \qquad \hat{r}_f \in \arg\min_{g\in\mathcal{G}} \hat{\varrho}_\varsigma(f,g). \tag{3.4}$$

Thus (3.3) defines respectively *the best mean and empirical hypothesis for* VaR *within* $\mathcal{F}$, and (3.4) defines *the best mean and empirical hypotheses for* $\mathrm{ES}-\mathrm{VaR}$ *within* $\mathcal{G}$ *conditioned to the hypothesis* $f$ *for* VaR ($f$ may not belong to $\mathcal{F}$). Similarly, we define the *best mean and empirical joint hypotheses for* $(\mathrm{VaR}, \mathrm{ES}-\mathrm{VaR})$ respectively by

$$(\widetilde{q},\widetilde{r}) \in \arg\min_{h=(f,g)\in\mathcal{H}} \tilde{\rho}_{\iota,\varsigma}(f,g), \qquad (\hat{q},\hat{r}) \in \arg\min_{h=(f,g)\in\mathcal{H}} \hat{\rho}_{\iota,\varsigma}(f,g).$$

### 3.1 The approximation error of the estimator of VaR

Algorithm 1 is based on the following assumption:

**Assumption 3.1.** *The function* $\iota : \mathbb{R} \to \mathbb{R}$ *in* (2.5) *is the identity function. We therefore omit* $\iota$ *and write*

$$\rho(y,v) = (1-\alpha)^{-1}(y-v)^+ + v,$$

*as well as* $\tilde{\rho}(\cdot)$ *and* $\hat{\rho}(\cdot)$ *instead of* $\rho_\iota(\cdot,\cdot), \tilde{\rho}_\iota(\cdot)$ *and* $\hat{\rho}_\iota(\cdot)$.

Assumption 3.1 implies the convexity of $\rho(y,\cdot)$ (for all $y$), which we exploit in several manners. In a sense, Assumption 3.1 is only an apparent restriction: notice that for any $(y,v) \in \mathbb{R}^2$

$$\rho_\iota(y,v) = \rho(\iota(y), \iota(v)),$$

which allows us to transport any conclusion under Assumption 3.1 to the respective conclusion for generic $\iota$, by "transferring" the hypotheses related to $(y, v)$ to hypotheses related to $(\iota(y), \iota(v))$.

The following assumption is a conditional version of Assumption A.4:

**Assumption 3.2.** *There exist functions $a, b : S \to \mathbb{R}$ such that*

$$F_{Y|X}(a(X)) < \alpha \leq F_{Y|X}(b(X)), \tag{3.5}$$

*on a set $\Omega_0$ of $\mathrm{P}$-measure one and such that $F_{Y|X(\omega)}(\cdot)$ is absolutely continuous in $[a(X(\omega)), b(X(\omega))]$ for every $\omega \in \Omega_0$.*

Notice that, under this assumption, $a(X) \leq q(X) \leq b(X)$ except on a set of measure zero.

**Assumption 3.3.** *(for a generic family $\mathcal{F}_1 \subset \mathcal{L}(S)$) Assumption 3.2 holds, and $\mathcal{F}_1 \subset \mathcal{L}(S)$ is such that*

1. *For every $f \in \mathcal{F}_1$, $a(X) \leq f(X) \leq b(X)$, except on a set $\Omega_0$ of $\mathrm{P}$-measure zero.*
2. *There exists $c_{\mathcal{F}_1} > 0$ such that, for every $f \in \mathcal{F}_1$,*

$$F'_{Y|X}(f(X)) \geq c_{\mathcal{F}_1}, \qquad \mathrm{P}\text{-}a.s..$$

Assumption 3.3 is needed to succeed in applying Taylor expansions towards the estimation of errors in our analysis.

**Lemma 3.1.** *Given $\mathcal{F} \subset \mathcal{L}(S)$, and under Assumption 3.1, define $\widetilde{q}$ by (3.3), let $\mathcal{F}_0 \subset \mathcal{F}$, and consider*

$$\mathcal{F}_0^* := \{tf + (1-t)q : (t, f) \in [0,1] \times \mathcal{F}_0\}, \qquad \mathcal{F}^* := \{tf + (1-t)q : (t, f) \in [0,1] \times \mathcal{F}\},$$

*If $\mathcal{F}_1 \equiv \mathcal{F}^*$ satisfies Assumption 3.3 and if*

$$C_{\mathcal{F}_0^*} := \sup_{f \in \mathcal{F}_0^*} \{||F'_{Y|X}(f(X))||_{\mathrm{P},\infty}\}, \tag{3.6}$$

*then the inequalities*

$$\begin{aligned} c_{\mathcal{F}^*}||\widetilde{q} - q||^2_{\mathrm{P}_X,2} &\leq 2(1-\alpha)(\tilde{\rho}(\widetilde{q}) - \tilde{\rho}(q)) \\ &\leq (2(2-\alpha) \inf_{f \in \mathcal{F}} ||f - q||_{\mathrm{P}_X,1}) \wedge (C_{\mathcal{F}_0^*} \inf_{f \in \mathcal{F}_0} ||f - q||^2_{\mathrm{P}_X,2}) \end{aligned} \tag{3.7}$$

*hold.*

***Proof.*** For any $f \in \mathcal{F}$, consider the function $[0,1] \to \mathbb{R}$ defined by

$$t \mapsto V_f(t) := \tilde{\rho}(q + t(f - q)),$$

which has a minimum at $t = 0$.

We use the definition of $F_{Y|X}(\cdot)$ and differentiation under the integral sign to obtain, for every $t \in [0,1]$

$$\begin{aligned} V_f''(t) =& \frac{\partial^2}{\partial t^2} \mathrm{E}\left[\int_{\mathbb{R}} \rho(y, q(X) + t(f(X) - q(X))) F_{Y|X}(dy)\right] \\ =& \frac{\partial}{\partial t} \mathrm{E}\left[(f(X) - q(X))((1-\alpha)^{-1}(F_{Y|X}(q(X) + t(f(X) - q(X))) - 1) + 1)\right] \\ =& \ \mathrm{E}\left[(f(X) - q(X))^2 F'_{Y|X}(q(X) + t(f(X) - q(X)))/(1-\alpha)\right] \\ \geq& \frac{c_{\mathcal{F}^*}}{1-\alpha} \mathrm{E}\left[(f(X) - q(X))^2\right]. \end{aligned} \tag{3.8}$$

This shows in particular that $V_f$ is twice continuously differentiable (from the right at $t = 0$) and convex. Applying Taylor's theorem and the fact that $V_f'(0) = 0$ we arrive at

$$\frac{c_{\mathcal{F}^*}}{2(1-\alpha)} ||f - q||^2_{\mathrm{P}_X,2} \leq \tilde{\rho}(f) - \tilde{\rho}(q). \tag{3.9}$$

Since this is valid for any $f \in \mathcal{F}$, it is valid for $f = \widetilde{q}$. This gives

$$\frac{c_{\mathcal{F}^*}}{2(1-\alpha)} ||\widetilde{q} - q||^2_{\mathrm{P}_X,2} \leq \tilde{\rho}(\widetilde{q}) - \tilde{\rho}(q). \tag{3.10}$$

The upper bound

$$\tilde{\rho}(\widetilde{q}) - \tilde{\rho}(q) \leq \frac{C_{\mathcal{F}_0^*}}{2(1-\alpha)} \inf_{f \in \mathcal{F}_0} ||f - q||^2_{\mathrm{P}_X,2} \tag{3.11}$$

follows from the inequality $\tilde{\rho}(\widetilde{q}) - \tilde{\rho}(q) \leq \tilde{\rho}(f) - \tilde{\rho}(q)$ (valid for any $f \in \mathcal{F}_0$) and an obvious modification of the previous argument starting from (3.8).

Finally, the upper bound

$$\tilde{\rho}(\widetilde{q}) - \tilde{\rho}(q) \leq \left(\frac{2-\alpha}{1-\alpha}\right) \inf_{f \in \mathcal{F}} ||f - q||_{\mathrm{P}_X,1} \tag{3.12}$$

follows via an elementary estimation using

$$|a^+ - b^+| \leq |a - b| \tag{3.13}$$

and the triangle inequality, together (again) with the inequality $\tilde{\rho}(\widetilde{q}) - \tilde{\rho}(q) \leq \tilde{\rho}(f) - \tilde{\rho}(q)$, valid for every $f \in \mathcal{F}$. The conclusion follows from (3.10), (3.11) and (3.12). ∎

**Remark 3.4.** *Notice that as $\mathcal{F}_0$ gets larger, $C_{\mathcal{F}_0^*}$ in* (3.6) *increases and* $\inf_{f \in \mathcal{F}_0} ||f - q||_{\mathrm{P}_X,2}$ *decreases: by making the bound* (3.7) *depend of $\mathcal{F}_0 \subset \mathcal{F}$ we leave open the room for a trade-off between these quantities.*

**Remark 3.5.** *If we strengthen Assumption 3.3 by requiring that for some $(c, C) \in (0,\infty) \times (0,\infty)$, and except on a set of* P*-measure zero*

$$c \leq F'_{Y|X}(y) \leq C, \ \textit{for every } y \in [a(X), b(X)], \tag{3.14}$$

*then the conclusion of Lemma 3.1 holds with* $(c_{\mathcal{F}^*}, C_{\mathcal{F}_0^*})$ *replaced by* $(c, C)$ *under the sole assumption that, for every* $f \in \mathcal{F}$,[5]

$$[f(X), q(X)] \cup [q(X), f(X)] \subset [a(X), b(X)], \quad \textit{except on a set of } \mathrm{P}\textit{-measure zero.} \tag{3.15}$$

*As will be illustrated in Examples 3.6 and 3.7, these observations allow weakening the dependence on* $\mathcal{F}$ *in the estimate* (3.7).

**Example 3.6.** *Assume* (3.14) *and, given* $\delta > 0$, *assume that* $\mathcal{F}$ *is such that* (3.15) *holds and*

$$\inf_{f \in \mathcal{F}} ||f - q||_{\mathrm{P}_X, 2} < \delta.$$

*Denoting by* $\widetilde{q}_\delta$ *the solution to the left-hand side of* (3.3), *an application of Remark 3.5 gives that*

$$c||\widetilde{q}_\delta - q||^2_{\mathrm{P}_X, 2} \leq \delta(2(2-\alpha) \wedge C\delta) \leq C\delta^2,$$

*leading to the estimate*

$$||\widetilde{q}_\delta - q||_{\mathrm{P}_X, 2} \leq \left(\frac{C}{c}\right)^{1/2} \delta. \tag{3.16}$$

**Example 3.7.** *To give a concrete instance of the previous example, assume that, for some* $A \leq B$,

$$q(X) \in [A, B]\,,\ \mathrm{P}_X\textit{-a.s.}$$

*(see also Remark A.3), assume that* (3.5) *and* (3.14) *hold with* $a(X) \equiv A$ *and* $b(X) \equiv B$, *and assume that there exists a finite or countable partition* $\{S_j\}_j \subset \mathcal{S}$ *of* $S$ *such that, for all* $j$,

$$||q||_{TV_{S_j}} := \sup_{(x,x') \in S_j \times S_j} |q(x) - q(x')| < \delta$$

*(for instance if* $q$ *is continous, as* $S$ *is a Polish space). Then* (3.16) *holds with*

$$\mathcal{F} = \{x \mapsto \sum_j a_j \mathbb{1}_{S_j}(x) : a_j \in [A, B], \forall j\}.$$

*Partitions* $\{S_j\}_j$ *as above can be available with only partial information on* $q$ *on cases of interests: consider for example the case in which* $S$ *is compact and* $q$ *is uniformly Lipschitz with a known Lipschitz constant.*

[5] $[u, v] \cup [v, u]$ is just the closed segment of the real line determined by $(u, v) \in \mathbb{R}^2$. Notice that (3.15) is exactly the same as 1. in Assumption 3.3 for $\mathcal{F}_1 = \mathcal{F}^*$.

### 3.2 A confidence interval for the estimator of VaR

Let us now give an upper bound for the error in probability associated to the empirical estimator $\hat{q}$ of $\widetilde{q}$. For this, we need to introduce the following measures of complexity applicable to the families of hypotheses used along our schemes:

**Definition 3.8.** *If $S$ is a Polish space, $\mathcal{H} \subset \mathcal{L}(S)$, and $X_{1:n}$ is a random sequence in $S$, the empirical Rademacher complexity $\mathcal{R}_{emp}(\mathcal{H}, X_{1:n})$ and the Rademacher complexity $\mathcal{R}_{ave}(\mathcal{H}, X_{1:n})$ of $\mathcal{H}$ at $X_{1:n}$ are defined as*

$$\mathcal{R}_{emp}(\mathcal{H}, X_{1:n}) = \mathrm{E}\left[\sup_{h\in\mathcal{H}} \sum_{k=1}^{n} U_k h(X_k) \Big| X_{1:n}\right], \qquad \mathcal{R}_{ave}(\mathcal{H}, X_{1:n}) = \mathrm{E}\left[\mathcal{R}_{emp}(\mathcal{H}, X_{1:n})\right]$$

*where $U_{1:n}$ is an i.i.d. Rademacher sequence* $\mathrm{P}[U_k = 1] = \mathrm{P}[U_k = -1] = 1/2$ *independent of $X_{1:n}$.*

The Rademacher complexities have the following property, which we will use later and whose proof is an easy exercise: if

$$co(\mathcal{H}) := \bigcup_m \left\{ \sum_{k=1}^{m} t_k h_k : h_{1:m} \in \mathcal{H}^m, t_{1:m} \in [0,1]^m, \sum_k t_k = 1 \right\} \tag{3.17}$$

is the convex hull of $\mathcal{H}$, and if

$$cobal(\mathcal{H}) = co(\mathcal{H} \cup -\mathcal{H}) \tag{3.18}$$

is the balanced convex hull of $\mathcal{H}$, then

$$\mathcal{R}_{emp}(co(\mathcal{H}), X_{1:n}) = \mathcal{R}_{emp}(\mathcal{H}, X_{1:n}), \qquad \mathcal{R}_{emp}(cobal(\mathcal{H}), X_{1:n}) \leq 2\mathcal{R}_{emp}(\mathcal{H}, X_{1:n}).$$

**Definition 3.9.** *If $S, \mathcal{H}$ and $X_{1:n}$ are as in Definition 3.8, and if $r \geq 0$, the covering number of $\mathcal{H}$ with respect to the empirical $L^1$-norm at $X_{1:n}$, $\mathcal{N}_1(\mathcal{H}, X_{1:n}, r)$, is defined as*

$$\mathcal{N}_1(\mathcal{H}, X_{1:n}, r) := \min\left\{ m \in \mathbb{N} : \exists\, g_{1:m} \in \mathcal{L}^m(S) : \sup_{h\in\mathcal{H}} \min_l \sum_{k=1}^{n} |h(X_k) - g_l(X_k)| < nr \right\}; \tag{3.19}$$

*with the convention* $\inf \emptyset = \infty$. *A sequence $g_{1:m}$ satisfying the condition in* (3.19) *is called an $r$-covering of $\mathcal{H}$ with respect to the empirical $L^1-$norm at $X_{1:n}$.*

In what follows, $(X, Y)_{1:n}$ is the sample (3.1) used to compute $\hat{q}$ and

$$\rho(\mathcal{F}) := \{(x, y) \mapsto \rho(y, f(x)) : f \in \mathcal{F}\},$$

is the family of instantaneous losses associated to $\mathcal{F}$.

**Lemma 3.2.** *Under the hypotheses of Lemma 3.1, and given $\delta \in (0,1)$, the bound*

$$c_{\mathcal{F}^*}||\hat{q}-q||^2_{\mathrm{P}_X,2} \leq \left(2(2-\alpha)\inf_{f\in\mathcal{F}}||f-q||_{\mathrm{P}_X,1}\right) \wedge \left(C_{\mathcal{F}_0^*}\inf_{f\in\mathcal{F}_0}||f-q||^2_{\mathrm{P}_X,2}\right)$$
$$+(1-\alpha)\left(\frac{2^5}{n}\right)^{1/2}\left(\sup_{f\in\mathcal{F}}||\rho(Y,f(X))||_{\mathrm{P},\infty}\left(\log\left(\frac{2}{\delta}\right)\right)^{1/2}+\left(\frac{2}{n}\right)^{1/2}\mathcal{R}_{ave}(\rho(\mathcal{F}),(X,Y)_{1:n})\right) \tag{3.20}$$

*holds with probability at least $1-\delta$. The right-hand side of* (3.20) *can be further upper bounded via the inequalities, valid for every $r>0$*

$$\begin{aligned}\mathcal{R}_{ave}(\rho(\mathcal{F}),D) \leq & ((2-\alpha)/(1-\alpha))\mathcal{R}_{ave}(\mathcal{F},X_{1:n}) \\ \leq & ((2-\alpha)/(1-\alpha))(r+\sqrt{n}\sup_{f\in\mathcal{F}}||f(X)||_{\mathrm{P},\infty}\mathrm{E}\left[\sqrt{2\log(\mathcal{N}_1(\mathcal{F},X_{1:n},r/n))}\right]).\end{aligned}$$

$$\begin{aligned}\mathcal{R}_{ave}(\rho(\mathcal{F}),D) \leq & r+\sqrt{n}\sup_{f\in\mathcal{F}}||\rho(Y,f(X))||_{\mathrm{P},\infty}\mathrm{E}\left[\sqrt{2\log(\mathcal{N}_1(\rho(\mathcal{F}),D,r/n))}\right] \\ \leq & r+\sqrt{n}\sup_{f\in\mathcal{F}}||\rho(Y,f(X))||_{\mathrm{P},\infty}\mathrm{E}\left[\sqrt{2\log(\mathcal{N}_1(\mathcal{F},X_{1:n},(1-\alpha)r/(2-\alpha)n))}\right].\end{aligned} \tag{3.21}$$

**Remark 3.10.** *If $\max\{||Y||_{\mathrm{P},\infty},\sup_{f\in\mathcal{F}}||f(X)||_{\mathrm{P},\infty}\} \leq B$ then, clearly,*

$$\sup_{f\in\mathcal{F}}||\rho(Y,f(X))||_{\mathrm{P},\infty} \leq \left(\frac{2-\alpha}{1-\alpha}\right)B.$$

***Proof.*** (of Lemma 3.2) According to (3.9), for every $f\in\mathcal{F}$

$$c_{\mathcal{F}^*}||f-q||^2_{\mathrm{P}_X,2} \leq 2(1-\alpha)(\tilde{\rho}(f)-\tilde{\rho}(q)),$$

implying in particular that

$$c_{\mathcal{F}^*}||\hat{q}-q||^2_{\mathrm{P}_X,2} \leq 2(1-\alpha)((\tilde{\rho}(\hat{q})-\tilde{\rho}(\widetilde{q}))+(\tilde{\rho}(\widetilde{q})-\tilde{\rho}(q)))$$

The term $2(1-\alpha)(\tilde{\rho}(\widetilde{q})-\tilde{\rho}(q))$ is upper bounded in (3.7). To upper bound $\tilde{\rho}(\hat{q})-\tilde{\rho}(\widetilde{q})$ in probability we apply the Rademacher bound Barrera (2022, (3.38)) taking $Z_k = (X_k,Y_k) \sim (X_1,Y_1)$ i.i.d. and the diagonal family

$$\rho(\mathcal{F})^{(n)}_{1:n} = \{((x_k,y_k))_{k\in 1:n} \mapsto (\rho(f)(x_k,y_k)/n)_{k\in 1:n} : f\in\mathcal{F}\}$$

to obtain the inequality (see also Barrera (2022, eqns. (2.25), (2.26))

$$\begin{aligned}\tilde{\rho}(\hat{q})-\tilde{\rho}(\widetilde{q}) \leq & 2((1/\sqrt{n})\sup_{f\in\mathcal{F}}||\rho(Y,f(X))||_{\mathrm{P},\infty}\sqrt{2\log(2/\delta)}+(2/n)\mathcal{R}_{ave}(\rho(\mathcal{F}),D)) \\ = & (2^3/n)^{1/2}\left(\sup_{f\in\mathcal{F}}||\rho(Y,f(X))||_{\mathrm{P},\infty}(\log(2/\delta))^{1/2}+(2/n)^{1/2}\mathcal{R}_{ave}(\rho(\mathcal{F}),D)\right)\end{aligned} \tag{3.22}$$

with probability at least $1-\delta$. We deduce (3.20) combining (3.22) with the above.

To prove the first inequality in (3.21), note that by Talagrand contraction lemma (Mohri, Rostamizadeh, and Talwalkar (2018, Lemma 4.2 p.78)), since

$$u \mapsto (1-\alpha)^{-1}u^+$$

is $(1-\alpha)^{-1}$-Lipschitz, then for any $(x,y)_{1:n} \subset (S \times \mathbb{R})^n$

$$\begin{aligned}
\mathcal{R}_{emp}(\rho(\mathcal{F}),(x,y)_{1:n}) \leq & \mathcal{R}_{emp}(\{(1-\alpha)^{-1}(y-f)^+ : f \in \mathcal{F}\},(x,y)_{1:n}) + \mathcal{R}_{emp}(\mathcal{F},x_{1:n}) \\
\leq & (1-\alpha)^{-1}\mathcal{R}_{emp}(\{(y-f) : f \in \mathcal{F}\},(x,y)_{1:n}) + \mathcal{R}_{emp}(\mathcal{F},x_{1:n}) \\
\leq & (1-\alpha)^{-1}\mathcal{R}_{emp}(\{y\},y_{1:n}) + \left(\frac{2-\alpha}{1-\alpha}\right)\mathcal{R}_{emp}(\mathcal{F},x_{1:n}) \\
= & \left(\frac{2-\alpha}{1-\alpha}\right)\mathcal{R}_{emp}(\mathcal{F},x_{1:n}),
\end{aligned}$$

which implies the first inequality in (3.21) by integration with respect to the law of $D$.

The second and third inequalities in (3.21) are a direct consequence of Barrera (2022, eqn. (3.47)) and the argument in Barrera (2022, eqn. (3.53)). The fourth follows easily from the fact that if $\mathcal{F}' \subset \mathcal{L}(S)$ is a $(1-\alpha)r/(2-\alpha)$ covering of $\mathcal{F}$ with respect to the empirical $L^1$-norm at $x_{1:n}$, then $\{(x,y) \mapsto y - f(x) | f \in \mathcal{F}'\}$ is an $r$-covering of $\rho(\mathcal{F})$ with respect to the empirical $L^1$-norm at $(x,y)_{1:n}$ (this can be proved using (3.13)).

Let us now introduce the following hypothesis, which covers the estimation error of $\hat{f}$ in Algorithm 1.

**Assumption 3.11.** *For given $0 < B_1 \leq B_2$,*

$$||Y||_{\mathrm{P},\infty} \leq B_2.$$

*In addition,* $\mathrm{VaR}(Y|X)$ *takes values in* $(-B_1, B_1]$ *and* $y \mapsto F_{Y|X(\omega)}[(-\infty, y]]$ *is* P*-a.e. differentiable, with derivative uniformly bounded away from* 0 *and* $\infty$ *in* $[-B_1, B_1]$*. That is,*

$$F_{Y|X}(-B_1) < \alpha \leq F_{Y|X}(B_1), \qquad \mathrm{P}\text{-}a.s.,$$

*and there exist* $0 < c_{B_1} \leq C_{B_1} < \infty$ *such that*

$$c_{B_1} \leq F'_{Y|X}(y) \leq C_{B_1}, \qquad \mathrm{P}\text{-}a.s.,$$

*for every* $y \in [-B_1, B_1]$.

Using Assumption 3.11, the following result follows easily from Lemma 3.2:

**Theorem 3.3.** *Under Assumption 3.11, let*

$$\mathcal{F}' \subset \mathcal{F} \subset co(\mathcal{F}') \subset \mathcal{L}(S) \tag{3.23}$$

*where $\mathcal{F}$ is a family of functions uniformly bounded by $B_1$ (see also* (3.17)*). Then the inequality*

$$c_{B_1}||\hat{q}-q||^2_{\mathrm{P}_X,2} \leq \left(2(2-\alpha)\inf_{f\in\mathcal{F}}||f-q||_{\mathrm{P}_X,1}\right) \wedge \left(C_{B_1}\inf_{f\in\mathcal{F}}||f-q||^2_{\mathrm{P}_X,2}\right)$$
$$+(2^2(2-\alpha)/\sqrt{n})\left(B_2\sqrt{2\log(2/\delta)}+2B_1\left(1+\mathrm{E}\left[\sqrt{2\log(N_1(\mathcal{F}',X_{1:n},B_1/\sqrt{n}))}\right]\right)\right) \tag{3.24}$$

*holds for every $\delta\in(0,1)$ with probability at least $1-\delta$.*

***Proof.*** As discussed in Remark 3.5, it is easy to see that the hypotheses of Lemma 3.1 hold for $c_{\mathcal{F}^*}=c_{B_1}$ and $C_{\mathcal{F}^*}=C_{B_1}$ in this case.

The inequalities in (3.21) and Barrera (2022, Remark 3.4) for

$$(\mathcal{H}'_{1:n},\mathcal{H}_{1:n})=(diag(\mathcal{F}')_{1:n},diag(\mathcal{F})_{1:n})$$

(see Barrera (2022, eqn. (2.3))) give that for every $\delta>0$

$$\mathcal{R}_{ave}(\rho(\mathcal{F}),(X,Y)_{1:n})$$
$$\leq((2-\alpha)/(1-\alpha))\left(\delta+\sqrt{n}\sup_{f\in\mathcal{F}}||f(X)||_{\mathrm{P},\infty}\mathrm{E}\left[\sqrt{2\log(N_1(\mathcal{F}',X_{1:n},r/n))}\right]\right). \tag{3.25}$$

Taking

$$\delta=B_1\sqrt{n}$$

and using (3.25) we obtain

$$\mathcal{R}_{ave}(\rho(\mathcal{F}),(X,Y)_{1:n})\leq((2-\alpha)/(1-\alpha))B_1\sqrt{n}(1+\mathrm{E}\left[\sqrt{2\log(N_1(\mathcal{F}',X_{1:n},B_1/\sqrt{n}))}\right]).$$

This inequality, when used to estimate the right-hand side of (3.20), gives the right hand side of (3.24). ∎

**Remark 3.12.** *As the proof shows, we obtain the same conclusion if $\mathcal{F}$ and $\mathcal{F}'$ are simply assumed to satisfy*

$$\mathcal{R}_{ave}(\mathcal{F},X_{1:n})\leq\mathcal{R}_{ave}(\mathcal{F}',X_{1:n}),$$

*in particular for $\mathcal{F}\subset(co(\mathcal{F}'))^+$ by a novel application of Talagrand's contraction lemma. Notice also that a slightly bigger upper bound is obtained in place of* (3.24) *(some terms are multiplied by 2) if we replace* (3.23) *by the less restrictive condition*

$$\mathcal{F}'\subset\mathcal{F}\subset cobal(\mathcal{F}')$$

*($cobal(\mathcal{F})$ is defined in* (3.18)*).*

### 3.3 A Rademacher confidence interval for the estimator of $\mathrm{ES}-\mathrm{VaR}$

In what follows, we will focus on the estimator $\hat{r}_{\hat{q}}$ of $r = s - q$ obtained under the following assumption corresponding to the scheme for approximating $r$ in Algorithm 1.

**Assumption 3.13.** *Assume that $\varrho_\varsigma \equiv \varrho^{(B)}$ in (3.4) (see below) is given by the square loss with truncation on the response*

$$\varrho^{(B)}(y, v, z) = (z - T_B((1-\alpha)^{-1}(y-v)^+))^2 \tag{3.26}$$

*for $B > 0$, and that $\mathcal{G}$ is a family of functions $S \to [0, B]$.*

As seen in Remark 2.3, the choice (3.26) corresponds to an approximation scheme (with an additional truncation) for the case $\varsigma(z) = z^2$. We will also consider the family $\varrho_f^{(B)}(\mathcal{G})$ defined (for $f$ fixed) by

$$\varrho_f^{(B)}(\mathcal{G}) := \{(x, y) \mapsto \varrho_f^{(B)}(g)(x, y) := \varrho^{(B)}(y, f(x), g(x)) | g \in \mathcal{G}\}.$$

Let us denote by $r_f$ ($f \in \mathcal{L}_+(S)$) any function satisfying

$$r_f(X) = \mathrm{E}\left[(1-\alpha)^{-1}(Y - f(X))^+ | X\right], \qquad \mathrm{P}\text{-}a.s.,$$

and let $r_f^{(B)} : S \to [0, B]$ be one of its truncated companions, defined by

$$r_f^{(B)}(X) = \mathrm{E}\left[T_B((1-\alpha)^{-1}(Y - f(X))^+) | X\right], \qquad \mathrm{P}\text{-}a.s..$$

For every $(f, g) \in \mathcal{L}(S) \times \mathcal{L}^+(S)$, we will define

$$h_{(f,g)}(X, Y) := \varrho^{(B)}(Y, f(X), g(X)) - \varrho^{(B)}(Y, f(X), r_f^{(B)}(X)),$$

which is the same as the function in Barrera (2022, Section 4 eqn.(4.5)) for the case in consideration.

**Lemma 3.4.** *For every $(f, f', B) \in \mathcal{L}_+(S) \times \mathcal{L}_+(S) \times (0, \infty]$ and every $p \geq 1$, the inequalities*

$$||r_f - r_f^{(B)}||_{\mathrm{P}_X,p} \leq ||((1-\alpha)^{-1}(y-f)^+ - B)^+||_{\mathrm{P}_X,p}$$
$$||r_f^{(B)} - r_{f'}^{(B)}||_{\mathrm{P}_X,p} \leq (1-\alpha)^{-1}||f - f'||_{\mathrm{P}_X,p}$$

*hold (with $r_f^{(\infty)} \equiv r_f$).*

***Proof***. The first inequality is a direct consequence of Jensen's inequality:

$$\mathrm{E}\left[|\mathrm{E}\left[(W - T_B W)|X\right]|^p\right] \leq \mathrm{E}\left[|W - T_B W|^p\right] = \mathrm{E}\left[((|W| - B)^+)^p\right],$$

valid for $p \geq 1$ and any integrable random variable $W$. As for the second, notice first that for every $(a, b, B) \in \mathbb{R} \times \mathbb{R} \times [0, \infty]$,

$$|T_B a - T_B b| \leq |a - b|. \tag{3.27}$$

Combining (3.27) with (3.13) and with Jensen's inequality we get, for every $p \geq 1$:

$$||r_f^{(B)} - r_{f'}^{(B)}||_{\mathrm{P}_X,p}^p = \mathrm{E}\left[|\mathrm{E}\left[T_B((1-\alpha)^{-1}(Y - f(X))^+) - T_B((1-\alpha)^{-1}(Y - f'(X))^+)|X\right]|^p\right]$$

$$\leq \mathrm{E}\left[\mathrm{E}\left[|T_B((1-\alpha)^{-1}(Y - f(X))^+) - T_B((1-\alpha)^{-1}(Y - f'(X))^+)|^p | X\right]\right]$$
$$\leq (1-\alpha)^{-p}||f - f'||_{\mathrm{P}_X,p}^p. \blacksquare$$

**Theorem 3.5.** *Under Assumption 3.13, given $f \in \mathcal{L}(S)$ and given*

$$\mathcal{G}' \subset \mathcal{G} \subset co(\mathcal{G}') \subset \mathcal{L}_+(S)$$

*where $\mathcal{G}$ is a family of functions uniformly bounded by $B$, the inequality*

$$\begin{aligned}||\hat{r}_f - r||_{\mathrm{P}_X,2} \leq & \inf_{g\in\mathcal{G}} ||g - r||_{\mathrm{P}_X,2} \\ & + 2((1-\alpha)^{-1}||f-q||_{\mathrm{P}_X,2} + ||((1-\alpha)^{-1}(y-q)^+ - B)^+||_{\mathrm{P}_{X,Y},2}) \\ & + B\left((2/\sqrt{n})\left(\sqrt{2\log(2/\delta)} + 8\left(1 + \mathrm{E}\left[\sqrt{2\log(N_1(\mathcal{G}', X_{1:n}, B/\sqrt{n}))}\right]\right)\right)\right)^{1/2}.\end{aligned} \tag{3.28}$$

*holds for every $\delta \in (0,1)$, with probability at least $1-\delta$ (and Remark 3.12 also applies).*

***Proof.*** In this proof, $||\cdot||$ denotes either the $L^2_{\mathrm{P}_X}$ seminorm on $\mathcal{L}_+(S)$ or the $L^2_{\mathrm{P}_{X,Y}}$ seminorm on $\mathcal{L}(S \times \mathbb{R})$, the appropriate choice will be always clear (any other norm will be made explicit).

For $f \in \mathcal{F}$, the triangle inequality gives

$$\begin{aligned}||\hat{r}_f - r|| \leq & ||\hat{r}_f - r_f^{(B)}|| + ||r_f^{(B)} - r_q^{(B)}|| + ||r_q^{(B)} - r|| \\ \leq & ||\hat{r}_f - r_f^{(B)}|| + (1-\alpha)^{-1}||f-q|| + ||((1-\alpha)^{-1}(y-q)^+ - B)^+||,\end{aligned} \tag{3.29}$$

by Lemma 3.4.

Now, if $(X', Y')$ is an independent copy of $(X, Y)$, then by the argument leading to Barrera (2022, Section 4, eqn. (4.15))[6]

$$\begin{aligned}& ||\hat{r}_f - r_f^{(B)}||^2 - \inf_{g\in\mathcal{G}} ||g - r_f^{(B)}||^2 \\ = & \mathrm{E}\left[\varrho^{(B)}(Y', f(X'), \hat{r}_f(X')) - \varrho^{(B)}(Y', f(X'), r_f(X'))|D\right] \\ \leq & \sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{k=1}^n (\mathrm{E}\left[\varrho^{(B)}(Y, f(X), g(X))\right] - \varrho^{(B)}(Y_k, f(X_k), g(X_k)))\right\} \\ & + \sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{k=1}^n (\varrho^{(B)}(Y_k, f(X_k), g(X_k)) - \mathrm{E}\left[\varrho^{(B)}(Y, f(X), g(X))\right])\right\}.\end{aligned}$$

We conclude as in the argument for Barrera (2022, eqn. (3.38)) that the inequality

$$\begin{aligned}& ||\hat{r}_f - r_f^{(B)}||^2 - \inf_{g\in\mathcal{G}} ||g - r_f^{(B)}||^2 \\ \leq & (2/\sqrt{n})(\sup_{g\in\mathcal{G}} ||\varrho_f^{(B)}(g)(X,Y)||_{\mathrm{P},\infty}\sqrt{2\log(2/\delta)} + (2/\sqrt{n})\mathcal{R}_{ave}(\varrho_f^{(B)}(\mathcal{G}), (X,Y)_{1:n}))\end{aligned} \tag{3.30}$$

holds for every $\delta \in (0,1)$ with probability at least $1-\delta$. In virtue again of the triangle inequality and the inequality

$$\sqrt{a^2 + b^2} \leq |a| + |b|,$$

[6] Taking $Z'_{1:n} = (X,Y)'_{1:n} \sim (X,Y)$ i.i.d., independent of $(X,Y)_{1:n}$ and $\lambda = 1$.

(3.30) and Lemma 3.4 imply that

$$
\begin{aligned}
||\hat{r}_f - r_f^{(B)}|| \leq & \inf_{g \in \mathcal{G}} ||g - r|| \\
& + (1-\alpha)^{-1} ||f - q|| + ||((1-\alpha)^{-1}(Y - q(X))^+ - B)^+|| \\
& + ((2/\sqrt{n})(\sup_{g \in \mathcal{G}} ||\varrho_f^{(B)}(g)(X,Y)||_{\mathrm{P},\infty} \sqrt{2\log(2/\delta)} + (2/\sqrt{n}) \mathcal{R}_{ave}(\varrho_f^{(B)}(\mathcal{G}), (X,Y)_{1:n})))^{1/2}.
\end{aligned} \tag{3.31}
$$

A combination of (3.29) and (3.31) leads to

$$
\begin{aligned}
||\hat{r}_f - r|| \leq & \inf_{g \in \mathcal{G}} ||g - r|| \\
& + 2((1-\alpha)^{-1} ||f - q|| + ||((1-\alpha)^{-1}(y - q)^+ - B)^+||) \\
& + ((2/\sqrt{n})(\sup_{g \in \mathcal{G}} ||\varrho_f^{(B)}(g)(X,Y)||_{\mathrm{P},\infty} \sqrt{2\log(2/\delta)} + (2/\sqrt{n}) \mathcal{R}_{ave}(\varrho_f^{(B)}(\mathcal{G}), (X,Y)_{1:n})))^{1/2}.
\end{aligned} \tag{3.32}
$$

Let us now upper bound the Rademacher complexity in (3.32). Since the function

$$u \mapsto u^2, \qquad u \in [0, 2B]$$

is Lipschitz with Lipschitz constant $4B$, Talagrand's contraction lemma gives

$$\mathcal{R}_{ave}(\varrho_f^{(B)}(\mathcal{G}), (X,Y)_{1:n}) \leq 4B \mathcal{R}_{ave}(\mathcal{G}, X_{1:n}) = 4B \mathcal{R}_{ave}(\mathcal{G}', X_{1:n}). \tag{3.33}$$

An application of Barrera (2022, (3.47)) together with (3.33) gives the inequality

$$\mathcal{R}_{ave}(\varrho_f^{(B)}(\mathcal{G}), (X,Y)_{1:n}) \leq 4B(r + B\sqrt{n} \mathrm{E}\left[\sqrt{2\log(\mathcal{N}_1(\mathcal{G}', X_{1:n}, r/n))}\right])$$

for every $r > 0$, which in turns implies that (taking $r = B\sqrt{n}$)

$$\mathcal{R}_{ave}(\varrho_f^{(B)}(\mathcal{G}), (X,Y)_{1:n}) \leq 4B^2\sqrt{n} \left(1 + \mathrm{E}\left[\sqrt{2\log(\mathcal{N}_1(\mathcal{G}', X_{1:n}, B/\sqrt{n}))}\right]\right) \tag{3.34}$$

(3.28) follows by a combination of (3.32), (3.34), and the bound

$$\sup_{g \in \mathcal{G}} ||\varrho_f^{(B)}(g)(X,Y)||_{\mathrm{P},\infty} \leq B^2. \blacksquare$$

### 3.4 A Vapnik-Chervonenkis confidence interval for the estimator of $\mathrm{ES} - \mathrm{VaR}$

We will proceed now to prove the following error bound for $||\hat{r}_f - r||_{\mathrm{P}_X,2}$ (see (2.13) and (3.4)):

**Theorem 3.6.** *Under Assumption 3.13, given $f \in \mathcal{L}_+(S)$ and $\mathcal{G} \subset \mathcal{L}_+(S)$, the inequality*

$$
\begin{aligned}
||\hat{r}_f - r||_{\mathrm{P}_X,2} \leq & \sqrt{(6\,\lambda - 5)}\, \inf_{g\in\mathcal{G}} ||g - r||_{\mathrm{P}_X,2} \\
& + (1 + \sqrt{(6\,\lambda - 5)})((1-\alpha)^{-1}||f - q||_{\mathrm{P}_X,2} + ||((1-\alpha)^{-1}(y-q)^+ - B)^+||_{\mathrm{P}_{X,Y},2}) \\
& + (2^7 \cdot 3)^{1/2} B((1/((\lambda-1)n))\,(\log(42) + \log(1/\delta) + \log(\mathrm{E}\,[\mathcal{N}_1(\mathcal{G}, X_{1:n}, B/(24\,n))]))\,)^{1/2}
\end{aligned}
\tag{3.35}
$$

*holds for every $\delta \in (0,1)$ with probability at least $1-\delta$, provided that*

$$
1 < \lambda \leq 13/12. \tag{3.36}
$$

***Proof.*** In this case we depart from the estimate (3.29) and we then estimate $||\hat{r}_f - r_f^{(B)}||$ via Barrera (2022, Theorem 4.2) , which depends on the functions

$$
\begin{aligned}
A :& S^n \times (1,\infty) \times (1,\infty) \times (0,\infty) \to (0,\infty] \\
a :& (1,\infty) \times (1,\infty) \times (0,\infty) \to (0,\infty] \\
\epsilon_n :& (1,\infty) \times (1,\infty) \to (0,\infty) \\
b :& (1,\infty) \times (1,\infty) \to (0,\infty)
\end{aligned}
$$

given by

$$
\begin{aligned}
A(x_{1:n}, c, \lambda, \epsilon) :=& 2(c+1)(2c+3)\mathcal{N}_1\left(T_B\mathcal{G}, x_{1:n}, \frac{1}{2^5}\frac{1}{B}\frac{1}{\lambda(c-1)+1}(1-\frac{1}{c})\epsilon\right), \\
a(c, \lambda, \epsilon) :=& \mathrm{E}\,[A(X_{1:n}, c, \lambda, \epsilon)]\,, \\
\epsilon_n(c, \lambda) :=& 8B^2(-(\lambda-1) + \sqrt{(\lambda-1)^2 + c(c+1)\lambda^2/n}) \\
b(c, \lambda) :=& \frac{1}{2^5 B^2}\frac{1}{(\frac{1}{3}(1-\frac{1}{c})(1-\frac{1}{\lambda}) + (2\lambda-1))^2}(1-\frac{1}{c})^3(1-\frac{1}{\lambda}).
\end{aligned}
$$

Indeed we arrive, by an argument as the one leading to (3.32), at the estimate

$$
\begin{aligned}
||\hat{r}_f - r|| \leq & \sqrt{(6\,\lambda - 5)}\, \inf_{g\in\mathcal{G}} ||g - r|| \\
& + (1 + \sqrt{(6\,\lambda - 5)})((1-\alpha)^{-1}||f - q|| + ||((1-\alpha)^{-1}(y-q)^+ - B)^+||) \\
& + \left(6\left(\epsilon_n(c,\lambda) \vee (\frac{1}{nb(c,\lambda)}(\log a(c, \lambda, \epsilon_n(c,\lambda)) + \log(2/\delta)))\right)\right)^{1/2}
\end{aligned}
\tag{3.37}
$$

with probability at least $1-\delta$. Restricting $\lambda$ to the range (3.36) and using the analysis leading to Barrera (2022, (4.41)), we deduce from (3.37) that

$$\begin{aligned}||\hat{r}_f - r|| \leq & \sqrt{(6\,\lambda - 5)}\inf_{g\in\mathcal{G}}||g-r|| \\ & +(1+\sqrt{(6\,\lambda-5)})((1-\alpha)^{-1}||f-q||+||((1-\alpha)^{-1}(y-q)^+ - B)^+||) \\ & +(2^7\cdot 3)^{1/2}B((1/((\lambda-1)n))\,(\log(42)+\log(1/\delta)+\log\left(\mathrm{E}\left[\mathcal{N}_1\left(\mathcal{G}, X_{1:n}, B/(24\,n)\right)\right]\right))^{1/2}\end{aligned}$$

holds with probability at least $1-\delta$. ∎

**Rademacher vs VC: from "small" to "big" data** To give a crude comparison between Theorems 3.5 and 3.6, first note that, since $\sqrt{6\lambda-5}\approx 1$ (under (3.36)), it is reasonable to limit the discussion to a comparison between the terms in the third line of the inequalities (3.28) and (3.35).

The ratio between these two terms is lower bounded (crudely) by

$$(2^3\cdot 3)^{-1/2}((\lambda-1)\sqrt{n})^{1/2}(\log\left(42\mathrm{E}\left[\mathcal{N}_1\left(\mathcal{G}, X_{1:n}, B/(24\,n)\right)\right]/\delta\right))^{-1/2},$$

which shows that (3.28) is worse (bigger) than (3.35) provided that

$$\begin{aligned}\sqrt{n} \geq & \frac{2^3\cdot 3}{\lambda-1}\log\left(42\mathrm{E}\left[\mathcal{N}_1\left(\mathcal{G}, X_{1:n}, B/(24\,n)\right)\right]/\delta\right) \\ \geq & 2^5\cdot 3^2(\log(42)+\log(1/\delta)),\end{aligned} \tag{3.38}$$

where in the last inequality we used the upper bound for $\lambda$ in (3.36).

The first inequality in (3.38) is an exact (but crude) criterion on the sample size indicating an interval where (3.35) is preferable to (3.28). The inequality between the first and the third terms in (3.38) can be understood as an "heuristic" criterion for this preference, indicating in particular the heuristic boundary

$$n \geq (2^5\cdot 3^2\cdot\log(42))^2$$

between "small-medium" and "big" data, where we pass from the Rademacher to the VC regime.

### 3.5 A Posteriori Monte Carlo Validation of VaR and ES learners

Assuming one has access to the generative process of the data, as it is the case in most quantitative finance problems, one can in fact estimate distances of any guesses to the groundtruth (conditional) VaR and ES without directly computing the latter, using a simple twin-simulation trick.

**Proposition 3.7.** *Let $\check{q}$ and $\check{s}$ be two Borel functions of $x$ (tentative approximations $\check{q}(X)$ and $\check{s}(X)$ of $q(X)=\mathrm{VaR}(Y|X)$ and $s(X)=\mathrm{ES}(Y|X)$ at the confidence level $\alpha$). Introducing two conditionally independent copies[7] $Y^{(1)}$ and $Y^{(2)}$ of $Y$ given $X$ and*

[7] i.e. for any bounded Borel functions $\varphi$ and $\psi$, we have $\mathrm{E}[\varphi(Y^{(1)})|X]=\mathrm{E}[\varphi(Y^{(2)})|X]=\mathrm{E}[\varphi(Y)|X]$ and $\mathrm{E}[\varphi(Y^{(1)})\psi(Y^{(2)})|X]=\mathrm{E}[\varphi(Y^{(1)})|X]\mathrm{E}[\psi(Y^{(2)})|X]$.

*denoting* $Y^{(1)} \wedge Y^{(2)} = \min\{Y^{(1)}, Y^{(2)}\}$*, we have*

$$\begin{aligned} &\|\mathrm{P}[Y \geq \check{q}(X)|X] - 1 + \alpha\|_{\mathrm{P},2} = \\ &\qquad \sqrt{(1-\alpha)(1-\alpha-2\mathrm{P}(Y > \check{q}(X))) + \mathrm{P}[Y^{(1)} \wedge Y^{(2)} > \check{q}(X)]}, \end{aligned} \tag{3.39}$$

$$\|\check{s}(X) - s(X)\|_{\mathrm{P},2} = \|\check{s}(X) - \check{q}(X) - \mathrm{E}[(1-\alpha)^{-1}(Y - \check{q}(X))^+|X]\|_{\mathrm{P},2} + \epsilon, \tag{3.40}$$

*where*

$$\begin{aligned} \|\check{s}(X) - \check{q}(X) - \mathrm{E}[(1-\alpha)^{-1}(Y - \check{q}(X))^+|X]\|_{\mathrm{P},2}^2 &= \|\check{s}(X) - \check{q}(X)\|_{\mathrm{P},2}^2 \\ &+ \frac{1}{(1-\alpha)^2}\mathrm{E}[(Y^{(1)} - \check{q}(X))^+(Y^{(2)} - \check{q}(X))^+] \\ &- \frac{2}{1-\alpha}\mathrm{E}[(\check{s}(X) - \check{q}(X))(Y - \check{q}(X))^+] \end{aligned} \tag{3.41}$$

*and, assuming that* $F'_{Y|X}(Y) \geq c$ *holds* P*-a.s for some* $c > 0$*,*

$$0 \leq \epsilon \leq \frac{1}{c}(1 + (1-\alpha)^{-1})\|\mathrm{P}[Y \geq \check{q}(X)|X] - 1 + \alpha\|_{\mathrm{P},2}, \tag{3.42}$$

*which is in turn given by* (3.39)*.*

***Proof.*** We have

$$\|\mathrm{P}[Y \geq \check{q}(X)|X] - 1 + \alpha\|_{\mathrm{P},2} = \sqrt{\mathrm{E}[\mathrm{P}[Y \geq \check{q}(X)|X]^2] + (1-\alpha)^2 - 2(1-\alpha)\mathrm{P}[Y \geq \check{q}(X)]},$$

where

$$\mathrm{E}[\mathrm{P}[Y \geq \check{q}(X)|X]^2] = \mathrm{E}[\mathrm{P}[Y^{(1)} \geq \check{q}(X)|X]\mathrm{P}[Y^{(2)} \geq \check{q}(X)|X]] = \mathrm{P}[Y^{(1)} \wedge Y^{(2)} \geq \check{q}(X)].$$

Thus (3.39) follows. For the ES, note that with $\rho(y, v) = (1-\alpha)^{-1}(y - v)^+ + v$ so that $\mathrm{E}[\rho(Y, q(X))|X] = s(X)$:

$$\|\check{s}(X) - s(X)\|_{\mathrm{P},2}^2 = \|\mathrm{E}[Z|X]\|_{\mathrm{P},2}^2,$$

where $Z := \check{s}(X) - \rho(Y, q(X))$ satisfies by the conditional Jensen inequality:

$$\|\mathrm{E}[Z|X]\|_{\mathrm{P},2}^2 = \mathrm{E}[(\mathrm{E}[Z|X])^2] \leq \mathrm{EE}[Z^2|X]) = \mathrm{E}[Z^2] = \|\check{s}(X) - \rho(Y, q(X))\|_{\mathrm{P},2}^2.$$

An application of the triangular inequality yields

$$\|\check{s}(X) - \rho(Y, q(X))\|_{\mathrm{P},2} \leq \|\check{s}(X) - \rho(Y, \check{q}(X))\|_{\mathrm{P},2} + \|\rho(Y, q(X)) - \rho(Y, \check{q}(X))\|_{\mathrm{P},2}.$$

One then uses the $(1 + (1-\alpha)^{-1})$-Lipschitz regularity of $\rho$ with respect to its second argument and the assumed $\frac{1}{c}$-Lipschitz regularity of $F_{Y|X}^{-1}$ to deduce (3.40) from the above, for $\epsilon$ satisfying (3.42). Using the twin-simulation trick again, we get (3.41). ∎

The expectations and probabilities in (3.39) and (3.41) can then be estimated via a simply dedoubled (twin) Monte Carlo simulation (see Algorithm 2), as opposed to a

plain nested Monte Carlo that would be required to explicitly attempt to approximate conditional expectations. Moreover the accuracy of the twin Monte-Carlo estimates can be controlled by computing confidence intervals. Noting that $1-\alpha = \mathrm{P}[Y \geq q(X)|X]$ holds almost P surely, the distance in (3.39) can be interpreted as a distance in $p$-values between the quantile estimate $\check{q}(X)$ and the true quantile $q(X)$, as opposed to a distance directly between values of conditional quantile estimators. If the approximation $\check{q}$ is sufficiently good, i.e. if this distance is sufficiently small (as compared to $1-\alpha$), then (3.41) can be used as a proxy for $\|\check{s}(X) - s(X)\|^2_{\mathrm{P},2}$: see Algorithm 2. Note however that, because of the $(1-\alpha)^{-1}$ factor in (3.42), the inequality in (3.42) becomes crude when $\alpha$ gets close to 1.

**name :** TwinVal
**input :** out-of-sample $\{(X_i, Y_i^{(1)}, Y_i^{(2)})\}_{i=1}^n$ with $Y_i^{(1)}, Y_i^{(2)}$ independent copies of $Y$ given $X = X_i$, a confidence level $\alpha$, corresponding estimates $\check{q}$ and $\check{s}$ of $q$ and $s$, tolerance levels $\delta^{\mathrm{var}}$ and $\delta^{\mathrm{es}}$
**output:** Quality of $\check{q}$ and $\check{s}$

Compute $(\epsilon^{\mathrm{var}})^2 = \frac{1}{n}\sum_{i=1}^n \left((1-\alpha)(1-\alpha-2\mathbb{1}_{Y_i^{(1)} > \check{q}(X_i)}) + \mathbb{1}_{Y^{(1)} \wedge Y_i^{(2)} > \check{q}(X_i)}\right)$
**if** $\epsilon^{\mathrm{var}} > \delta^{\mathrm{var}}$ **then**
  Reply already $\check{q}$ is bad
**else**
  Compute $(\epsilon^{\mathrm{es}})^2 = \frac{1}{n}\sum_{i=1}^n \Big[(\check{s}(X_i) - \check{q}(X_i))^2 + \frac{1}{(1-\alpha)^2}(Y_i^{(1)} - \check{q}(X_i))^+(Y_i^{(2)} - \check{q}(X_i))^+ - \frac{2}{1-\alpha}(\check{s}(X_i) - \check{q}(X_i))(Y_i^{(1)} - \check{q}(X_i))^+]\Big]$
  **if** $\epsilon^{\mathrm{es}} > \delta^{\mathrm{es}}$ **then**
    Reply $\check{q}$ is good but $\check{s}$ is bad
  **else**
    Reply $\check{q}$ and $\check{s}$ are good
  **end**
**end**

**Algorithm 2:** Twin Monte Carlo validation for VaR and ES.

In the case where the twin Monte Carlo estimates for the right-hand-sides in (3.39) and (3.41), after having been confirmed to be accurate by drawing enough samples, are not good enough, one can improve the numerical optimization, in first attempt, and then act on the hypothesis space. For instance, in the case of the next section of the paper where hypothesis spaces of neural networks are used, one can improve the corresponding stochastic gradient descent (e.g. switching from Algorithm 3 to Algorithm 8) by changing the optimizer and/or its hyperparameters, in first attempt, and then try to train with more layers/units or better architectures.

**name :** SGDOpt
**input :** $\{(X_i, Y_i)\}_{i=1}^n$, a partition $B$ of $\{1 \dots n\}$, a number of epochs $E \in \mathbb{N}^\star$, a learning rate $\eta > 0$, initial weight (matrix) $\widehat{W}$ and bias (vector) $\widehat{b}$ parameters, and a loss function $\rho = \rho(W, b, \text{batch})$
**output:** Trained parameters $\widehat{W}$ and $\widehat{b}$

```
for epoch = 1, ..., E do                                  // loop over epochs
    for batch ∈ B do                                      // loop over batches
        Ŵ ← Ŵ − η∇_W ρ(Ŵ, b̂, batch)
        b̂ ← b̂ − η∇_b ρ(Ŵ, b̂, batch)
    end
end
```

**Algorithm 3:** Stochastic gradient descent in a neural net hypothesis space.

# 4 Learning Using Neural Networks

## 4.1 Error bound of the single-$\alpha$ learning algorithm with one-layer neural networks

We apply the previous developments to the estimation of errors from Algorithm 1 when one-hidden-layer neural networks with bounded weights are used to define the hypothesis spaces. We consider the following families of functions:

**Definition 4.1.** *Let $\sigma : \mathbb{R} \to [0, 1]$ be a nondecreasing measurable function that is applied element-wise when supplied with a vector as input and let $(d, M, B) \in \mathbb{N} \times \mathbb{N} \times (0, \infty)$. Denote by $\tilde{\mathcal{F}}(d, B, m, \sigma) \subset \mathcal{L}_{\mathbb{R}^d}$ the family of neural networks on $S = \mathbb{R}^d$ with $m$ (or less) units, one hidden layer, activation function $\sigma$ and Lasso regularization bound $B$, defined as follows*

$$\tilde{\mathcal{F}}(d, B, m, \sigma) = \{\mathbb{R}^d \ni x \mapsto c_0 + \sum_{k=1}^{m} c_k \sigma(a_k \cdot x + b_k) \in \mathbb{R} \mid$$
$$(a_{1:m}, b_{1:m}) \in (\mathbb{R}^d)^m \times \mathbb{R}^m, c_{0:m} \in \mathbb{R}^{m+1} \text{ with } \sum_{k=0}^{m} |c_k| \leq B\}.$$

It is clear that $\tilde{\mathcal{F}}(d, B, m, \sigma)$ is totally bounded by $B$. Notice also that for all $m \in \mathbb{N}^\star$

$$\tilde{\mathcal{F}}(d, B, 1, \sigma) \subset \tilde{\mathcal{F}}(d, B, m, \sigma) \subset co(\tilde{\mathcal{F}}(d, B, 1, \sigma)) = cobal(\tilde{\mathcal{F}}(d, B, 1, \sigma)),$$

where $co(\cdot)$ and $cobal(\cdot)$ are defined in (3.17) and (3.18).

We have from Barrera (2022, Example 3.2) for all $0 < r < \frac{B}{2}$:

$$\log(\mathcal{N}_1(\tilde{\mathcal{F}}(d, B, m, \sigma), X_{1:n}, r)) \leq ((2d + 5)m + 1)(1 + \log(12) + \log(B/r) + \log(m + 1))$$

This estimate can be combined with Theorem 3.3 to give an error estimate for Algorithm 1. In the context of this algorithm, we simplify the notation by writing

$$Y_{h_k}(\omega) = h_k(X(\omega), Y(\omega)), \qquad q_{h_k}(x) = h_k(x, q(x)) \qquad (k = 1, 2),$$
$$r_{h_2}(x) = h_2(x, r(x)) \tag{4.1}$$

where $q$ and $r = s - q$ are defined as in (2.4).

**Theorem 4.1.** *With the notation of Algorithm 1 and in* (4.1)*, and for* $\tilde{\mathcal{F}} = \tilde{\mathcal{F}}(d, B_1, m, \sigma)$*, if* $Y_{h_1}$ *satisfies Assumption 3.11, then the inequality*

$$
\begin{aligned}
c_{B_1}||\hat{f} - q_{h_1}||^2_{\mathrm{P}_X,2} \leq & \left(2(2-\alpha)\inf_{f\in\tilde{\mathcal{F}}}||f - q_{h_1}||_{\mathrm{P}_X,1}\right) \wedge \left(C_{B_1}\inf_{f\in\tilde{\mathcal{F}}}||f - q_{h_1}||^2_{\mathrm{P}_X,2}\right) \\
& + \frac{4(2-\alpha)}{\sqrt{n}}\left(B_2\sqrt{2\log\left(\frac{2}{\delta}\right)}\right. \\
& \left. +2B_1\left(1 + \sqrt{2((2d+5)m+1)(1+\log(12(m+1)\sqrt{n}))}\right)\right)
\end{aligned}
$$

*holds for every* $\delta \in (0,1)$ *with probability at least* $1-\delta$*.*

**Remark 4.2.** *The discussion in Padilla, Tansey, and Chen (2020) implies that the rates following from these bounds cannot be improved in general, but as proved in (Chen, 2007, Example 3.2.2), the dimension of the feature space can play a role in a variety of examples.*

Analogous reasoning, using this time Theorems 3.5 and 3.6 and the observations in Remark 3.12, lead to the following bound on the error of $\hat{g}$ in Algorithm 1:

**Theorem 4.2.** *With the notation of Algorithm 1 and in* (4.1)*, for*[8] $\mathcal{G} = (\tilde{\mathcal{F}}(d, B, m, \sigma))^+$*, the inequality*

$$
\begin{aligned}
&||\hat{g} - r_{h_2}||_{\mathrm{P}_X,2} \leq \sqrt{(6\,\lambda - 5)}\inf_{g\in\mathcal{G}}||g - r_{h_2}||_{\mathrm{P}_X,2} \\
&+ (1+\sqrt{(6\,\lambda-5)})((1-\alpha)^{-1}||f - q_{h_2}||_{\mathrm{P}_X,2} + ||((1-\alpha)^{-1}(y_{h_2} - q_{h_2})^+ - B)^+||_{\mathrm{P}_{X,Y},2}) \\
&+ B\left(\sqrt{2/\sqrt{n}}\right)\times \\
&\quad\left(2^4\sqrt{(\log(2\cdot 3\cdot 7/\delta) + ((2d+5)m+1)(1+\log(2^5\cdot 3^2(m+1)n))/((\lambda-1)\sqrt{n})}\right. \\
&\quad\left. \wedge \sqrt{\sqrt{2\log(2/\delta)} + 2^3\left(1 + \sqrt{2(d+3)(1+\log(2^3\cdot 3\sqrt{n}))}\right)}\right)
\end{aligned}
$$

*holds with probability at least* $1-\delta$*, for every* $1 < \lambda \leq 13/12$ *and every* $f \in \mathcal{F}$*.*

More generally, we consider feed-forward neural networks with more than one layer in what follows. We define $\mathcal{F}^{d,o,m,n}$, to be the set of functions of the form $\mathbb{R}^d \ni x \mapsto \zeta^{d,o}_{l+1}(x, W, b) \in \mathbb{R}^o$, where:

$$
\begin{aligned}
\zeta^{d,o}_0(x, W, b) &= x \\
\zeta^{d,o}_i(x, W, b) &= \sigma(W_i\zeta^{d,o}_{i-1}(x, W, b) + b_i), \forall i \in \{1, \ldots, l\} \\
\zeta^{d,o}_{l+1}(x, W, b) &= W_{l+1}\zeta^{d,o}_l(x, W, b) + b_{l+1}
\end{aligned}
$$

[8] with $(\mathcal{H})^+ = \{(h)^+ : h \in \mathcal{H}\}$ for any set of functions $\mathcal{H}$.

and $W_1 \in \mathbb{R}^{m\times d}, W_2 \ldots, W_n \in \mathbb{R}^{m\times m}, W_{l+1} \in \mathbb{R}^{o\times m}$, $b_1, \ldots, b_l \in \mathbb{R}^m, b_{l+1} \in \mathbb{R}^o$. The function $\sigma$ is called an activation function. We also choose the Softplus activation function, i.e. $\sigma(x) = \log(1 + \exp(x))$.

In what follows, we assume a finite i.i.d sample of $(X, Y)$ given by $D_n \coloneqq \{(X_i, Y_i)\}_{1\le i\le n}$.

### 4.2 Learning the VaR

In this part, the goal is to find an approximation of $q(X) = \mathrm{VaR}(Y|X)$, at the confidence level $\alpha$, as a function of $X$, represented by a neural network from $\mathcal{F}^{d,1,m,l}$, for given $m$ and $l$. More precisely, we aim to solve the following optimization problem (cf. (2.9) and (2.5)):

$$\tilde{q} \in \arg\min_{q\in\mathcal{F}^{d,1,m,l}} \mathrm{E}[(Y - q(X))^+ + (1-\alpha)q(X)]$$

or, equivalently, find weights

$$(\tilde{W}^{\mathrm{var}}, \tilde{b}^{\mathrm{var}}) \in \arg\min_{W,b} \mathrm{E}[(Y - \zeta_{l+1}^{d,1}(X, W, b))^+ + (1-\alpha)\zeta_{l+1}^{d,1}(X, W, b)]. \tag{4.2}$$

Problem (4.2) is then solved numerically by applying a stochastic gradient descent (Algorithm 3 or an accelerated version of it, noting that the gradients there are quickly and exactly computed by automatic differentiation) to a finite-sample formulation of the problem (cf. step 3 in Algorithm 1):

$$(\widehat{W}^{\mathrm{var}}, \widehat{b}^{\mathrm{var}}) \in \arg\min_{W,b} \frac{1}{n}\sum_{i=1}^{n}[(Y_i - \zeta_{l+1}^{d,1}(X_i, W, b))^+ + (1-\alpha)\zeta_{l+1}^{d,1}(X_i, W, b)]. \tag{4.3}$$

This specification of Algorithm 1 regarding the VaR (see the step 3 there) is detailed in Algorithm 4 (the corresponding treatment of ES is deferred to Section 4.3). Once (4.3) has been solved numerically (a procedure to which we will refer to as training in what follows), we obtain an approximation of $\mathrm{VaR}(Y|X)$, at the confidence level $\alpha$, given by $\widehat{q}(X)$, where

$$\widehat{q}(x) \coloneqq \zeta_{l+1}^{d,1}(x, \widehat{W}^{\mathrm{var}}, \widehat{b}^{\mathrm{var}}),\ x \in \mathbb{R}^d$$

(see Algorithm 4).

**name :** VaRAlg
**input :** $\{(X_i, Y_i)\}_{i=1}^n$, a partition $B$ of $\{1 \ldots n\}$, a quantile level $\alpha$, a number of epochs $E \in \mathbb{N}^\star$, a learning rate $\eta > 0$, initial values for the network parameters $\widehat{W}$ and $\widehat{b}$, and neural network output function $\zeta_{l+1}^{d,1}(x, W, b)$
**output:** Trained parameters of VaR network

1 define $\rho^{\mathrm{var}}(W, b, \mathrm{batch}) = \dfrac{1}{|\mathrm{batch}|}\displaystyle\sum_{i\in\mathrm{batch}} [(Y_i - \zeta_{l+1}^{d,1}(X_i, W, b))^+ + (1-\alpha)\zeta_{l+1}^{d,1}(X_i, W, b)]$

2 $(\widehat{W}^{\mathrm{var}}, \widehat{b}^{\mathrm{var}}) \leftarrow \mathrm{SGDOpt}(\{(X_i, Y_i)\}_{i=1}^n, B, E, \eta, \widehat{W}, \widehat{b}, \rho^{\mathrm{var}})$

**Algorithm 4:** Neural network regression for learning the VaR.

Given that the training is done for a single fixed confidence level $\alpha$, we refer to this approach as the *single-$\alpha$ learning* (or single-$\alpha$ for brevity in the numerics). Under this approach, if one is interested in finding the conditional VaR for another confidence level, one has to repeat the training procedure using the new confidence level in the learning problem (4.3).

### 4.3 Learning the ES using a two-step approach

Our next aim is to find an approximation of the ES$(Y|X)$, at the confidence level $\alpha$, as a function of $X$ that is represented by a neural network from $\mathcal{F}^{d,1,m,l}$, for given $m$, $n$. Assuming a representation, or approximation, $\check{q}$ of the VaR of $Y$ given $X$ at the confidence level $\alpha$, which we will call VaR candidate, the goal is to solve the following problem (cf. (2.12)):

$$\tilde{s} \in \arg \min_{s \in \mathcal{F}^{d,1,m,l}} \mathrm{E}[((1-\alpha)^{-1}(Y - \check{q}(X))^+ + \check{q}(X) - s(X))^2]$$

for which we can write a finite-sample version in parameter space as follows (cf. the step 4 in Algorithm 1):

$$(\widehat{W}^{\mathrm{es}}, \widehat{b}^{\mathrm{es}}) \in \arg \min_{W,b} \frac{1}{n} \sum_{i=1}^{n} [((1-\alpha)^{-1}(Y_i - \check{q}(X_i))^+ + \check{q}(X_i) - \zeta_{l+1}^{d,1}(X_i, W, b))^2]. \quad (4.4)$$

This specification of Algorithm 1 regarding the ES (see the step 4 there) is detailed in the second part of Algorithm 5.

Alternatively, using a transfer learning trick, one can deduce an ES approximation very quickly using a VaR candidate that is in neural network form. Namely, one can look for an ES approximator using a neural network with the same architecture as the one used for the VaR, set the weights of all hidden layers to those of the VaR network and then freeze them. The training of the ES approximator then falls down to a linear regression to determine the weights of the output layer, as detailed in the first part of Algorithm 5. We show in Section 6 that such a scheme is enough to obtain good approximations, while also being very fast (a fraction of a second in our experiments) if one uses highly optimized linear algebra routines such as the ones implemented by cuBLAS for Nvidia GPUs.

In either case, the ensuing estimate of $s$ is

$$\widehat{s}(x) := \zeta_{l+1}^{d,1}(x, \widehat{W}^{\mathrm{es}}, \widehat{b}^{\mathrm{es}}), \ x \in \mathbb{R}^d$$

(see Algorithm 5).

**name :** ESAlg
**input :** $\{(X_i, Y_i)\}_{i=1}^n$, a partition $B$ of $\{1 \dots n\}$, a quantile level $\alpha$, a number of epochs $E \in \mathbb{N}^\star$, a learning rate $\eta > 0$, initial values for the network parameters $\widehat{W}$ and $\widehat{b}$ and neural network output function $\zeta_{l+1}^{d,1}(X_i, W, b)$
**output:** Trained parameters of ES network $\widehat{W}^{\text{es}}$ and $\widehat{b}^{\text{es}}$

// Learn the corresponding VaR
$\widehat{W}^{\text{var}}, \widehat{b}^{\text{var}} \leftarrow \text{VaRAlg}(\{(X_i, Y_i)\}_{i=1}^n, B, \alpha, E, \eta, \widehat{W}, \widehat{b})$
**if** *linear regression* **then**
  // Remind $(W_i, b_i)$ denote the weight and bias of $i$-th layer

$$(\{\widehat{W}_i^{\text{es}}\}_{i=1}^l, \{\widehat{b}_i^{\text{es}}\}_{i=1}^l) \leftarrow (\{\widehat{W}_i^{\text{var}}\}_{i=1}^l, \{\widehat{b}_i^{\text{var}}\}_{i=1}^l)$$

$$(\widehat{W}_{l+1}^{\text{es}}, \widehat{b}_{l+1}^{\text{es}}) \leftarrow \text{argmin}_{W_{l+1}, b_{l+1}} \sum_{i=1}^n \Big[ (1-\alpha)^{-1} (Y_i - \zeta_{l+1}^{d,1}(X_i, \widehat{W}^{\text{var}}, \widehat{b}^{\text{var}}))^+ + \zeta_{l+1}^{d,1}(X_i, \widehat{W}^{\text{var}}, \widehat{b}^{\text{var}}) - \zeta_{l+1}^{d,1}\Big(X_i, (\{\widehat{W}_i^{\text{var}}\}_{i=1}^l, W_{l+1}), (\{\widehat{b}_i^{\text{var}}\}_{i=1}^l, b_{l+1})\Big)\Big]^2$$

**else**
  define $\rho^{\text{es}}(W, b, \text{batch}) = \frac{1}{|\text{batch}|} \sum_{i \in \text{batch}} [((1-\alpha)^{-1}(Y_i - \zeta_{l+1}^{d,1}(X_i, \widehat{W}^{\text{var}}, \widehat{b}^{\text{var}}))^+ + \zeta_{l+1}^{d,1}(X_i, \widehat{W}^{\text{var}}, \widehat{b}^{\text{var}}) - \zeta_{l+1}^{d,1}(X_i, W, b))^2]$
  $(\widehat{W}^{\text{es}}, \widehat{b}^{\text{es}}) \leftarrow \text{SGDOpt}(\{(X_i, Y_i)\}_{i=1}^n, B, E, \eta, \widehat{W}^{\text{var}}, \widehat{b}^{\text{var}}, \rho^{\text{es}})$
**end**

**Algorithm 5:** Neural network regressions for learning the ES in two steps.

# 5 Multi-$\alpha$ learning for VaR

In this part we are interested in learning $\text{VaR}(Y|X)$ for multiple confidence levels $\alpha \in (0, 1)$ using a single empirical error minimization. This can help give insights into the sensitivity of $\text{VaR}(Y|X)$ with respect to the confidence level, or into the full distribution of the law of $Y$ given $X$ (e.g. approximated by a histogram representation).

Although one could also formulate multi-$\alpha$ learning versions for ES à la Section 5.3, we have found numerically that it significantly degrades the learning and thus we stick to the VaR in what follows. However, for multi-$\alpha$ ES, the transfer learning trick of Section 4.3 is still a valuable alternative, whether it is done $\alpha$ by $\alpha$, as each run of it is very fast, or globally across $\alpha$'s based on either of the multi-$\alpha$ VaR approaches below.

## 5.1 Related literature

The simultaneous learning of conditional quantiles for multiple confidence levels and the problem of quantile crossing, i.e. the violation of the monotonicity with respect to the confidence level, are early addressed in He (1997); Koenker (2004); Takeuchi, Le, Sears, and Smola (2006). We refer the reader to Moon, Jeon, Lee, and Kim (2021) for a review of more recent references. To deal with the quantile crossing problem, two strategies for constraints can be considered.

The first strategy is to use hypothesis spaces of functions nondecreasing with respect to the confidence level. Meinshausen and Ridgeway (2006) introduce quantile regression forests. In this model the predicted quantile of a new point is based on the

empirical percentile of the group (i.e. the terminal leaf of each tree) where this point belongs, hence, the monotonicity of the quantile estimates is satisfied by construction. Regarding neural networks, Hatalis, Lamadrid, Scheinberg, and Kishore (2017) propose a specific initialization scheme for the weights of the output layer, which does not prevent quantile crossings, but appears to reduce them significantly in their experiments. Cannon (2018) considers the confidence level as an additional explanatory variable and then explores a network such that the estimate is monotone with a defined covariate (confidence level), imposing the non-crossing. Gasthaus, Benidis, Wang, Rangapuram, Salinas, Flunkert, and Januschowski (2019) and Padilla, Tansey, and Chen (2020) use a (deep) network with multiple outputs, constrained by design to be positive, which are expected to approximate quantile increments. The latter resembles our multi-$\alpha$(III) approach in Section 12, especially when the increments are constrained to be positive. Under our multi-$\alpha$(III) approach, however, we sample the confidence level uniformly on a given interval and we further interpolate linearly with respect to the confidence level before insertion of the output of the neural network in the training loss (cf. (5.2)-(5.3)), in order to have a conditional quantile function that is valid for all quantile levels in the interval.

The second strategy is to consider explicitly the non-crossing constraints during the learning phase of the model in form of either hard constraints (that the model must strictly satisfy) or soft constraints (i.e. penalization). Once the non-crossing hard constraints are employed, the model is usually learned using primal-dual optimization algorithms. The latter are applicable in a wide class of models, e.g. support vector regression (Takeuchi, Le, Sears, and Smola, 2006; Sangnier, Fercoq, and d'Alché Buc, 2016) and spline regression (Bondell, Reich, and Wang, 2010), but notably not in the case of the family of (deep) neural networks, because of the computational cost and the poor scalability of projected gradient descent. Therefore, the non-crossing constraints are more preferably embedded in the training of neural networks via a penalty term, based in (Moon, Jeon, Lee, and Kim, 2021) on a finite difference of the output of the neural network (that approximates the value-at-risk) for two confidence levels. In Section 5.3 we use a similar penalization strategy, where, instead of penalizing the negative part of a finite difference, we penalize the negative part of the partial derivative of the network with respect to the confidence level. The partial derivative gives more information about the local behavior around training points and we can penalize its negative part at every $\alpha$ that appears at the training stage, e.g. for several thousands values of $\alpha$ in our numerics below, as opposed to penalizing negative increments at a few fixed values of $\alpha$ in (Moon, Jeon, Lee, and Kim, 2021). Our approach also spares one hyperparameter, namely the size of the discrete increment in confidence levels used for the finite differences.

### 5.2 Extension of the bounds to multi-$\alpha$ learning

The various proofs and bounds presented in this paper for a fixed $\alpha \in [0,1]$ can be extended to the multi-$\alpha$ learning framework where $\alpha$ is now a random variable supported on $I = [\underline{\alpha}, \overline{\alpha}] \subset (0,1)$ (with Lebesgue sigma-algebra $\mathcal{I}$) treated as a covariate alongside $X$: see Table 1 for the changes that need to be done in order to have similar results in this new framework. The implementation of this approach using neural networks is

| **Single-$\alpha$** | **Multi-$\alpha$** |
|---|---|
| $D = \{(X_j, Y_j)\}_{j=1}^n$ is an i.i.d sample of $(X, Y)$ | $D = \{(\alpha_j, X_j, Y_j)\}_{j=1}^n$ is an i.i.d sample of $(\alpha, X, Y)$ |
| $\mathcal{S} \otimes \mathcal{R}$ | $\mathcal{I} \otimes \mathcal{S} \otimes \mathcal{R}$ |
| $\rho(y, v) = (1-\alpha)^{-1}(y-v)^+ + v$ | $\rho(\alpha, y, v) = (1-\alpha)^{-1}(y-v)^+ + v$ |
| $\tilde{\rho}(f) = \mathrm{E}[\rho(Y, f(X))]$ | $\tilde{\rho}(f) = \mathrm{E}[\rho(\alpha, Y, f(\alpha, X))]$ |
| $\mathcal{F} \subset \ell(S)$ | $\mathcal{F} \subset \ell([\underline{\alpha}, \overline{\alpha}] \times S)$ |
| $\tilde{q} \in \operatorname{argmin}_{f \in \mathcal{F}} \mathrm{E}[\rho(Y, f(X))]$ | $\tilde{q} \in \operatorname{argmin}_{f \in \mathcal{F}} \mathrm{E}[\rho(\alpha, Y, f(\alpha, X))]$ |
| $\hat{q} \in \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{k=1}^n \rho(Y_k, f(X_k))$ | $\hat{q} \in \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{k=1}^n \rho(\alpha_k, Y_k, f(\alpha_k, X_k))$ |
| $\tilde{\rho}(\tilde{q}) - \tilde{\rho}(q) \geq \frac{c_{\mathcal{F}^*}}{2(1-\alpha)} \lVert \tilde{q} - q \rVert^2_{\mathrm{P}_X, 2}$ | $\tilde{\rho}(\tilde{q}) - \tilde{\rho}(q) \geq \frac{c_{\mathcal{F}^*}}{2} \mathrm{E}[\frac{(\tilde{q}(\alpha, X) - q(\alpha, X))^2}{1-\alpha}]$ $\geq \frac{c_{\mathcal{F}^*}}{2(1-\underline{\alpha})} \lVert \tilde{q} - q \rVert^2_{\mathrm{P}_{(\alpha,X)}, 2}$ |
| $\tilde{\rho}(\tilde{q}) - \tilde{\rho}(q) \leq \frac{C_{\mathcal{F}_0^*}}{2(1-\alpha)} \lVert \tilde{q} - q \rVert^2_{\mathrm{P}_X, 2}$ | $\tilde{\rho}(\tilde{q}) - \tilde{\rho}(q) \leq \frac{C_{\mathcal{F}_0^*}}{2} \mathrm{E}[\frac{(\tilde{q}(\alpha, X) - q(\alpha, X))^2}{1-\alpha}]$ $\leq \frac{C_{\mathcal{F}_0^*}}{2(1-\overline{\alpha})} \lVert \tilde{q} - q \rVert^2_{\mathrm{P}_{(\alpha,X)}, 2}$ |
| $\tilde{\rho}(\tilde{q}) - \tilde{\rho}(q) \leq \frac{2-\alpha}{1-\alpha} \inf_{f \in \mathcal{F}} \lVert f - q \rVert^2_{\mathrm{P}_X, 1}$ | $\tilde{\rho}(\tilde{q}) - \tilde{\rho}(q) \leq \inf_{f \in \mathcal{F}} \mathrm{E}[\left(\frac{2-\alpha}{1-\alpha}\right) \lvert f(\alpha, X) - q(\alpha, X) \rvert]$ $\leq \frac{2-\overline{\alpha}}{1-\overline{\alpha}} \inf_{f \in \mathcal{F}} \lVert f - q \rVert^2_{\mathrm{P}_{(\alpha,X)}, 1}$ |
| $\rho(\mathcal{F})_{1:n}^{(n)} = \{(X_k, Y_k)_{k \in 1:n} \mapsto$ $(\rho(Y_k, f(X_k))/n)_{k \in 1:n}, f \in \mathcal{F}\}$ | $\rho(\mathcal{F})_{1:n}^{(n)} = \{(\alpha_k, X_k, Y_k)_{k \in 1:n} \mapsto$ $(\rho(\alpha_k, Y_k, f(X_k))/n)_{k \in 1:n}, f \in \mathcal{F}\}$ |

Table 1: Main changes required to adapt the previous results and proofs from a single-quantile to a multi-quantile regression setup.

discussed in Section 5.3.

Hereafter we randomize $\alpha$ and assume $\alpha \sim \mathcal{U}([\underline{\alpha}, \overline{\alpha}])$. We then consider a finite i.i.d sample $\alpha_1, \ldots, \alpha_n$ of $\alpha$, independent of $D_n$.

### 5.3 Multi-$\alpha$ learning using neural networks

**Learning with a continuum of $\alpha$'s** The finite-sample training problem for this approach can be stated as follows:

$$(\widehat{W}^{\mathrm{vars}}, \widehat{b}^{\mathrm{vars}}) \in \arg\min_{W,b} \frac{1}{n} \sum_{i=1}^n [(Y_i - \zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b))^+ + (1-\alpha_i)\zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b)],$$

where $[a, x]$ is the vector obtained by concatenating the vector $x$ to the real $a$. One can also approximately impose the non-crossing of the quantiles by penalizing the sample average of the negative part of the partial derivative $\frac{\partial}{\partial \alpha} \zeta_{l+1}^{d+1,1}([\alpha, X], W, b)$:

$$\begin{aligned}(\widehat{W}^{\mathrm{vars}}, \widehat{b}^{\mathrm{vars}}) \in \arg\min_{W,b} \frac{1}{n} \sum_{i=1}^n [&(Y_i - \zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b))^+ \\ &+ (1-\alpha_i)\zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b) + \lambda(\frac{\partial}{\partial \alpha} \zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b))^-],\end{aligned} \tag{5.1}$$

where $\lambda > 0$ determines the strength of the penalization. An approximation for $\mathrm{VaR}(Y|X)$ for any $\alpha \in (\underline{\alpha}, \overline{\alpha})$ is then given by $\zeta_{l+1}^{d+1,1}([\alpha, X], \widehat{W}^{\mathrm{vars}}, \widehat{b}^{\mathrm{vars}})$. Notice that one can compute the derivative in (5.1) fast in closed-form given our neural network parametrization, as $\frac{\partial}{\partial\alpha}\zeta_{l+1}^{d+1,1}([\alpha, X], W, b) = W_{l+1}\frac{\partial}{\partial\alpha}\zeta_n^{d+1,1}([\alpha, X], W, b)$, where

$$\frac{\partial}{\partial\alpha}\zeta_0^{d+1,1}([\alpha, X], W, b) = [1, 0_d] \text{ and, for } i = 1, \ldots, l,$$
$$\frac{\partial}{\partial\alpha}\zeta_i^{d+1,1}([\alpha, X], W, b) = (W_i\frac{\partial}{\partial\alpha}\zeta_{i-1}^{d+1,1}([\alpha, X], W, b)) \odot \sigma'(W_i\zeta_{i-1}^{d+1,1}([\alpha, X], W, b) + b_{i-1}).$$

Here $\odot$ is an element-wise product and $\sigma'$ is the derivative of $\sigma$ (applied element-wise). Given the computations of $\zeta_{l+1}^{d+1,1}([\alpha, X], W, b)$ and $\frac{\partial}{\partial\alpha}\zeta_{l+1}^{d+1,1}([\alpha, X], W, b)$ share many common sub-expressions, the recursions can be done at the same time, i.e. at each $i \in \{0, \ldots, l+1\}$, compute $\zeta_i^{d+1,1}([\alpha, X], W, b)$ and then reuse the common sub-expressions to compute also $\frac{\partial}{\partial\alpha}\zeta_i^{d+1,1}([\alpha, X], W, b)$. In the numerics, we refer to this approach with multi-$\alpha$(I) if we use a non-zero $\lambda$, and multi-$\alpha$(II) otherwise: see Algorithm 6. The ensuing approximation of $\mathrm{VaR}(Y|X)$ at the (random) confidence level $\alpha$, is given by $\widehat{q}_\alpha(X)$, where

$$\widehat{q}_a(x) \coloneqq \zeta_{l+1}^{d+1,1}(a, x, \widehat{W}^{\mathrm{vars}}, \widehat{b}^{\mathrm{vars}}), \ a \in [\underline{\alpha}, \overline{\alpha}], x \in \mathbb{R}^d$$

(see Algorithm 6).

**name :** MultiContinousVaRAlg
**input :** $\{(X_i, Y_i)\}_{i=1}^n$, a partition $B$ of $\{1 \ldots n\}$, a quantile upper bound level $\overline{\alpha}$, and lower bound level $\underline{\alpha}$, a number of epochs $E \in \mathbb{N}^\star$, a learning rate $\eta > 0$, a regularisation parameter $\lambda \geq 0$, initial values for the network parameters $\widehat{W}$ and $\widehat{b}$ and neural network output function $\zeta_{l+1}^{d+1,1}([a, x], W, b)$
**output:** Trained parameters of multi-VaR network $\widehat{W}$ and $\widehat{b}$

`// Sample quantile levels` $\alpha$
$\alpha_i \sim \mathrm{Uniform}(\underline{\alpha}, \overline{\alpha})$ for $i = 1 \ldots n$
`// Define a loss function`
**if** *non-crossing quantile regularisation* **then**
  `// multi-`$\alpha$`(I)`
  define $\rho^{\mathrm{vars}}(W, b, \mathrm{batch}) = \frac{1}{|\mathrm{batch}|}\sum_{i\in\mathrm{batch}}[(Y_i - \zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b))^+ + (1 - \alpha_i)\zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b) + \lambda(\frac{\partial}{\partial\alpha}\zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b))^-]$
  where $\frac{\partial}{\partial\alpha}\zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b))^+]$ can be quickly computed as in Section 5.3
**else**
  `// multi-`$\alpha$`(II)`
  define $\rho^{\mathrm{vars}}(W, b, \mathrm{batch}) = \frac{1}{|\mathrm{batch}|}\sum_{i\in\mathrm{batch}}[(Y_i - \zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b))^+ + (1 - \alpha_i)\zeta_{l+1}^{d+1,1}([\alpha_i, X_i], W, b)]$
**end**
$(\widehat{W}^{\mathrm{vars}}, \widehat{b}^{\mathrm{vars}}) \leftarrow \mathrm{SGDOpt}(\{(X_i, Y_i)\}_{i=1}^n, B, E, \eta, \widehat{W}, \widehat{b}, \rho^{\mathrm{vars}})$

**Algorithm 6:** Learning multi continuous VaR.

**Learning via a discrete set of $\alpha$'s and linear interpolation** Another approach for multi-$\alpha$ learning is to use a finite set of confidence levels $\alpha^{(1)} < \cdots < \alpha^{(K)}$ in $[\underline{\alpha}, \overline{\alpha}]$ in conjunction via linear interpolation. More precisely, we solve

$$(\widehat{W}^{vars}, \widehat{b}^{vars}) \in \arg\min_{W,b} \frac{1}{n} \sum_{i=1}^{n} \Big[ \Big( Y_i - \Sigma(\alpha_i, \zeta_{l+1}^{d,K}(X_i, W, b)) \Big)^+ + (1-\alpha_i) \Sigma(\zeta_{l+1}^{d,K}(\alpha_i, X_i, W, b)) \Big], \tag{5.2}$$

where, for $y = (y_0, \cdots, y_{K-1})^\top$ and $a \in \mathbb{R}$,

$$\Sigma(a, y) = y_0 + \sum_{j=1}^{K-1} \mathbf{1}_{\alpha^{(j)} \leq a} \frac{(\alpha^{(j+1)} \wedge a - \alpha^{(j)})}{\alpha^{(j+1)} - \alpha^{(j)}} y_j. \tag{5.3}$$

In (5.3), $[\zeta_{l+1}^{d,K}(x, W, b)]_0$ is a predictor of the value-at-risk of lowest grid level $\alpha^{(1)}$, whereas, for each $j \geq 1$, $[\zeta_{l+1}^{d,K}(x, W, b)]_j$ is a predictor of the increment between the value-at-risks of levels $\alpha^{(j)}$ and $\alpha^{(j+1)}$.

Notice that one can impose the monotonicity by design by adding a positive activation function $\sigma$ to each neuron in the output layer of $\zeta_{l+1}^{d+1,K}$, except for the first neuron, e.g. by replacing

$$y_j \text{ with } \sigma(y_j), \text{ for all } j \in 1, \ldots, K-1,$$

in (5.3). However we haven't found doing so to be satisfactory numerically and thus we keep the formulation in (5.3) as is. In the numerics, we refer to this approach as multi-$\alpha$(III).

The ensuing approximation of $\mathrm{VaR}(Y|X)$ at the (random) confidence level $\alpha$ is given by $\widehat{q}_\alpha(X)$, where

$$\widehat{q}_a(x) \coloneqq \Sigma(a, \zeta_{l+1}^{d,K}(x, \widehat{W}^{vars}, \widehat{b}^{vars})), \ a \in [\underline{\alpha}, \overline{\alpha}], x \in \mathbb{R}^d$$

(see Algorithm 7).

We now test the proposed procedures on a Gaussian toy-example and a dynamic initial margin (DIM) case-study. Any minimization of loss functions over $\mathcal{F}^{d,D,m,n}$ or similar sets of neural networks is done using the Adam algorithm of Kingma and Ba (2014) over the parameters $W$ and $b$ along with mini-batching: see Algorithm 8 (to be compared with Algorithm 3).

**name :** MultiDiscreteVaRAlg// `multi-`$\alpha$`(III)`

**input :** $\{(X_i, Y_i)\}_{i=1}^n$, a partition $B$ of $\{1 \dots n\}$, an increasing quantile level sequence $\alpha^{(1)} < \cdots < \alpha^{(K)}$, a number of epochs $E \in \mathbb{N}^\star$, a learning rate $\eta > 0$, initial values for the network parameters $\widehat{W}$ and $\widehat{b}$, neural network output function $\zeta_{l+1}^{d,K}(x, W, b)$

**output:** Trained parameters of multi-VaR network $\widehat{W}$ and $\widehat{b}$

`// Sample quantile levels` $\alpha$

$\alpha_i \sim \text{Uniform}(\underline{\alpha}, \overline{\alpha})$ for $i = 1 \dots n$

`// Define a loss function`

define $\Sigma(y, a) = y_0 + \sum_{j=1}^{K-1} \mathbf{1}_{\alpha^{(j)} \leq a} \frac{(\alpha^{(j+1)} \wedge a - \alpha^{(j)})}{\alpha^{(j+1)} - \alpha^{(j)}} y_j$

define $\rho^{vars}(W, b, \text{batch}) =$

$$\frac{1}{|\text{batch}|} \sum_{i \in \text{batch}} \left[ \left( Y_i - \Sigma(\zeta_{l+1}^{d,K}(X_i, W, b), \alpha_i) \right)^+ + (1 - \alpha_i) \Sigma(\zeta_{l+1}^{d,K}(X_i, W, b), \alpha_i) \right]$$

$(\widehat{W}^{vars}, \widehat{b}^{vars}) \leftarrow \text{SGDOpt}(\{(X_i, Y_i)\}_{i=1}^n, B, E, \eta, \widehat{W}, \widehat{b}, \rho^{vars})$

**Algorithm 7:** Learning multi discrete VaR.

**name :** SGDOpt `// Adam variant`

**input :** $\{(X_i, Y_i)\}_{i=1}^n$, a partition $B$ of $\{1 \dots n\}$, a number of epochs $E \in \mathbb{N}^\star$, a learning rate $\eta > 0$, initial weight (matrix) $\widehat{W}$ and bias (vector) $\widehat{b}$ parameters, and a loss function $\rho = \rho(W, b, \text{batch})$

**output:** Trained parameters $\widehat{W}$ and $\widehat{b}$

`// Set exponential decay rates for the first and second moment estimates and a small number`

$\beta_1 \leftarrow 0.9$; $\beta_2 \leftarrow 0.999$; $\epsilon \leftarrow 1e-8$

$t \leftarrow 1$; $m_0 \leftarrow 0$; $v_0 \leftarrow 0$

**for** *epoch* $= 1, \dots, E$ **do** `// loop over epochs`

  **for** *batch* $\in B$ **do** `// loop over batches`

    $g_t \leftarrow \nabla_{(W,b)} \rho(\widehat{W}, \widehat{b}, \text{batch})$ ; `// Get gradient of parameter`

    $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$ ; `// Update biased first moment estimate`

    $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) (g_t)^2$ ; `// Update biased second moment estimate`

    $m \leftarrow \frac{m_t}{1 - (\beta_1)^t}$ ; `// Compute bias-corrected first moment estimate`

    $v \leftarrow \frac{v_t}{1 - (\beta_2)^t}$ ; `// Compute bias-corrected second moment estimate`

    $(\widehat{W}, \widehat{b}) \leftarrow (\widehat{W}, \widehat{b}) - \frac{\eta}{\sqrt{v} + \epsilon} m$ ; `// Update parameters`

    $t \leftarrow t + 1$

  **end**

**end**

**Algorithm 8:** Adam algorithm learning neural network parameters.

All neural networks have 3 hidden layers, and twice their input dimensionality as the number of neurons per hidden layer. In both examples below, for the multi-$\alpha$(I) and multi-$\alpha$(II) learning approaches, we use the bounds $(1-\underline{\alpha}, 1-\overline{\alpha}) = (10^{-4}, 0.15)$. For the multi-$\alpha$(III) approach, we use a uniform interpolation grid $1 - \alpha^{(k)} = 10^{-3} + k \frac{0.15 - 10^{-3}}{20}$, with $k \in \{0, \dots, 20\}$.

# 6 Conditionally Gaussian Toy Model

In our toy example, we apply the above algorithms to the data generating process $(X, Y)$ such that $X$ is a standard multivariate normal vector and, conditional on $X$, $Y$ is normally distributed. Namely,

$$X \sim \mathcal{N}(0, I_d), \text{ for some } d \in \mathbb{N}^\star, \text{ and } (Y|X) \sim \mathcal{N}(P(X), Q(X)^2),$$

where $P$ and $Q$ are multivariate polynomials of degree 2, i.e. for some coefficients $\lambda$ and $\mu$ we have $P(x) = \lambda_0 + \sum_{i=1}^d \lambda_i x_i + \sum_{1 \le i < j \le d} \lambda_{i,j} x_i x_j$ and $Q(x) = \mu_0 + \sum_{i=1}^d \mu_i x_i + \sum_{1 \le i < j \le d} \mu_{i,j} x_i x_j$, for every $x = (x_1, \dots, x_d) \in \mathbb{R}^d$.

Then, denoting by $\Phi$ the cdf of the standard normal distribution and by $\varphi$ its pdf, we have:

$$q(X) = \mathrm{VaR}(Y|X) = P(X) + |Q(X)|\Phi^{-1}(\alpha)$$
$$s(X) = \mathrm{ES}(Y|X) = P(X) + (1-\alpha)^{-1}|Q(X)|\varphi(\Phi^{-1}(\alpha)),$$

which will serve us as ground-truth values.

## 6.1 Numerical Results

We use a dimension of $d = 25$ for the state space of $X$, leading to $1 + d + \frac{d(d+1)}{2} = 351$ monomials in the multivariate polynomials $P$ and $Q$. The coefficients $\lambda$ and $\mu$ of those monomials are drawn independently from a standard normal distribution. For this example, we use $2^{19} = 524288$ training points and the same number of testing points for computing the errors. For the Adam algorithm, we used 2000 epochs, mini-batching with batches of size $2^{15} = 32768$, a learning rate $\gamma = 0.01$, and the rest of the parameters kept at their default values in Kingma and Ba (2014).

Tables 2, as also 6, 7 and 8 below in the DIM case, suggest that the multi-$\alpha$ approaches are competitive compared to the single-$\alpha$ approach by yielding acceptable errors for confidence levels below 99%, while requiring only one single training, as opposed to the single-$\alpha$ approach which requires one training per target confidence level. For very extreme confidence levels, like 99.9%, the multi-$\alpha$(III) approach outperforms all the other approaches. This can be explained by the fact that, even if the target confidence level is hard to reach given a limited training set, the lower confidence levels in the interpolation grid contribute to inferring the VaR for the target confidence level. Table 3 confirms that one can rely on the twin-simulation trick of Section 3.5 to draw mostly similar conclusions as in Table 2, without the need to have access to the goundtruth estimators. Note that we computed upper-bounds of 95% confidence intervals for (3.39), instead of the estimates directly in order to be conservative and take into account the potentially high variance in the indicator functions that need to be simulated in order to estimate (3.39). Table 4 demonstrates the effectiveness of the penalization term (for $\lambda$ simply set to 1) in the multi-$\alpha$(I) approach to mitigate the quantiles crossing problem. Table 4 also shows that the other multi-$\alpha$ learning approaches, even without directly penalizing the crossing of the quantiles, behave better than a single-$\alpha$ learning in terms of the crossing of the quantiles.

For the ES learning in the Gaussian toy-example, for brevity, we denote by "LR using M VaR" an ES learning using linear regression only for the output layer, as

| $\alpha$ | 0.999 | 0.995 | 0.99 |
|---|---|---|---|
| multi-$\alpha$(I) | 0.151 (0.004) | 0.060 (0.002) | **0.039** (0.001) |
| multi-$\alpha$(II) | 0.161 (0.004) | 0.065 (0.002) | 0.042 (0.002) |
| multi-$\alpha$(III) | **0.061** (0.002) | **0.051** (0.002) | 0.043 (0.001) |
| Single-$\alpha$ | 0.612 (0.043) | 0.062 (0.001) | 0.044 (0.001) |
| $\alpha$ | 0.98 | 0.95 | 0.9 |
| multi-$\alpha$(I) | **0.029** (0.001) | 0.023 (0.001) | 0.018 (0.001) |
| multi-$\alpha$(II) | 0.031 (0.001) | 0.024 (0.001) | 0.019 (0.001) |
| multi-$\alpha$(III) | 0.037 (0.001) | 0.029 (0.001) | 0.025 (0.001) |
| Single-$\alpha$ | 0.032 (0.001) | **0.021** (0.001) | **0.016** (0.001) |

Table 2: Means (standard deviations) of RMSE errors of learned conditional VaR estimators against groundtruth values in the Gaussian toy-example across 32 runs. Errors are normalized by dividing by the standard deviation of the groundtruth VaR.

| $\alpha$ | 0.999 | 0.995 | 0.99 |
|---|---|---|---|
| multi-$\alpha$(I) | 0.00020 (0.000010) | 0.00021 (0.000009) | 0.00027 (0.000008) |
| multi-$\alpha$(II) | 0.00023 (0.000013) | 0.00024 (0.000013) | 0.00029 (0.000013) |
| multi-$\alpha$(III) | **0.00003** (0.000002) | **0.00008** (0.000003) | **0.00024** (0.000008) |
| Single-$\alpha$ | 0.00008 (0.000003) | 0.00020 (0.000007) | 0.00035 (0.000008) |
| $\alpha$ | 0.98 | 0.95 | 0.9 |
| multi-$\alpha$(I) | **0.00046** (0.000009) | **0.00157** (0.000020) | 0.00379 (0.000060) |
| multi-$\alpha$(II) | **0.00046** (0.000009) | **0.00157** (0.000030) | 0.00398 (0.000086) |
| multi-$\alpha$(III) | 0.00057 (0.000015) | 0.00171 (0.000030) | 0.00428 (0.000066) |
| Single-$\alpha$ | 0.00066 (0.000008) | 0.00171 (0.000029) | **0.00343** (0.000069) |

Table 3: Means (standard deviations) across 32 runs of the upper-bounds of 95% confidence intervals of $L_2$ $p$-value error estimates, i.e. as defined in (3.39), of learned conditional VaR estimators in the Gaussian toy-example.

described in Section 4.3, and a VaR learned using the method M as the candidate VaR. For example, LR using single-$\alpha$ VaR refers to the linear regression approach for learning the ES, by using a VaR that is learned with the single-$\alpha$ approach as the VaR candidate. To demonstrate the effectiveness of this linear regression approach, we also introduce an ES that is learned by neural regression, by using a neural network in (4.4), without freezing any weights and using the groundtruth VaR as the VaR candidate. Table 5 shows that our linear regression approach for the ES outperforms the neural regression, no matter which approach is used for learning the embedded VaR candidate. The relative performance of the different linear regression approaches in Table 5 is explained by the relative performance of the VaR learning approaches, given that the VaR learning error contributes to the ES learning error.

| $E$ | $q_{0.999}(X) < q_{0.995}(X)$ | $q_{0.995}(X) < q_{0.99}(X)$ | $q_{0.99}(X) < q_{0.98}(X)$ |
|---|---|---|---|
| multi-$\alpha$(I) | **0.000004** (0.000001) | **0.000005** (0.000002) | **0.000008** (0.000003) |
| multi-$\alpha$(II) | 0.000016 (0.000008) | 0.000017 (0.000007) | 0.000020 (0.000008) |
| multi-$\alpha$(III) | 0.000461 (0.000107) | 0.000164 (0.000037) | 0.002765 (0.000619) |
| Single-$\alpha$ | 0.111117 (0.003184) | 0.251983 (0.006574) | 0.213348 (0.005818) |
| $E$ | $q_{0.98}(X) < q_{0.97}(X)$ | $q_{0.97}(X) < q_{0.96}(X)$ | $q_{0.96}(X) < q_{0.95}(X)$ |
| multi-$\alpha$(I) | **0.000022** (0.000007) | **0.000073** (0.000017) | **0.000367** (0.000059) |
| multi-$\alpha$(II) | 0.000032 (0.000008) | 0.000080 (0.000012) | 0.000405 (0.000096) |
| multi-$\alpha$(III) | 0.016378 (0.003258) | 0.159370 (0.011163) | 0.011956 (0.002695) |
| Single-$\alpha$ | 0.272327 (0.005291) | 0.316263 (0.006022) | 0.336678 (0.004992) |

Table 4: Empirical estimates (and corresponding standard deviations) of $\mathrm{P}(E)$, for the events $E$ listed in the first row, for learned conditional VaR estimators in the Gaussian toy-example across 32 runs.

| $\alpha$ | 0.999 | 0.995 | 0.99 |
|---|---|---|---|
| NNR using true VaR | 0.408 (0.013) | 0.106 (0.002) | 0.076 (0.002) |
| LR using single-$\alpha$ VaR | 0.536 (0.037) | 0.062 (0.001) | 0.045 (0.001) |
| LR using multi-$\alpha$(I) VaR | 1.900 (0.166) | 0.068 (0.004) | **0.037** (0.002) |
| LR using multi-$\alpha$(II) VaR | 2.382 (0.174) | 0.082 (0.006) | 0.041 (0.002) |
| LR using multi-$\alpha$(III) VaR | **0.126** (0.005) | **0.057** (0.002) | 0.050 (0.002) |
| $\alpha$ | 0.98 | 0.95 | 0.9 |
| NNR using true VaR | 0.054 (0.001) | 0.041 (0.001) | 0.034 (0.001) |
| LR using single-$\alpha$ VaR | 0.034 (0.001) | **0.025** (0.001) | **0.021** (0.001) |
| LR using multi-$\alpha$(I) VaR | **0.031** (0.001) | **0.025** (0.001) | 0.022 (0.001) |
| LR using multi-$\alpha$(II) VaR | 0.032 (0.001) | 0.026 (0.001) | 0.023 (0.001) |
| LR using multi-$\alpha$(III) VaR | 0.043 (0.002) | 0.036 (0.001) | 0.030 (0.001) |

Table 5: Means (standard deviations) of RMSE errors of learned conditional ES estimators against groundtruth values in the Gaussian toy-example across 32 runs. Errors are normalized by dividing by the stdev of the groundtruth ES.

# 7 Dynamic Initial Margin Case Study

A financial application of the quantile learning framework is the learning of a path-wise, dynamic initial margin (DIM) in the context of XVA computations (see e.g. Albanese, Crépey, Hoskinson, and Saadeddine (2021, Section 5)). Let there be given respectively $\mathbb{R}^d$ valued and real valued stochastic proceesses $X = (X_t)_{t\geq 0}$ and $\mathrm{MtM} = (\mathrm{MtM}_t)_{t\geq 0}$, where $X$ is Markov and $X_t$ represents the state of the market at time $t$ (e.g. diffused market risk factors), whereas $\mathrm{MtM}_t$ represents the mark-to-market (price) of the portfolio of the bank at time $t$—cumulative price including the cash flows cumulated up to

time $t$, such that $\mathrm{MtM}_{t+\delta} - \mathrm{MtM}_t$ is $\sigma(X_s, t \le s \le t+\delta)$ measurable. We ignore risk-free discounting in the notation (while preserving it in the numerical experiments). The initial margin of the bank at time $t$ at the confidence level $\alpha$, denoted by $\mathrm{IM}_t$, defined as

$$\mathrm{IM}_t := \mathrm{VaR}\left(\mathrm{MtM}_{t+\delta} - \mathrm{MtM}_t \,|X_t\right). \tag{7.1}$$

Hence, having simulated paths of $X$ and MtM, one can estimate the initial margin at each simulation grid time $t > 0$, i.e. the DIM process, using quantile learning at each $t$.

**Estimating $\mathrm{IM}_t$ using a nested Monte Carlo** Alternatively, given $t > 0$, one can consider a brute force nested Monte Carlo scheme based on $n_{\mathrm{outer}}$ i.i.d samples

$$(X_t^{(1)}, \mathrm{MtM}_t^{(1)}), \dots, (X_t^{(n_{\mathrm{outer}})}, \mathrm{MtM}_t^{(n_{\mathrm{outer}})})$$

of $(X_t, \mathrm{MtM}_t)$ and, for each $i \in \{1, \dots, n_{\mathrm{outer}}\}$, $K$ i.i.d sub-samples

$$\{\mathrm{MtM}_{t+\delta}^{(i,1),[1]}, \dots, \mathrm{MtM}_{t+\delta}^{(i,n_{\mathrm{inner}}),[1]}\}, \dots, \{\mathrm{MtM}_{t+\delta}^{(i,1),[K]}, \dots, \mathrm{MtM}_{t+\delta}^{(i,n_{\mathrm{inner}}),[K]}\}$$

of $\mathrm{MtM}_{t+\delta}$ conditional on $X_t = X_t^{(i)}$. We can then use these sub-simulations to estimate the conditional quantile in (7.1), for each realization $X_t^{(\nu)}$ of $X_t$. For GPU memory limitation reasons, and in order to avoid having to store simulations on the global memory, we chose to do so via one stochastic approximation algorithm per (conditional on) each outer simulation node. More precisely, for every $i \in \{1, \dots, n_{\mathrm{outer}}\}$, we define iteratively over $k \in \{1, \dots, K\}$:

$$\mathrm{IM}_t^{(i),[k+1]} := \mathrm{IM}_t^{(i),[k]} + \gamma(prop^{(i),[k]} - 1 + \alpha)$$

where $\gamma$ is a positive learning rate (see below) and

$$prop^{(i),[k]} := \frac{1}{n_{\mathrm{inner}}} \sum_{j=1}^{n_{\mathrm{inner}}} \mathbf{1}_{\{\mathrm{MtM}_{t+\delta}^{(i,j),[k]} - \mathrm{MtM}_t^{(i)} \ge \mathrm{IM}_t^{(i),[k]}\}},$$

One then iterates over $k$, simultaneously for all $i$ in parallel, until convergence in order to obtain an approximation of $\mathrm{IM}_t$ at each outer realization of $X_t$. This corresponds to a value-at-risk stochastic approximation algorithm, namely the batched version of (Barrera, Crépey, Diallo, Fort, Gobet, and Stazhynski, 2019, Algorithm 0), run conditionally on each outer simulation node at time $t$ (cf. (Barrera, Crépey, Diallo, Fort, Gobet, and Stazhynski, 2019, Section 5.3.1)). To speed-up the convergence, we take $\gamma$ to be of the order of the conditional standard deviation of $\mathrm{MtM}_{t+\delta} - \mathrm{MtM}_t$, itself estimated via the same nested Monte Carlo procedure, and we use a Gaussian VaR as the initial value (i.e. $\mathrm{IM}_t^{(0),[k]}$), computed using conditional expectation and standard deviation estimates using the inner samples at the first iteration. $n_{\mathrm{inner}} = 1024$ samples for the sub-simulations and $K = 256$ iterations are then enough to achieve an error in $p$-value, as computed using (3.39), of roughly $0.5(1-\alpha)$ in our experiments.

### 7.1 Numerical Results

We consider a portfolio composed of 100 interest rate swaps with randomly drawn characteristics and final maturity 10 years, assessed in the market model of Abbas-Turki, Crépey, and Saadeddine (2022, Appendix B), i.e. a multi-factor market model with 10 short-rate processes representing 10 economies and 9 cross-currency rate processes. Given that swap coupons can depend on short-rates at previous fixing dates, we also include in the regression basis the same short-rates but observed at the latest previous fixing date, which leads in total to a dimensionality of $d = 29$ for the state vector $X_t$ at a given time $t > 0$, with 100 time steps uniformly spread between time 0 and the final maturity of the portfolio equal to 10 years. We use $2^{22} = 4194304$ simulated paths (generated in 25 seconds using the code developed in Abbas-Turki, Crépey, and Saadeddine (2022, Appendix B)) for training and $2^{14}$ simulated paths, independent of the former, for evaluating the nested Monte Carlo benchmark and computing the errors. We leverage the transfer learning trick used in Abbas-Turki, Crépey, and Saadeddine (2022, Appendix B), which consists in doing the training starting from the latest time-step and then proceeding backwards by reusing the solution obtained at each successive time-step $t_{k+1}$ as an initialization for the learning to be done at time $t_k$. This allows us to use only 16 training epochs. As in the Gaussian toy-example, we use mini-batching. The batch size is taken to be $2^{17} = 131072$, we use a learning rate of 0.001, and the rest of the Adam parameters are kept at their default values.

To illustrate that the quantile learning approach allows one to learn an entire stochastic process (dynamic initial margin), we plot the mean and 5-th/95-th percentiles of the learned IM process at each time-step for the different quantile learning schemes in Figure 1. The sawtooth-like behaviour in the paths of $(\mathrm{IM}_t)_{t\geq 0}$ that is visible in the plots in Figure 1 is expected, due to the recurring cash-flows inherent to interest rate swaps (Andersen, Pykhtin, and Sokol, 2017).

Tables 6, 7 and 8 (using the nested Monte Carlo as a benchmark) confirms the conclusions of Table 2 regarding the competitiveness of the multi-$\alpha$ approaches.

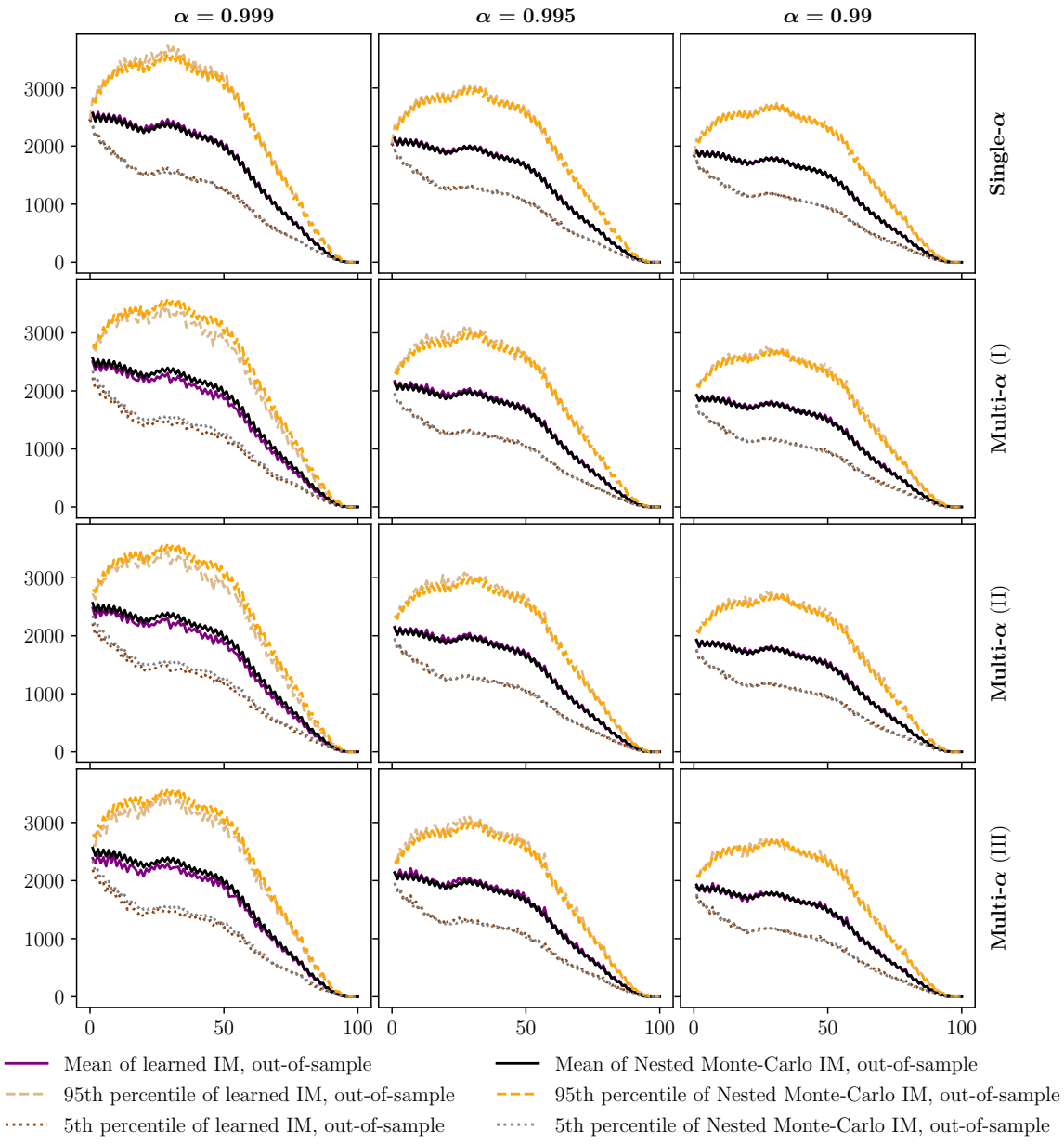


Figure 1: Mean and 5-th/95-th percentiles of both the learned and the nested Monte Carlo IM at different time steps and for different values of $\alpha$ and learning approaches. The learning approach used for the plots in each row is indicated on the right, and each column corresponds to one value of $\alpha$ which is indicated at the top of each column. Statistics are computed using out-of-sample trajectories of the diffused risk-factors, and the time steps are on the $x$-axis.

| $\alpha$ | 0.999 | 0.995 | 0.99 | 0.98 | 0.95 | 0.9 |
|---|---|---|---|---|---|---|
| multi-$\alpha$(I) | 0.265 | 0.160 | 0.109 | 0.065 | 0.058 | **0.056** |
| multi-$\alpha$(II) | 0.261 | 0.155 | 0.107 | 0.066 | **0.057** | **0.056** |
| multi-$\alpha$(III) | **0.128** | 0.185 | 0.102 | 0.133 | 0.116 | 0.074 |
| Single-$\alpha$ | 0.134 | **0.074** | **0.070** | **0.056** | 0.066 | 0.065 |

Table 6: RMSE errors of learned $\mathrm{IM}_t$ estimators against nested Monte Carlo estimators, for $t = 2.5$years. Errors are normalized by dividing by the standard deviation of the nested Monte Carlo benchmark.

| $\alpha$ | 0.999 | 0.995 | 0.99 | 0.98 | 0.95 | 0.9 |
|---|---|---|---|---|---|---|
| multi-$\alpha$(I) | 0.204 | 0.166 | 0.131 | 0.072 | 0.061 | 0.069 |
| multi-$\alpha$(II) | 0.212 | 0.162 | 0.127 | 0.072 | 0.062 | 0.069 |
| multi-$\alpha$(III) | **0.150** | 0.123 | **0.067** | 0.065 | 0.066 | 0.068 |
| Single-$\alpha$ | 0.165 | **0.095** | 0.070 | **0.057** | **0.060** | **0.066** |

Table 7: RMSE errors of learned $\mathrm{IM}_t$ estimators against nested Monte Carlo estimators, for $t = 5$years. Errors are normalized by dividing by the stdev of the nested Monte Carlo benchmark.

| $\alpha$ | 0.999 | 0.995 | 0.99 | 0.98 | 0.95 | 0.9 |
|---|---|---|---|---|---|---|
| multi-$\alpha$(I) | 0.292 | 0.119 | 0.122 | 0.095 | 0.073 | 0.072 |
| multi-$\alpha$(II) | 0.296 | 0.118 | 0.118 | 0.091 | 0.071 | 0.070 |
| multi-$\alpha$(III) | 0.157 | 0.118 | 0.090 | 0.089 | 0.079 | 0.086 |
| Single-$\alpha$ | **0.119** | **0.088** | **0.082** | **0.068** | **0.061** | **0.061** |

Table 8: RMSE errors of learned $\mathrm{IM}_t$ estimators against nested Monte Carlo estimators, for $t = 7.5$years. Errors are normalized by dividing by the stdev of the nested Monte Carlo benchmark.

## Conclusion

The numerical experiments of Sections 6 and 7 suggest that learning multiple quantiles (multi-$\alpha$(I), multi-$\alpha$(II) or multi-$\alpha$(III)), although counter-intuitive at first, can help better target extreme quantiles than a standard single quantile learning approach. This can be explained by the fact that multiple quantile approaches leverage the information given by nearby quantiles and thus are better at extrapolating at the extremes. The multi-$\alpha$(I) approach is remarkably good at ensuring, via soft-constraints on the derivative with respect to the quantile level, monotonicity (avoiding quantile crossings), in cases where consistency among different quantile levels is desired. Our experiments also show that one can successfully use these quantile estimation methods in an XVA

or dynamic risk calculation setting, where the computation times may be greatly accelerated by replacing nested Monte Carlo estimations by quantile and expected-shortfall learnings.

# A Value-at-Risk and Expected Shortfall Representations

In this appendix we recall various elicitability results underlying our VaR and ES learning algorithms.

A cumulative distribution function (cdf) $F : \mathbb{R} \to [0,1]$ is by definition (Stieltjes) integrable if

$$\int_{\mathbb{R}} |y| \, F(dy) < \infty. \tag{A.1}$$

If $Y$ is a random variable with distribution function $F$ (i.e. $\mathrm{P}[Y \leq t] = F(t)$, $t \in \mathbb{R}$), then (A.1) holds if and only if $Y$ is P-integrable (the left-hand side of (A.1) is then $\mathrm{E}\,[|Y|]$).

**Definition A.1.** *Let $F : \mathbb{R} \to [0,1]$ be an integrable cdf and let $\alpha \in (0,1)$. The* value-at-risk *(*VaR*) and* expected shortfall *(*ES*) of $F$ at the confidence level $\alpha$ are defined respectively by*

$$\mathrm{VaR}(F) := \inf F^{-1}([\alpha, 1])\,, \ \mathrm{ES}(F) = \frac{1}{1 - F(\mathrm{VaR}(F)-)} \int_{[\mathrm{VaR}(F),\infty)} y \, F(dy). \tag{A.2}$$

*(see* (2.2) *for the definition of $F(y_0-)$). If $Y$ is an integrable random variable on the probability space $(\Omega, \mathcal{A}, \mathrm{P})$, we write*

$$\mathrm{VaR}(Y) := \mathrm{VaR}(F_Y), \qquad \mathrm{ES}(Y) := \mathrm{ES}(F_Y)$$

*where $F_Y(t) = \mathrm{P}\,[Y \leq t]$ is the distribution function of $Y$.*

**Remark A.2.** *If $Y$ is an integrable random variable, then it is easy to see that*

$$\mathrm{VaR}(Y) = \inf\{t : \mathrm{P}\,[Y \leq t] \geq \alpha\}, \qquad \mathrm{ES}(Y) = \mathrm{E}\,[Y | Y \geq \mathrm{VaR}(Y)] \tag{A.3}$$

*(the conditional expectation is with respect to* P*). In particular,*

$$\mathrm{VaR}(Y) \leq \mathrm{ES}(Y), \tag{A.4}$$

*with equality if and only if*

$$\mathrm{P}\,[Y \leq \mathrm{VaR}(Y)] = 1. \tag{A.5}$$

*The versions of* (A.3)*,* (A.4) *and* (A.5) *for abstract distribution functions $F$ are clear mutatis mutandis.*

**Remark A.3.** *It is necessary to assume that our random variables are bounded (possibly after transformation as explained in Sections 2.2 and Appendix B) in order to obtain nonasymptotic bounds in the errors induced by the methods to approximate* VaR *and* ES *presented here (see for instance* (3.20)*).*

*This entails no loss of generality for* VaR*. To see why, let $Y$ be any integrable random variable defined on $(\Omega, \mathcal{A}, \mathrm{P})$, let $I \subset \mathbb{R}$ be a (possibly infinite) interval supporting $Y$ ($\mathrm{P}[Y \in I] = 1$), let $-\infty < a < b < \infty$, and let $h : I \to (a, b)$ be any increasing bijective, Borel measurable function. Then by monotonicity*

$$\mathrm{VaR}(h(Y)) = h(\mathrm{VaR}(Y)),$$

*which allows reducing the approximation of* $\mathrm{VaR}(Y)$ *to the bounded case: to approximate* $\mathrm{VaR}(Y)$*, approximate* $\mathrm{VaR}(h(Y))$ *and compose with $h^{-1}$. The error bounds provided in this paper, which apply to* $\mathrm{VaR}(h(Y))$*, can then be translated into error bounds on the approximation of* $\mathrm{VaR}(Y)$ *using ad hoc analytic properties of $h$.*

*As for* ES*, notice that for such $h$*

$$\mathrm{ES}(h(Y))\mathbb{1}_{\{h(Y)\geq \mathrm{VaR}(h(Y))\}} = \mathrm{E}\left[h(Y)|\mathbb{1}_{\{h(Y)\geq \mathrm{VaR}(h(Y))\}}\right] = \mathrm{E}\left[h(Y)|\mathbb{1}_{\{Y\geq \mathrm{VaR}(Y)\}}\right].$$

*From this it follows that if $h$ is in addition convex [concave] on $I \cap [\mathrm{VaR}(Y), \infty)$, then*[9]

$$\mathrm{ES}(Y) \leq [\geq] h^{-1}(\mathrm{ES}(h(Y))). \tag{A.6}$$

*The inequality* (A.6) *for convex [concave] $h$ shows that $h^{-1}(\mathrm{ES}(h(Y)))$ is a conservative [risky] estimate of* $\mathrm{ES}(Y)$*. Notice that such conservative* ES *estimates are only available when $Y$ is assumed upper bounded, for there is no convex, increasing and bounded bijection with domain $[a, \infty)$. Note also that if $h$ is an increasing affine transformation, then* $\mathrm{ES}(h(Y)) = h(\mathrm{ES}(Y))$.

It is convenient for what follows to present the discussion in terms of distribution functions. We start by noticing that if $F$ has an $\alpha$-quantile, namely if

$$F(y) = \alpha \qquad \textit{for some } y,$$

then $\mathrm{VaR}(F)$ is the minimum of such $y$'s. In this case (and this case only)

$$F(\mathrm{VaR}(F)) = \alpha. \tag{A.7}$$

By the intermediate value theorem, such $y$ exists in $[a, b]$ if

[9] If $Z$ is an integrable random variable on $(\Omega, \mathcal{A}, \mathrm{P})$ and $\mathcal{A}_0 \subset \mathcal{A}$ is a sigma-algebra, then for every convex, bijective and bimeasurable function $h : \mathbb{R} \to \mathbb{R}$,

$$h^{-1}(\mathrm{E}[h(Z)|\mathcal{A}_0]) \geq \mathrm{E}\left[h^{-1}(h(Z))|\mathcal{A}_0\right] = \mathrm{E}[Z|\mathcal{A}_0].$$

If $\mathcal{A}_0 = \sigma(\mathbb{1}_{\{Y\geq a\}})$ and the invertible, bimeasurable function $h : \mathbb{R} \to \mathbb{R}$ is convex in the interval $J = I \cap [a, \infty)$ where $\mathrm{P}[Y \in I] = 1$, then $h_0 = h\mathbb{1}_{I\cap[a,\infty)} + h_1\mathbb{1}_{\mathbb{R}\setminus(I\cap[a,\infty))}$ is convex, invertible and bimeasurable in $\mathbb{R}$ for an appropriate $h_1 : \mathbb{R} \to \mathbb{R}$, and $\mathrm{E}[h(Y)|\mathcal{A}_0] = \mathrm{E}[h_0(Y)|\mathcal{A}_0] = \mathrm{E}[h_0(Y)|Y \geq a]\,\mathbb{1}_{\{Y\geq a\}}$. Even more, $\mathrm{E}[h_0(Y)|Y \geq a] = \mathrm{E}[h(Y)|Y \geq a]$ because $h_0|_{I\cap[a,\infty)} = h|_{I\cap[a,\infty)}$. The argument for concave $h$ is similar.

**Assumption A.4.** *There exists an interval $[a, b]$ where $F$ is continuous and*

$$F(a) < \alpha \leq F(b).$$

The following operator will allow us to characterize VaR and ES as minimizers of a suitable functional.

**Definition A.5.** *Given a Polish space $S$, a (Borel measurable) function $h : S \times \mathbb{R} \to \mathbb{R}$ and a distribution function $F$, we define $(h * F) : S \to \mathbb{R}$ by*

$$(h * F)(x) = \int_{\mathbb{R}} h(x, y)\, F(dy)$$

*provided that $h(x, \cdot)$ is $F$-integrable for all $x$. When necessary, we will write $(h*F)(\cdot) = h(\cdot, y) * F(dy)$.*

Recall (2.5) and (2.6). Our methods are built over the following results of Rockafellar and Uryasev (2000)[10], whose easy proof we give for the sake of completeness:

**Lemma A.1.** *If $F$ is an integrable distribution function satisfying Assumption A.4, then the set of minimizers of the function $(\rho_\iota * F)|_{[a,b]}$ is the set of $\alpha$-quantiles of $F$ within $[a, b]$, and given $c > 0$,*

$$\mathrm{ES}(F) = \frac{1}{c} \min_v (\rho_{c\cdot} * F)|_{[a,b]}(v), \tag{A.8}$$

*where $c\cdot$ denotes the function $y \mapsto cy$.*

***Proof.*** Since $\iota$ is increasing and continuous, and since $F$ is absolutely continous in $[a, b]$, the identity

$$\begin{aligned}(\rho_\iota * F)'(v) =& \frac{d}{dv}\left((1-\alpha)^{-1}\int_v^\infty (\iota(y) - \iota(v))F(dy) + \iota(v)\right)\\ =& \iota'(v)(1 - (1-\alpha)^{-1}(1 - F(v))).\end{aligned}$$

holds for $v \in [a, b]$. It follows in particular that the (continuously differentiable) function $(\rho_\iota * F)|_{[a,b]}$ has critical points in the set of $\alpha$-quantiles of $F$ within $[a, b]$. Since $F$ is increasing, these critical points are the minimizers of $\rho_\iota * F$.

With this, (A.8) is a straigthforward consequence of the definition (A.2) of $\mathrm{ES}(F)$ together with (A.7): given any $\alpha$-quantile $q$ of $F$ within $[a, b]$, and since $F$ is constant in $[\mathrm{VaR}(F), q)$,

$$\begin{aligned}\mathrm{ES}(F) =& (1-\alpha)^{-1}\int_q^\infty yF(dy) = (1-\alpha)^{-1}\int_{\mathbb{R}} (y-q)^+ F(dy) + q\\ =& \frac{1}{c}(\rho_{c\cdot} * F)|_{[a,b]}(q) = \frac{1}{c}\min_v (\rho_{c\cdot} * F)|_{[a,b]}(v),\end{aligned}$$

where for the last equality we used the first part already proved. ∎

[10] where we only added $\iota$ for the sake of data transformation to boundedness.

Notice that the estimation of ES via (A.8) implies the estimation of an integral with respect to $F$. It is desirable, in order to propose distribution-free methods for the estimation of ES, to have characterizations of this risk measure as a *minimizer* (rather than a minimum). The following theorem presents the first one, which works given a corresponding $\alpha$-quantile:

**Lemma A.2.** *If $F$ is an integrable distribution function and if $q$ is an $\alpha$-quantile of $F$, then $\mathrm{ES}(F) - q \in [0, \infty)$ is the unique minimizer of $\varrho_\varsigma(y, q, \cdot) * F(dy)|_{[0,\infty)}$.*

***Proof.*** In this case,

$$\frac{d}{dz}(\varrho_\varsigma(y, q, z) * F(dy)) = \varsigma''(z) \left( z - (1-\alpha)^{-1} \int_{\mathbb{R}} (y-q)^+ F(dy) \right),$$

which changes from negative to positive at $z = \mathrm{ES}(F) - q$ because $\varsigma''(z) > 0$: this follows as in the proof of (A.8). ∎

Inspired by Corollary 5.5 in Fissler and Ziegel (2016), we finally present the following "joint" loss, which is basically a combination of (2.5) and (2.6), for the elicitability of $(\mathrm{VaR}, \mathrm{ES})$ based on the loss function (2.8).

**Lemma A.3.** *For every integrable cdf $F$ satisfying Assumption A.4, $(F^{-1}(\alpha) \cap [a, b]) \times \{\mathrm{ES}(F)\}$ is the set of minimizers of the function*

$$\rho_{\iota,\varsigma}(y, \cdot, \cdot) * F(dy) : [a, b] \times \mathbb{R} \to \mathbb{R}. \tag{A.9}$$

***Proof.*** The derivative of (A.9) with respect to $v$ is

$$(\iota'(v) - \varsigma'(z))(1 - (1-\alpha)^{-1}(1 - F(v))) \tag{A.10}$$

which equals zero if and only if $v \in F^{-1}(\alpha)$ by the assumptions on $\iota'$ and $\varsigma'$.

By a similar calculation and using $\varsigma'' \neq 0$, the derivative of (A.9) with respect to $z$ is zero if and only if

$$z = v + (1-\alpha)^{-1} \int_{\mathbb{R}} (y - v)^+ F(dy),$$

which, as justified in the proof of Lemma A.2, gives $z = \mathrm{ES}(F)$ if $v \in F^{-1}(\alpha)$.

It follows that $(F^{-1}(\alpha) \cap [a, b]) \times \{\mathrm{ES}(F)\}$ is the set of critical points of (A.9). The fact that these critical points are indeed minimizers of (A.9) follows by an argument akin to the proof of Lemma A.1 (consider $z = \mathrm{ES}(Y)$ fixed and the expression (A.10) for the derivative with respect to $v$). ∎

# B The Role of Data Transformations and Truncations

The functions $h_k(x, y)$ $(k = 1, 2)$ in Algorithm 1 serve at least two purposes: to uniformly bound and normalize the data, in particular to make it fit to the theory of (Barrera, 2022), and to open the room for profiting from a priori information about the conditional distributions of $Y$ given $X$.

Let us discuss the functions involved in the estimation of ES: the reason for restricting ourselves to conditionally affine transformations

$$h(x,y) = \tau(x)y + \nu(x) \qquad (a(x) > 0) \tag{B.1}$$

is that, as explained in Remark A.3, only these satisfy (in general) the equation

$$\mathrm{ES}(h(X,Y)|X) = h(X, \mathrm{ES}(Y|X)), \tag{B.2}$$

thus allowing us to compute $\mathrm{ES}(Y|X)$ by solving the right hand side of (B.2) for $X$ fixed (which corresponds to the definition of $\hat{r}$ in Algorithm 1).

Notice that, conditionally affine transformations (B.1) are the ones used for "centering and normalizing": typically, one would use $h_2(x,y) = (y - \hat{\mu}(x))/\hat{\sigma}(x)$ where $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ are estimates of the conditional mean and variance of $Y$ given that $X = x$.

It may be convenient to say some additional words about this traditional normalization: if $Z \in L^1_{\mathrm{P}}$ has $\alpha$-quantiles, then integrating the inequality

$$\mathrm{VaR}(Z)\mathbb{1}_{\{Z \geq \mathrm{VaR}\}} \leq Z\mathbb{1}_{\{Z \geq \mathrm{VaR}(Z)\}} \tag{B.3}$$

and applying Hölder's inequality we obtain the following: for every $p \in [1, \infty]$ ($p' = p/(p-1)$)

$$\mathrm{VaR}(Z)(1-\alpha) \leq ||Z||_{\mathrm{P},p}(1-\alpha)^{1/p'}. \tag{B.4}$$

Now, if $F_Z(t) := \mathrm{P}[Z \leq t]$ is continuous and strictly increasing in $[\mathrm{VaR}(Z), \mathrm{VaR}(Z)+\delta)$ (for some $\delta > 0$) then

$$-\mathrm{VaR}_\alpha(Z) = \mathrm{VaR}_{(1-\alpha)}(-Z)$$

where $\mathrm{VaR}_\beta(\cdot)$ indicates the corresponding VaR at level $\beta$ (Definition 2.1), and the previous argument with $-Z$ in place of $Z$ and $1-\alpha$ in place of $\alpha$ leads to

$$-\mathrm{VaR}(Z)\alpha \leq ||Z||_{\mathrm{P},p}\alpha^{1/p'}. \tag{B.5}$$

Interpreting (B.3), (B.5) in a conditional context and going back to our conventions we obtain that if $p > 1$ and $F_{Y|X}$ is continuous and increasing in $[\mathrm{VaR}(Y|X), \mathrm{VaR}(Y|X) + \delta(X))$ then

$$-\alpha^{-1} \leq \frac{(\mathrm{VaR}(Y|X))^p}{\mathrm{E}\left[|Y|^p|X\right]} \leq (1-\alpha)^{-1} \tag{B.6}$$

which combined with the identity[11]

$$\mathrm{ES}(Y|X) = (1-\alpha)^{-1}\int_\alpha^1 \mathrm{VaR}_\beta(Y|X)\, d\beta \tag{B.7}$$

[11] The equality (B.7) is known as Acerbi's formula. It was generalized to the case of noncontinuous distributions in (Acerbi and Tasche, 2002, Proposition 3.2). For the case in consideration a quick proof follows by the change of variable $y = F^{-1}_{Y|X}(\beta) = \mathrm{VaR}_\beta(F_{Y|X})$ in (2.3).

gives that

$$-p'(1-\alpha^{1/p'})(1-\alpha)^{-1} \leq \frac{\mathrm{ES}(Y|X)}{(\mathrm{E}\left[|Y|^p|X\right])^{1/p}} \leq p'(1-\alpha)^{-1/p}. \tag{B.8}$$

provided that $F_{Y|X}$ is strictly increasing and continuous in $[\mathrm{VaR}(Y|X), \infty)$.

The inequalities (B.4), (B.6) and (B.8) carry at least two important messages: first, the integrability properties of $Y$ are inherited by $\mathrm{VaR}(Y|X)$ and $\mathrm{ES}(Y|X)$ ($\mathrm{E}\left[|Y|^p\right] = \int ||Y||^p_{\mathrm{P}_x,p} \mathrm{P}_X(\mathrm{d}x)$); and second, the (conditional) moments of $Y$ control the value of these risk measures. It follows in particular that if $x \mapsto \hat{M}_p(x) > 0$ is (say) an estimate of $x \mapsto M_p(x) := ||Y||_{\mathrm{P}_x,p}$ and $C > 0$ is a constant such that

$$\mathrm{P}\left[M_p(X) \leq C\hat{M}_p(X)\right] = 1, \tag{B.9}$$

then the specification in Algorithm 1 given by

$$h_1(x,y) = h(y/\hat{M}_p(x))$$

where $h(y)$ is a continuous and increasing bounded function equal to the identity if

$$|y| \leq C(\alpha \wedge (1-\alpha))^{-1/p},$$

permits to assume that

$$B_1 = C(\alpha \wedge (1-\alpha))^{-1/p},$$

giving (by the definition of $h$) that

$$\hat{q}(x) = \hat{M}_p(x)\hat{f}(x).$$

As for the computation of ES$-$VaR, choosing the conditionally affine transformation

$$h_2(x,y) = y/\hat{M}_p(x)$$

permits to fix the bound

$$C(p'(1-\alpha)^{-1/p} + \alpha^{-1/p}). \tag{B.10}$$

for the hypotheses $\mathcal{G}$ and to truncate by any $B_3$ larger than or equal to (B.10) when carrying the regression in Step 4.

Following this line of reasoning, notice that the truncation by $B_3$ gives rise to a "tail error" of the form

$$\mathrm{E}\left[((|W| - B_3)^+)^2\right], \tag{B.11}$$

where $W = (1-\alpha)^{-1}(h_2(X,Y) - h_2(X, q(X)))^+$ is the random variable whose conditional expectation (given $X$) we are trying to estimate. To justify our belief in the necessity of *a priori* controls on tail bounds on $W$ (or $W|X$) for the estimation of ES (e.g. upper bounds to (B.11)), consider the following:

**Claim.** *For every strictly increasing, integrable distribution function $F$ and every $(C, \delta) \in \mathbb{R} \times (0, \infty)$, there exists an increasing and integrable distribution function $G$ coinciding with $F$ in $(-\infty, C]$ and such that* $\mathrm{ES}(F) + \delta < \mathrm{ES}(G)$[12].

According to this claim, no inference can be made in general about $\mathrm{ES}(F)$ only from information on $F(y)$ up to some upper bound $y \leq C < \infty$. Being this is the only kind of information available through finite observations $Y_1(\omega), \ldots, Y_n(\omega)$ of $Y \sim F$, it is not possible in general to infer statistical bounds on the approximation error for estimations of $\mathrm{ES}(F)$ which are based only on finite samples of $F$.[13]

## References


Abbas-Turki, L., S. Crépey, and B. Saadeddine (2022). Pathwise CVA regressions with oversimulated defaults. *Mathematical Finance*. Forthcoming (preprint on https://perso.lpsm.paris/∼crepey).

Acerbi, C. and D. Tasche (2002). On the coherence of expected shortfall. *Journal of Banking and Finance 26*, 1487–1503.

Albanese, C., S. Crépey, R. Hoskinson, and B. Saadeddine (2021). XVA analysis from the balance sheet. *Quantitative Finance 21*(1), 99–123.

Andersen, L., M. Pykhtin, and A. Sokol (2017). Rethinking the margin period of risk. *Journal of Credit Risk 13*(1), 1–45.

Barrera, D. (2022). Confidence intervals for nonparametric regression. arXiv:2203.10643.

Barrera, D., S. Crépey, B. Diallo, G. Fort, E. Gobet, and U. Stazhynski (2019). Stochastic approximation schemes for economic capital and risk margin computations. *ESAIM: Proceedings and Surveys 65*, 182–218.

Bondell, H., B. Reich, and H. Wang (2010). Noncrossing quantile regression curve estimation. *Biometrika 97*(4), 825–838.

Cannon, A. J. (2018). Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic environmental research and risk assessment 32*(11), 3207–3225.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics 6*, 5549–5632.


[12] This can be proved easily via the following observation: assume without loss of generality that $\mathrm{VaR}(F) < C$, consider a random variable $Y$ with the distribution $F$ and random variables

$$Y_k := Y \mathbb{1}_{\{Y \leq C\}} + (C + 2^k (Y - C)) \mathbb{1}_{\{Y > C\}},$$

and notice that $\lim_k \mathrm{ES}(Y_k) = \infty$ by the monotone convergence theorem. The sought for $G$ corresponds to some of these $Y_k$.

[13] This is also an obstruction to obtaining in general, from finite samples of $(X, Y)$, a function satisfying (B.9): we have seen that this implies bounds for ES in the case of continuous distributions.

Dimitriadis, T. and S. Bayer (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics 13*(1), 1823–1871.

Fissler, T. and J. Ziegel (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics 44*(4), 1680–1707.

Fissler, T., J. Ziegel, and T. Gneiting (2016). Expected shortfall is jointly elicitable with value at risk—implications for backtesting. *Risk Magazine*, January.

Gasthaus, J., K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski (2019). Probabilistic forecasting with spline quantile function rnns. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1901–1910. PMLR.

Hatalis, K., A. J. Lamadrid, K. Scheinberg, and S. Kishore (2017). Smooth pinball neural network for probabilistic forecasting of wind power. *arxiv:1710.01720*.

He, X. (1997). Quantile curves without crossing. *The American Statistician 51*(2), 186–192.

Kallenberg, O. (2006). *Foundations of modern probability*. Springer.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics 9*, 155–176.

Koenker, R. and B. J. Park (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics 71*(1-2), 265–283.

Meinshausen, N. and G. Ridgeway (2006). Quantile regression forests. *Journal of Machine Learning Research 7*(6).

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018). *Foundations of machine learning*. MIT press.

Moon, S. J., J.-J. Jeon, J. S. H. Lee, and Y. Kim (2021). Learning multiple quantiles with neural networks. *Journal of Computational and Graphical Statistics*, 1–11.

Padilla, O. H. M., W. Tansey, and Y. Chen (2020). Quantile regression with relu networks: Estimators and minimax rates. *arXiv:2010.08236*.

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

Rockafellar, R. T. and J. O. Royset (2013). Superquantiles and their applications to risk, random variables, and regression. In *Theory Driven by Influential Applications*, pp. 151–167. Informs.

Rodrigues, F. and F. C. Pereira (2020). Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems. *IEEE transactions on neural networks and learning systems 31*(12), 5377–5389.

Sangnier, M., O. Fercoq, and F. d'Alché Buc (2016). Joint quantile regression in vector-valued RKHSs. In *Neural Information Processing Systems*.

Shen, G., Y. Jiao, Y. Lin, J. L. Horowitz, and J. Huang (2021). Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv:2107.04907*.

Takeuchi, I., Q. Le, T. Sears, and A. Smola (2006). Nonparametric quantile estimation. *7*, 1231–1264.