

An Integrated Approach to Importance Sampling and Machine Learning for Efficient Monte Carlo Estimation of Distortion Risk Measures in Black Box Models

Sören Bettels Stefan Weber
Leibniz Universität Hannover

April 22, 2026*

Abstract

Distortion risk measures are important tools for quantifying downside risk under uncertainty. Their accurate estimation becomes challenging when the underlying loss variable is generated by a computationally expensive simulation model without analytical tractability. In this paper, we develop an importance sampling method for the efficient Monte Carlo estimation of distortion risk measures in such black-box models. The method combines importance sampling with machine learning approximations in order to reduce computational cost. Its performance is illustrated in numerical experiments for several distortion risk measures and simulation models.

Keywords: distortion risk measures; importance sampling; quantile estimation; asset-liability management; monetary risk measures

1 Introduction

Many real-world simulation models are highly complex. Random inputs are mapped to outputs through functions whose evaluation is computationally expensive. Such models often form the basis for risk measurement and risk control in firms and other systems, where a careful analysis of rare events is essential. Important industry examples are the internal models used by banks and insurance companies for risk management and solvency regulation.

The goal of this paper is to develop an importance sampling (IS) algorithm for the computation of distortion risk measures (DRMs), an important class of downside risk measures, when the mapping from model inputs to outputs is computationally expensive. We refer to such models as *black-box models*. They are characterized by high computational complexity and, in some cases, by an opaque input–output structure. Our approach combines machine

*Corresponding author: Stefan Weber, House of Insurance & Institute of Actuarial and Financial Mathematics, Leibniz Universität Hannover, Welfengarten 1, 30167 Hannover, Germany. e-mail: stefan.weber@insurance.uni-hannover.de

learning techniques (ML) with two main ingredients: efficient IS for quantile estimation, as developed by Glynn [1996] and Ahn and Shyamalkumar [2011], and representations of distortion risk measures as mixtures of quantiles; see Dhaene et al. [2012].

The quantitative analysis of downside risk has been studied systematically since the 1990s. Foundational axiomatic contributions were provided by Artzner et al. [1999], Föllmer and Schied [2002], and Frittelli and Gianin [2002]. An important and broad class within this framework is given by distortion risk measures (DRMs). It includes many distribution-based coherent risk measures, but also widely used non-convex risk measures. Prominent examples are value at risk, average value at risk, also known as expected shortfall, and range value at risk. The class also contains Wang's premium principles; see Wang [1995] and Wang [1996]. The main contributions of the paper are as follows:

- (i) We develop an importance sampling algorithm, combined with machine learning techniques, for the Monte Carlo estimation of DRMs in black-box models. The construction of the sampling scheme requires a discretization of the DRM into a mixture of quantiles, suitable measure changes for quantiles at different levels, and an efficient allocation of the available sampling budget across these levels. Machine learning is used to approximate the black-box model and thereby reduce computational cost.
- (ii) We analyze the performance of the method in a range of case studies. For DRMs that place substantial weight on very extreme tail events, we also propose and test an iterative refinement of the algorithm. Finally, we apply the method to a simple asset-liability management model of an insurance company.

Literature

More extensive treatments of risk measures and distortion risk measures can be found in Föllmer and Schied [2016] and Föllmer and Weber [2015]. DRMs are closely related to Choquet integrals, introduced by Choquet [1954] and studied in detail by Denneberg [1994]. They have been investigated, for example, in Wang [1995], Wang [1996], Kusuoka [2001], Acerbi [2002], Dhaene et al. [2006], Song and Yan [2006], Song and Yan [2009a], Song and Yan [2009b], Weber [2018], and Kim and Weber [2022]. The representation theorem for DRMs used in this paper is taken from Dhaene et al. [2012].

Standard references on Monte Carlo simulation and importance sampling include Glasser-

man [2003] and Asmussen and Glynn [2007]. Importance sampling techniques for rare-event simulation are developed, for example, in Rubino and Tuffin [2009], Bucklew [2004], Blanchet and Glynn [2008], Dupuis and Wang [2004], Hult and Nyquist [2016], Asmussen et al. [2000], and Juneja and Shahabuddin [2006].

More closely related to the present paper are the following contributions. The asymptotic properties of the IS quantile estimators used here are studied in Glynn [1996] and extended in Ahn and Shyamalkumar [2011]. Glynn [1996] investigates the IS estimation of quantiles and proposes four estimators for which asymptotic normality is established. In applications, the choice of IS distributions from an exponential family is motivated by large-deviations arguments. Building on this work, Ahn and Shyamalkumar [2011] study IS for $V@R$ and $AV@R$ and prove asymptotic normality under weaker assumptions. Arief et al. [2021] consider rare-event simulation in black-box systems, with a focus on probability estimation. Glasserman et al. [2002] analyze the IS estimation of $V@R$ for heavy-tailed risk factors using exponential measure changes. Dunkel and Weber [2007] study the estimation of utility-based shortfall risk by combining stochastic approximation and IS. Brazauskas et al. [2008] consider the estimation of conditional value at risk without using IS; among other results, they prove consistency of the estimator and construct confidence intervals.

Sun and Hong [2009] study IS for value at risk and average value at risk, exploiting the OCE representation of average value at risk due to Rockafellar and Uryasev [2000] and Rockafellar and Uryasev [2002]; see also Ben-Tal and Teboulle [2007]. Their measure changes are selected from an exponential family. Beutner and Zähle [2010] develop a modified functional delta method for the estimation of DRMs and derive asymptotic distributions and approximate confidence intervals from a primarily statistical perspective. They do not consider IS. Pandey et al. [2021] combine a trapezoidal rule with quantile estimation in order to estimate spectral risk measures and prove error bounds in probability. IS is not considered there either. Estimators of DRMs are also closely related to L -estimators; for background, see Stigler [1974] and Serfling [1980].

Surveys on machine learning techniques and their applications include Shalev-Shwartz and Ben-David [2014] and Mohri et al. [2018]. Applications of black-box models in finance are discussed, for example, in Huang et al. [2020].

Outline

The paper is organized as follows. Section 2 introduces distortion risk measures, the quantile estimators used in the paper, and their asymptotic distribution, and develops the proposed importance sampling method for DRMs. Section 3 studies the performance of the method in a range of case studies. Section 4 presents an application to a simple asset-liability management model of an insurance firm. Auxiliary material is collected in an online appendix, including background on distortion risk measures, asymptotic results for IS quantile estimators, a brief review of the machine learning tools used in the paper, supplementary computations and proofs, and additional figures based on the case studies.

2 Efficient Estimation of DRMs in Black Box Models

2.1 Setting the Scene

Accurate risk measurement in complex systems is an important task. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an atomless probability space and let $X : \Omega \rightarrow \mathbb{R}^d$ be a random vector. The random output of the system is modeled by $Y = h(X)$, where $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is a measurable function. We assume that Y can be generated by simulation, but that h is highly complex and analytically intractable. By contrast, the distribution of the input vector X is explicitly known and can be modified in order to improve estimation efficiency. In particular, we assume that h can still be evaluated under such modified input distributions, although each evaluation remains computationally expensive.

The problem is to determine $\rho(h(X))$ by simulation, where ρ is a monetary risk measure. We use the actuarial sign convention under which $h(X)$ records losses as positive and gains as negative. More specifically, we assume that $\rho = \rho_g$ is a DRM associated with a distortion function $g : [0, 1] \rightarrow [0, 1]$ of the form $\rho_g(Y) = \int_{-\infty}^0 [g(\mathbb{P}(Y > y)) - 1] dy + \int_0^{\infty} g(\mathbb{P}(Y > y)) dy$. For further details, we refer to Appendix A.1. By Dhaene et al. [2012], see also Bettels et al. [2024], DRMs can be written as mixtures of quantiles, i.e.,

$$\rho_g(Y) = c_1 \int_{[0,1]} q_Y^+(1-u) dg_1(u) + c_2 \int_{[0,1]} q_Y(1-u) dg_2(u), \quad (1)$$

where g_1 and g_2 are right- and left-continuous distortion functions, respectively, $c_1 + c_2 = 1$, $c_1, c_2 \in [0, 1]$, $g = c_1 g_1 + c_2 g_2$, and $q_Y^+(u) = \sup\{y \mid F_Y(y) \leq u\}$, $q_Y(u) = \inf\{y \mid F_Y(y) \geq u\}$. Equation (1) is the starting point of the Monte Carlo simulation scheme.

The risk estimation problem has two aspects. First, the quantiles appearing as integrands in eq. (1) must be estimated efficiently. Second, the integrals, i.e., the mixtures of quantiles, must be discretized. We address the first issue by developing an importance sampling technique for quantile estimation based on a machine learning approximation of the function h . We address the second by constructing a sample allocation rule along the discretization that leads to good performance. The next sections explain the design and implementation of the resulting Monte Carlo algorithm for the estimation of DRMs. Algorithm 1 implements a more specific version of this general approach. In particular, $\alpha \in (0, 1)$ denotes a user-specified tail threshold that determines the quantile region on which the discretization is concentrated.

2.2 Quantile Estimation with Importance Sampling

We begin with a quantile estimation technique based on importance sampling, as proposed and studied in Glynn [1996] and Ahn and Shyamalkumar [2011]. In this section, we first assume that the function h is known. The incorporation of machine learning techniques is discussed later in Section 2.4.

Let F denote the distribution function of X , and let F^* be another distribution function on \mathbb{R}^d such that F is absolutely continuous with respect to F^* . We are interested in the u -quantile of $Y = h(X)$, $u \in (0, 1)$, when X has distribution function F . If $(X_i)_{i=1, \dots, N}$ are sampled independently from F^* , then an importance sampling estimator of $q_Y(u)$ is given by

$$\hat{q}_{F^*, N}(u) := \inf \left\{ x \in \mathbb{R} \mid \frac{1}{N} \sum_{h(X_i) > x} \frac{dF}{dF^*}(X_i) \leq 1 - u \right\}, \quad u \in (0, 1). \quad (2)$$

Conditions for the asymptotic normality of this estimator are given in Glynn [1996] and Ahn and Shyamalkumar [2011]; for convenience, we state them in Appendix A.4. More precisely, let G and G^* denote the distribution functions of $Y = h(X)$ when X has distribution function F and F^* , respectively. Then Ahn and Shyamalkumar [2011] prove the following result.

Theorem 2.1. *Suppose that Assumption A.12 holds. Then, for every $u \in (0, 1)$,*

$$\sqrt{N}(\hat{q}_{F^*, N}(u) - q_Y(u)) \xrightarrow{d} \mathcal{N} \left(0, \frac{\mathbf{E}_{F^*} \left[\left(\frac{dF}{dF^*}(X) \right)^2 \mathbb{1}_{\{h(X) > q_Y(u)\}} \right] - (1 - u)^2}{G'(q_Y(u))^2} \right)$$

as $N \rightarrow \infty$.

Algorithm 1 Importance Sampling DRM Estimation Algorithm

1: **Input:** Distortion function g , threshold α , pivot sample size M , sample size N , size of partition m .

2: **Output:** Estimation of $\rho_g(Y)$

3: **function** MAIN:

4: Set $\alpha_i = i\alpha/m$ for $i \in \{0, \dots, m\}$ and $\alpha_{m+1} = 1$;

5: Sample $X \leftarrow (X_1, \dots, X_M)$ from F and set $Y \leftarrow (h(X_1), \dots, h(X_M))$;

6: **for** $i \in \{0, \dots, m\}$ **do**

7: Set $aux \leftarrow$ empirical quantile of sample Y at level $1 - \alpha_i$;

8: Set ϑ_i such that

$$aux = \frac{\sum_{j=1}^M Y_j \exp(\vartheta_i Y_j)}{\sum_{j=1}^M \exp(\vartheta_i Y_j)};$$

9: **for** $i \in \{0, \dots, m\}$ **do**

10: Set $aux \leftarrow$ empirical quantile of sample Y at level $1 - \alpha_i$;

11: Set $aux_c \leftarrow \frac{1}{M} \sum_{j=1}^M \frac{dF}{dF_{\vartheta_i}}(X_j) \mathbf{1}_{\{Y_j > aux\}}$;

12: Set c_i such that

$$c_i \leftarrow \frac{aux_c - \alpha_i^2}{G'(aux)^2} \cdot (g(\alpha_{i+1}) - g(\alpha_i));$$

13: **for** $i \in \{0, \dots, m\}$ **do**

14: Set

$$p_i \leftarrow \frac{\sqrt{c_i}}{\sum_{j=0}^m \sqrt{c_j}};$$

15: Choose \hat{h} as the regression selected by a k -fold validation and calibration from X, Y ;

16: Set F_i for $i \in \{0, 1, \dots, m\}$ such that

$$dF_i = \exp\left(\vartheta_i \hat{h}(x) - \hat{\psi}(\vartheta_i)\right) dF;$$

17: Sample $\theta_1, \dots, \theta_N$ as i.i.d. copies of θ such that $\mathbf{P}(\theta = i) = p_i$ for $i \in \{0, 1, \dots, m\}$;

18: Sample $X' \leftarrow (X'_1, \dots, X'_N)$ such that $X'_i \sim F_{\theta_i}$ and set $Y' \leftarrow (h(X'_1), \dots, h(X'_N))$;

19: Set $estimate \leftarrow 0$;

20: **for** $i \in \{0, \dots, m\}$ **do**

21: **Option 1:** Compare the variances of $\hat{q}_{F_i, N_i}(1 - \alpha_i)$ and $\hat{q}_{F^*, N}(1 - \alpha_i)$;

22: Set $\hat{q}_Y(1 - \alpha_i)$ as the better performing estimator;

23: **Option 2:** Set $\hat{q}_Y(1 - \alpha_i) \leftarrow \hat{q}_{F^*, N}(1 - \alpha_i)$;

24: Set $estimate \leftarrow estimate + \hat{q}_Y(1 - \alpha_i) \cdot (g(\alpha_{i+1}) - g(\alpha_i))$;

25: **Return:** $estimate$;

Theorem 2.1 can now be used to construct a sampling distribution F^* that improves the efficiency of the Monte Carlo simulation. A classical choice for large rare outcomes is given by exponential tilting. In our setting, X is sampled, while the quantity of interest is the upper tail of $Y = h(X)$. We therefore consider the family of sampling distributions

$$dF_{\vartheta}(x) = \exp(\vartheta h(x) - \psi(\vartheta)) dF(x), \quad (3)$$

where $\vartheta \in \Theta \subseteq \mathbb{R}$ for some suitable neighborhood Θ of 0, and $\psi(\vartheta) := \log(\mathbf{E}_F[\exp(\vartheta h(X))])$.

To reduce the asymptotic variance in Theorem 2.1, Sun and Hong [2009] minimize a suitable upper bound and obtain the condition

$$q_Y(u) = \mathbf{E}_{F_\vartheta}[h(X)], \quad (4)$$

which is used to select a suitable parameter ϑ . Under appropriate technical assumptions, they show that the resulting measure change reduces the variance of the estimator. For convenience, we review this argument in Appendix A.4. Implementing eq. (4), however, requires knowledge of the target quantile $q_Y(u)$ itself. Moreover, in our black-box setting, the exact structure of h and $\psi(\vartheta)$ is not available in tractable form. Section 2.4 therefore develops an algorithmic approach based on machine learning and MCMC to address these difficulties.

2.3 Discretization and Optimal Allocation of the Sampling Budget

The estimation of eq. (1) requires a discretization of the two integrals involving the left- and right-continuous distortion functions. This distinction matters only if $q_Y^+(1-u) \neq q_Y(1-u)$ for some u . In the numerical implementation, we assume that $q_Y^+(1-u) = q_Y(1-u)$ at all quantile levels appearing in the discretization. This is consistent with Theorem 2.1 and Assumption A.12, which underlie the quantile approximation used below. What is needed is that the distribution function of $Y = h(X)$ be locally increasing at the quantile levels under consideration, so that the relevant quantiles are unique. A sufficient condition is that Y admit a density that is strictly positive in a neighborhood of these quantiles. This is an assumption on the law of Y , not directly on the law of X . If the distribution function has a flat part containing one of the quantile levels under consideration, then the corresponding lower and upper quantiles differ. This occurs in particular for discrete distributions, whose distribution functions are constant between jumps at the atoms.

We consider the approximation

$$\hat{\rho}_g(Y) = \sum_{i=0}^m \hat{q}_{F_{\vartheta_i^*}, N_i}(1 - \alpha_i)(g(\alpha_{i+1}) - g(\alpha_i)) \quad (5)$$

of $\rho_g(Y)$, based on a partition $0 = \alpha_0 < \alpha_1 < \dots < \alpha_m < \alpha_{m+1} = 1$. Here $\hat{q}_{F_{\vartheta_i^*}, N_i}(1 - \alpha_i)$ are the importance sampling quantile estimators from Section 2.2, constructed from the sampling distributions $F_{\vartheta_i^*}$ and the allocated sample sizes N_i . For instance, the partition $(\alpha_i)_{i=0, \dots, m+1}$

may be chosen uniformly over the region where g increases, or one may take $\alpha_i = g^{-1}\left(\frac{i}{m+1}\right)$, $i = 0, \dots, m+1$, in order to place more grid points where g assigns more weight.

We assume that Assumption A.12 holds and that, for each i , the sample size N_i is large enough to ensure that $\hat{q}_{F_{\vartheta_i^*}, N_i}(1 - \alpha_i)$ is finite according to eq. (8). Define the estimator of the quantile function by

$$\hat{q}_Y(1 - u) := \sum_{i=0}^m \mathbb{1}_{\{u \in [\alpha_i, \alpha_{i+1})\}} \hat{q}_{F_{\vartheta_i^*}, N_i}(1 - \alpha_i).$$

Then $\hat{\rho}_g(Y) = \int_0^1 \hat{q}_Y(1 - u) dg(u)$. This leads to two questions. First, how should the total sample size N be allocated across the quantiles at the different levels? Second, should one use individual importance sampling for each quantile, or a single common measure change with pooled samples?

2.3.1 Sample Allocation to Quantiles

Using Jensen's inequality, Fubini's theorem and Theorem 2.1, the MSE of the estimator can approximately be bounded above as follows (see also Section A.7.1):

$$\begin{aligned} \mathbb{E} [(\rho_g(Y) - \hat{\rho}_g(Y))^2] &= \mathbb{E} \left[\left(\int_0^1 (q_Y(1 - u) - \hat{q}_Y(1 - u)) dg(u) \right)^2 \right] \leq \int_0^1 \mathbb{E} [(q_Y(1 - u) - \hat{q}_Y(1 - u))^2] dg(u) \\ &\approx \sum_{i=0}^m \int_{\alpha_i}^{\alpha_{i+1}} (q_Y(1 - u) - q_Y(1 - \alpha_i))^2 dg(u) + \underbrace{\frac{\mathbb{E}_{\vartheta_i^*} \left[\frac{dF}{dF_{\vartheta_i^*}}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1 - \alpha_i))^2}}_{:= \frac{V(1 - \alpha_i, F_{\vartheta_i^*})}{N_i}} (g(\alpha_{i+1}) - g(\alpha_i)) \\ &=: \mathcal{E}(F_{\vartheta_0^*}, F_{\vartheta_1^*}, \dots, F_{\vartheta_m^*}, \tilde{N}) \text{ with } \tilde{N} = (N_i)_{i=0,1,\dots,m}. \end{aligned} \quad (6)$$

In the latter sum only the second summand $\frac{V(1 - \alpha_i, F_{\vartheta_i^*})}{N_i} (g(\alpha_{i+1}) - g(\alpha_i))$ depends on N_i . Hence, when minimizing the approximate upper bound, the optimal allocation is obtained by minimizing $\sum_{i=0}^m \frac{V(1 - \alpha_i, F_{\vartheta_i^*})}{N_i} (g(\alpha_{i+1}) - g(\alpha_i))$ under the constraint $\sum_{i=0}^m N_i = N$. This leads (up to rounding) to the solution

$$N_i^* = N \frac{\sqrt{c_i}}{\sum_{j=0}^m \sqrt{c_j}}, \quad i = 0, 1, \dots, m, \quad (7)$$

where $c_j := V(1 - \alpha_j, F_{\vartheta_j^*}) (g(\alpha_{j+1}) - g(\alpha_j))$, $j \in \{0, 1, \dots, m\}$. The derivation of this result can be found in Appendix A.7.1.

If the total sample size N is not known in advance, eq. (7) determines the fraction $p_i := \frac{N_i^*}{N}$

of the samples generated for each quantile. The total collection of all samples for these quantiles can also be viewed as samples from the mixture sampling distribution $F^* := \sum_{i=0}^m p_i \cdot F_{\vartheta_i^*}$, where $F_{\vartheta_i^*}$ are the sampling distributions for each individual quantile constructed in Section 2.2.

2.3.2 Efficient Use of the Samples in the Estimation of Multiple Quantiles

The estimation of DRMs according to eq. (5) requires the estimation of quantiles at the levels $1 - u$ for $u = \alpha_0, \dots, \alpha_m$. We discuss whether individual importance sampling should be used, or a single common measure change for pooled samples is preferred. We assume in our comparison that the generation of individual samples is costly, but that the evaluation of the quantile estimators for given samples is comparatively inexpensive. We suppose that samples are allocated to the individual quantiles according to eq. (7) and that F^* is the mixture sampling distribution in Section 2.3. For each i , estimators of $q_Y(1 - \alpha_i)$ are $\hat{q}_{F_{\vartheta_i^*}, N_i^*}(1 - \alpha_i)$ and $\hat{q}_{F^*, N}(1 - \alpha_i)$ with approximate variances

$$\begin{aligned} \frac{V(1 - \alpha_i, F_{\vartheta_i^*})}{N_i^*} &= \frac{\mathbb{E}_{F_{\vartheta_i^*}} \left[\frac{dF}{dF_{\vartheta_i^*}}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2}{p_i \cdot N \cdot G'(q_Y(1 - \alpha_i))^2}, \\ \frac{V(1 - \alpha_i, F^*)}{N} &= \frac{\mathbb{E}_{F^*} \left[\frac{dF}{dF^*}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2}{NG'(q_Y(1 - \alpha_i))^2}, \end{aligned}$$

respectively, if the conditions of Theorem 2.1 are satisfied. Hence, by comparing

$$\mathbb{E}_F \left[\frac{dF}{dF_{\vartheta_i^*}}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2 \quad \text{to} \quad p_i \cdot \left(\mathbb{E}_F \left[\frac{dF}{dF^*}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2 \right),$$

the preferred estimator can be selected. For simplicity, we propose using the mixture distribution for all quantile levels and implement it in all case studies in Sections 3 & 4. The corresponding estimator is

$$\hat{\rho}_g(Y) = \sum_{i=0}^m \hat{q}_{F^*, N}(1 - \alpha_i)(g(\alpha_{i+1}) - g(\alpha_i)).$$

2.4 Machine Learning and Implementation

The objective of the simulation is to estimate $\rho_g(Y)$ for $Y = h(X)$, with the main challenge being the high cost of evaluating h . In order to apply importance sampling to this situation, we propose an algorithm with the following steps: First, pivot samples are used to compute parameters that govern importance sampling on the basis of exponential changes of measure. Second, the costly function h is approximated by some auxiliary function \hat{h} that simplifies the

measure changes and allows to accelerate the generation of the corresponding samples. This step typically involves acceptance-rejection methods where the auxiliary function \hat{h} allows to avoid the costly evaluation of $Y = h(X)$. Third, quantile estimators are computed for these samples for the original random variable $Y = h(X)$. We discuss additional design nuances and details of the approach in Appendix A.5. Let $(X_1, h(X_1)), \dots, (X_M, h(X_M))$ be pivot samples. According to eq. (4), an estimator $\hat{\vartheta}_i^*$ of ϑ_i^* can be obtained by solving $\hat{q}_{F,M}(1 - \alpha_i) = \frac{\sum_{j=1}^M h(X_j) \exp(\hat{\vartheta}_i^* h(X_j))}{\sum_{j=1}^M \exp(\hat{\vartheta}_i^* h(X_j))}$ numerically; here, $\hat{q}_{F,M}(1 - \alpha_i)$ signifies the crude quantile estimator. A plug-in estimator of $\psi(\vartheta_i^*)$ is, for example, given by $\hat{\psi}_i = \log\left(\frac{1}{M} \sum_{j=1}^M \exp(\hat{\vartheta}_i^* h(X_j))\right)$.

To facilitate the generation of additional samples, we use ML-techniques trained on the pivot samples to approximate h by a less costly function \hat{h} . More specifically, we consider linear and polynomial predictors, linear, polynomial and Gaussian support vector machines (SVMs) and k -nearest neighbor (k -NN) regressions in numerical case studies. To compare methods and parameters, k -fold validation is applied; we determine an approximation that yields the smallest MSE across splits of the training sample. For a brief review of the methods, see Appendix A.6 and Shalev-Shwartz and Ben-David [2014].

Importance sampling is based on the approximation \hat{h} , i.e., in (3) we use a Radon-Nikodym density proportional to $\exp(\hat{\vartheta}_i^* \hat{h} - \hat{\psi}_i) =: \hat{f}_{\hat{\vartheta}_i^*}$. The latter function might not itself be a probability density due to a potentially incorrect normalization, since $\hat{\psi}_i$ was estimated from h instead of \hat{h} . The correct normalization constant could be estimated at this stage, or Markov Chain Monte Carlo (MCMC) can directly be used to generate more samples. Unless the proposal kernel is the independence kernel, MCMC typically does not preserve the independence of simulations, but appears to be quite efficient in our case studies. The Metropolis-Hastings algorithm with random walk proposal allows to either produce samples from a density proportional to $\hat{f}_{\hat{\vartheta}_i^*}$ or proportional to the mixture $p_0 \hat{f}_{\hat{\vartheta}_0^*} + \dots + p_m \hat{f}_{\hat{\vartheta}_m^*}$ with $(p_i)_{i=0,1,\dots,m}$ according to Section 2.3.1.

In the implementation, the mixture weights are approximated by plug-in estimates \hat{p}_i based on the pivot sample. More precisely, after estimating $\hat{\vartheta}_i^*$ and the crude quantile $\hat{q}_{F,M}(1 - \alpha_i)$, we set

$$\hat{a}_i = \frac{1}{M} \sum_{j=1}^M \frac{dF}{dF_{\hat{\vartheta}_i^*}}(X_j) \mathbb{1}_{\{h(X_j) > \hat{q}_{F,M}(1 - \alpha_i)\}},$$

estimate $G'(\hat{q}_{F,M}(1 - \alpha_i))$ from the pivot sample, and define the plug-in quantities

$$\hat{c}_i = \frac{\hat{a}_i - \alpha_i^2}{\widehat{G'}(\hat{q}_{F,M}(1 - \alpha_i))^2} (g(\alpha_{i+1}) - g(\alpha_i)), \quad \hat{p}_i = \frac{\sqrt{\hat{c}_i}}{\sum_{j=0}^m \sqrt{\hat{c}_j}}.$$

In the numerical implementation used for the reported experiments, the plug-in weights were computed with $\widehat{G}'(\hat{q}_{F,M}(1-\alpha_i))$ in the denominator rather than $\widehat{G}'(\hat{q}_{F,M}(1-\alpha_i))^2$. Accordingly, the empirical mixture weights reported in the simulations correspond to this plug-in variant of the allocation rule. This earlier implementation choice was used consistently throughout all reported experiments and does not materially affect the qualitative conclusions of the numerical study.

For the importance sampling quantile estimator (2) we need to evaluate the likelihood ratio which requires knowledge of the normalizing constants. To be more precise, we have $\frac{dF}{d\hat{F}_{\hat{\vartheta}_i^*}} = \frac{z_i}{\exp(\hat{\vartheta}_i^* \hat{h}(x) - \hat{\psi}_i)}$, $z_i = \int \exp(\hat{\vartheta}_i^* \hat{h}(x) - \hat{\psi}_i) F(dx)$ and $\frac{dF}{d(\sum_{i=0}^m p_i \hat{F}_{\hat{\vartheta}_i^*})} = \frac{z}{\sum_{i=0}^m p_i \exp(\hat{\vartheta}_i^* \hat{h}(x) - \hat{\psi}_i)}$, $z = \sum_{i=0}^m p_i z_i$. The normalizing factors z_i and z could be approximated using simulations, but this might be costly given the desired accuracy. In low dimensions, we can alternatively either use a trapezoidal rule with the $N + M$ samples as grid points or apply an adaptive quadrature rule explained in Shampine [2008]. Since F is the original distribution of the factor X , a suitable original model design might also ensure the applicability of such an approach for high dimensions d . Another strategy for estimating the normalizing constant relies on estimating the density function from the samples drawn; in this case, we assume that F has density f with respect to d -dimensional Lebesgue measure. Consider, for example, the mixture distribution, and let \hat{f}_{mix} be the estimated density, e.g., via kernel density estimation. Then for all $x \in \mathbb{R}^d$ we have that $z \approx \frac{\sum_{i=0}^m p_i \exp(\hat{\vartheta}_i^* \hat{h}(x)) f(x)}{\hat{f}_{\text{mix}}(x)}$. Thus, z can be estimated by computing the right hand side for several x and taking an average. In the implementations we chose for each application the method that performed best in the test cases. While the suggested approach works quite well in the considered numerical experiments, future research needs to further optimize the algorithm to guarantee good performance for high-dimensional random vectors X . A successful strategy could be to choose tractable pairs of ML hypothesis classes on the one hand and the factor sampling distribution F on the other hand that facilitate the implementation of measure changes.

2.5 Time Efficiency of IS and Crude Estimation Methods

The proposed IS algorithm is computationally more involved than a crude estimator of the same DRM. In this section, we identify conditions under which the IS method is nevertheless preferable. More precisely, we analyze the case in which samples are drawn i.i.d. from the

mixture IS distribution and compare the corresponding error bound in eq. (6) with that of crude Monte Carlo; the case of individual IS is deferred to Appendix A.11. The key point is that, if the IS estimator achieves a smaller MSE and the evaluation of h is sufficiently costly, then the resulting gain in statistical efficiency can outweigh the additional computational overhead.

We write $T_{CR}(N_{CR}, m)$ and $T_{IS}(M, N_{IS}, m)$ for the computation times required to estimate the DRM $\rho_g(Y)$ by the crude and IS methods, respectively. Here N_{CR} is the sample size of the crude estimator, while M and N_{IS} denote the pivot and importance sampling sample sizes. The parameter m is the partition size. We assume that N_{CR} , M , and N_{IS} are chosen such that both methods satisfy the same error bound in eq. (6), and that $N_{CR} > M + N_{IS}$. The following lemma states this relation explicitly.

Lemma 2.2. *Let $F^* = \sum_{i=0}^m p_i F_i$, where F_i are distribution functions for $i \in \{0, 1, \dots, m\}$. Assume that F is absolutely continuous with respect to each F_i , and that Assumptions A.12 hold. If samples are drawn i.i.d. from F^* , then*

$$\mathcal{E}(F, \dots, F, \tilde{N}_{CR}) = \mathcal{E}(F^*, \dots, F^*, \tilde{N}_{IS}) \iff N_{CR} = N_{IS} \frac{\sum_{i=0}^m V(1 - \alpha_i, F) (g(\alpha_{i+1}) - g(\alpha_i))}{\sum_{i=0}^m V(1 - \alpha_i, F^*) (g(\alpha_{i+1}) - g(\alpha_i))}.$$

Proof. When all samples from F^* are used for the quantile estimation, we have $\tilde{N}_{CR} = (N_{CR}, \dots, N_{CR})^T$ and $\tilde{N}_{IS} = (N_{IS}, \dots, N_{IS})^T$. Hence:

$$\begin{aligned} \mathcal{E}(F, \dots, F, \tilde{N}_{CR}) &= \mathcal{E}(F^*, \dots, F^*, \tilde{N}_{IS}) \\ \Leftrightarrow \frac{1}{N_{CR}} \sum_{i=0}^m V(1 - \alpha_i, F) (g(\alpha_{i+1}) - g(\alpha_i)) &= \frac{1}{N_{IS}} \sum_{i=0}^m V(1 - \alpha_i, F^*) (g(\alpha_{i+1}) - g(\alpha_i)) \\ \Leftrightarrow N_{CR} &= N_{IS} \cdot \frac{\sum_{i=0}^m V(1 - \alpha_i, F) (g(\alpha_{i+1}) - g(\alpha_i))}{\sum_{i=0}^m V(1 - \alpha_i, F^*) (g(\alpha_{i+1}) - g(\alpha_i))}. \end{aligned}$$

□

For comparing the computational costs of the crude and IS algorithms, we define the following quantities:

$t_S(N)$: computation time required to sample $(X_i)_{i \in \{1, \dots, N\}}$ from F .

$t_h(N)$: computation time required to evaluate $(h(X_i))_{i \in \{1, \dots, N\}}$.

$t_\rho(N, m)$: computation time required to compute the quantile estimates and $\hat{\rho}_g(Y)$.

$t_{\text{MIX}}(M, m)$: computation time required to compute ϑ_i , $\psi(\vartheta_i)$, and p_i for $i \in \{0, \dots, m\}$ from the samples $(X_i)_{i \in \{1, \dots, M\}}$.

$t_{\text{kFold}}(M)$: computation time required for k -fold validation based on the samples $(X_i)_{i \in \{1, \dots, M\}}$.

$t_{\text{MH}}(N)$: computation time required to sample $(X_i)_{i \in \{1, \dots, N\}}$ from the IS distribution using the Metropolis–Hastings algorithm.

$t_{\text{Norm}}(m)$: computation time required to estimate the normalizing constant.

We assume that the computation times are increasing in their arguments, and that $t_S(\cdot)$ and $t_h(\cdot)$ are linear functions. With these definitions, we can express the estimation methods together with their computation times. For the purpose of comparison, the estimation methods are decomposed into the following steps:

Crude Method:

$t_S(N_{CR})$: Draw samples $(X_i)_{i \in \{1, \dots, N_{CR}\}}$ from F .

$t_h(N_{CR})$: Evaluate h to obtain $(h(X_i))_{i \in \{1, \dots, N_{CR}\}}$.

$t_\rho(N_{CR}, m)$: Compute the quantile estimators $\hat{q}_{F, N_{CR}}(1 - \alpha_i)$, $i \in \{0, \dots, m\}$, and the estimate $\hat{\rho}_g(Y)$.

IS Method:

$t_S(M)$: Draw samples $(X_i)_{i \in \{1, \dots, M\}}$ from F .

$t_h(M)$: Evaluate h to obtain $(h(X_i))_{i \in \{1, \dots, M\}}$.

$t_{M_{ix}}(M, m)$: Estimate ϑ_i^* , $\psi(\vartheta_i^*)$, and p_i for $i \in \{0, \dots, m\}$.

$t_{k\text{Fold}}(M)$: Perform k -fold validation based on $(h(X_i))_{i \in \{1, \dots, M\}}$.

$t_{\text{MH}}(N_{IS})$: Generate samples $(X'_i)_{i \in \{1, \dots, N_{IS}\}}$ from the IS distribution.

$t_h(N_{IS})$: Evaluate h to obtain $(h(X'_i))_{i \in \{1, \dots, N_{IS}\}}$.

$t_{\text{Norm}}(m)$: Estimate the normalizing constant.

$t_\rho(N_{IS}, m)$: Compute the quantile estimators $\hat{q}_{F^*, N_{IS}}(1 - \alpha_i)$, $i \in \{0, \dots, m\}$, and the estimate $\hat{\rho}_g(Y)$.

The total computation times of the crude and IS estimations are

$$T_{CR}(N_{CR}, m) := t_S(N_{CR}) + t_h(N_{CR}) + t_\rho(N_{CR}, m),$$

$$T_{IS}(M, N_{IS}, m) := t_S(M) + t_h(M) + t_{M_{ix}}(M, m) + t_{k\text{Fold}}(M) + t_{\text{MH}}(N_{IS})$$

$$\begin{aligned}
& + t_h(N_{IS}) + t_{Norm}(m) + t_\rho(N_{IS}, m) \\
= & t_S(M) + t_h(M + N_{IS}) + t_{Mix}(M, m) + t_{kFold}(M) \\
& + t_{MH}(N_{IS}) + t_{Norm}(m) + t_\rho(N_{IS}, m).
\end{aligned}$$

Proposition 2.3. *Suppose $N_{CR} > M + N_{IS}$. If*

$$t_h(N_{CR} - (M + N_{IS})) > t_{Mix}(M, m) + t_{kFold}(M) + t_{MH}(N_{IS}) + t_{Norm}(m),$$

then

$$T_{CR}(N_{CR}, m) - T_{IS}(M, N_{IS}, m) > 0,$$

that is, the crude method requires more computation time than IS.

Proof. Here we use that $t_S(\cdot)$ and $t_h(\cdot)$ are assumed to be linear, so in particular $t_S(N_{CR}) - t_S(M) = t_S(N_{CR} - M)$ and $t_h(N_{CR}) - t_h(M + N_{IS}) = t_h(N_{CR} - (M + N_{IS}))$. We may write

$$\begin{aligned}
T_{CR}(N_{CR}, m) - T_{IS}(M, N_{IS}, m) &= t_S(N_{CR} - M) + t_h(N_{CR} - (M + N_{IS})) \\
&\quad - [t_{Mix}(M, m) + t_{kFold}(M) + t_{MH}(N_{IS}) + t_{Norm}(m)] \\
&\quad + [t_\rho(N_{CR}, m) - t_\rho(N_{IS}, m)].
\end{aligned}$$

Since $N_{CR} > M + N_{IS}$, the monotonicity of $t_S(\cdot)$ and $t_\rho(\cdot)$ yields $t_S(N_{CR} - M) > 0$ and $t_\rho(N_{CR}, m) - t_\rho(N_{IS}, m) > 0$. Hence, if

$$t_h(N_{CR} - (M + N_{IS})) > t_{Mix}(M, m) + t_{kFold}(M) + t_{MH}(N_{IS}) + t_{Norm}(m),$$

then $T_{CR}(N_{CR}, m) - T_{IS}(M, N_{IS}, m) > 0$, as claimed. \square

3 Case Studies

In this section, we apply the developed method to various test models and distributions. The goal is to experimentally evaluate the variance reduction achieved by the proposed algorithm compared to importance sampling in the exact model, which is known in closed form for the test cases. We compare the root mean square errors (RMSEs) when estimating different DRMs that model both risk-averse and risk-seeking attitudes.

3.1 Simulation Design

We consider the distortion function $g_{\alpha,\gamma}(u) = \mathbb{1}_{\{u \in [0,\alpha]\}} \left(\frac{u}{\alpha}\right)^\gamma + \mathbb{1}_{\{u \in (\alpha,1]\}}$ with $\gamma \in \{1/2, 1, 2\}$ illustrated in Figure 1, see Example A.6 in the Appendix for more details. The concave function $g_{\alpha,1/2}$ defines a convex DRM that models a risk-averse attitude. Conversely, the function $g_{\alpha,2}$ is convex on the interval $[0, \alpha]$ and models a risk-seeking attitude. The function $g_{\alpha,1}$ corresponds to the Average Value at Risk (AV@R) at level α . The AV@R, also known as Expected Shortfall, is particularly important in practice, since it serves as the foundation for various solvency regimes. For additional details and references, see Appendix A.2.

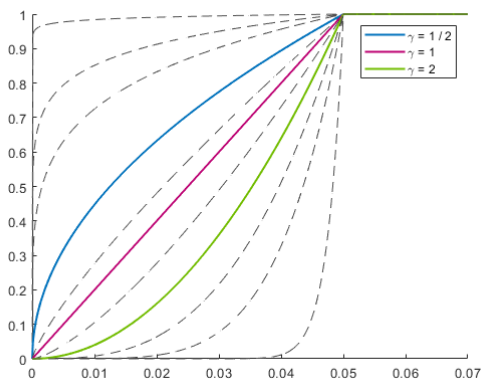


Figure 1: Different distortion functions $g_{\alpha,\gamma}(u)$ with $\alpha = 0.05$ and $\gamma \in \{0.01, 0.1, 0.2, 0.5, 0.8, 1, 1.4, 2, 3, 5, 20\}$. The distortion functions used in the case studies are highlighted.

In our numerical experiments, we repeat Algorithm 1 for DRM estimation R times to obtain data that can be further analyzed. For clear benchmarking, we specify the hypothesis class used in each of these experiments and compare the results across hypothesis classes. Thus, we do not perform the k -fold validation in line 15 of Algorithm 1 in each repetition, but only recalibrate within any previously selected class. Additionally, we identify the winner of a k -fold validation with $k = 20$ based on $M = 2,000$ pivot samples. Numerical tests in the context of our case studies show that this determination of a ML hypothesis class is quite robust, i.e., different sets of $M = 2,000$ pivot samples typically lead to the selection of the same class.

We consider the following functions and distributions:

- (1) *Identity of Normal*: We set $X \sim \mathcal{N}(0, 1)$ and $h(x) = x$, implying $Y \sim \mathcal{N}(0, 1)$. The k -fold cross validation from the pivot samples suggests using a linear regression to approximate h .
- (2) *Sum of Normals*: We consider $X_1, X_2 \sim \mathcal{N}(0, 1)$ with $\text{Corr}(X_1, X_2) = 0.3$ and $h(x_1, x_2) =$

$x_1 + x_2$. The k -fold cross validation suggests a linear regression.

- (3) *Product of Normals*: Let $X_1, X_2 \sim \mathcal{N}(2, 1)$ with $\text{Corr}(X_1, X_2) = -0.3$ and $h(x_1, x_2) = x_1 \cdot x_2$. The k -fold validation identifies the SVM with a polynomial kernel of degree 2 as the optimal choice for \hat{h} .
- (4) *Sum of Squared Normals*: Consider the independent random variables $X_1, X_2, X_3, X_4 \sim \mathcal{N}(0, 1)$ and let $h(x_1, x_2, x_3, x_4) = x_1^2 + x_2^2 + x_3^2 + x_4^2$. Then $h(X_1, X_2, X_3, X_4)$ follows a χ^2 distribution with 4 degrees of freedom. The k -fold validation suggests a polynomial SVM with degree 2.
- (5) *Sine and Uniform*: Set $X \sim \text{unif}(0, 1)$ and $h(x) = x \sin(2.5 \cdot \pi \cdot x)$. This example is used in Altman [1992] to illustrate k -NN regression. The k -fold validation suggests polynomial regression with degree 5 as an approximation of h .
- (6) *Logistic Transformation and Exponential*: Letting $X \sim \exp(1)$, then $h(X)$ with $h(x) = -\log(e^{-x}/(1 - e^{-x}))$ follows a Logistic(0, 1) distribution. The k -fold validation suggests either a SVM with Gaussian kernel or k -NN regression with $k = 1$.

For each of these functions and distributions, we perform numerical experiments for different ML hypothesis classes used to approximate h in the IS algorithms. In particular, this analysis is also performed for the winner of the k -fold validations. Each experiment is repeated R times. In all cases, we implement a crude estimation with $M + N$ samples as well as an estimation based on the IS method with M pivot samples and N samples from the mixture distribution defined in Section 2.3. As a benchmark, we determine an “exact value” by a crude estimation with 10,000,000 samples. From the replications we calculate the estimated RMSE for all cases.

3.2 Results

3.2.1 Distribution of the Samples

To illustrate the measure change, we consider model (5) in Figure 2. As mentioned before, option 2 in line 23 of Algorithm 1 was implemented in all case studies. An analogous analysis for models (1)-(4) & (6) can be found in Figures 10 & 11. We consider the DRMs $\rho_{g_\alpha, \gamma}(Y)$, $\alpha = 0.05, \gamma \in \{1/2, 1, 2\}$. The figures show the true density of Y . Additionally, 200 samples from the mixture distribution (with values on the x-axis) are plotted along the probability density. The labeled quantile $q_Y(0.95)$ indicates the threshold above which samples are relevant

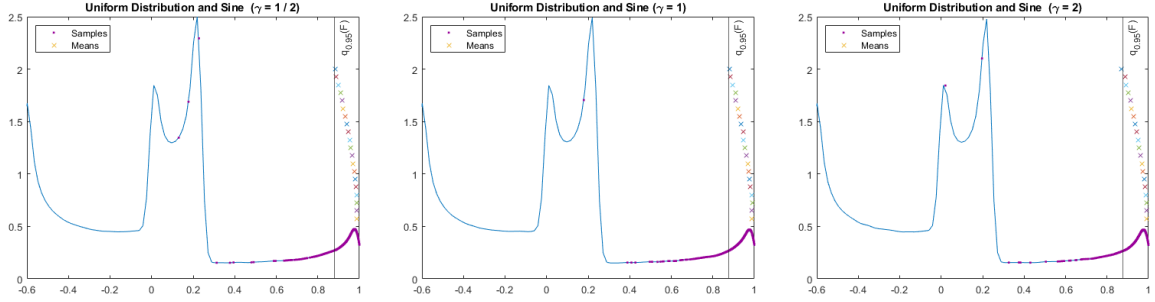


Figure 2: An example of 200 samples drawn from the mixture distribution plotted on the underlying distribution of the model Y for the case study (5). To approximate the mixture weights and optimal mixture components $M = 20,000$ pivot samples were drawn.

for the estimation of $\rho_{g_{\alpha,\gamma}}(Y)$. The crosses mark the expectations of the individual importance sampling components of the mixture distribution. By design, samples and expectations are in the right tail of the distribution in the area relevant for the estimation of the DRM.

3.2.2 Efficiency of the Estimations

In this section, we discuss the efficiency of the algorithm for case studies (1)-(6). The ML approximation used in importance sampling is fixed, and we compare different hypothesis classes. The calibration of the ML regressions and the construction of the importance sampling measure change are based on $M = 2,000$ pivot samples. We choose $m = 20$ and $N = 20,000$ and run $R = 1,000$ replications of the simulation for estimating the RMSE. Figure 3 & 4 show the ratio of the RMSE between the crude estimate and the proposed importance sampling method for the DRMs $\rho_{g_{\alpha,\gamma}}$ with $\gamma \in \{1/2, 1, 2\}$ and $\alpha \in [0.01, 0.3]$ for models (1) to (6). The absolute estimated RMSE for the different estimation methods is shown in Figure 16 & 17 in Appendix A.14.

The plots confirm that the proposed importance sampling algorithm can successfully reduce the RMSE in all cases. The efficiency of the algorithm depends significantly on the chosen ML regression model. A poorly chosen approximation can even lead to a higher error than the crude method. Interestingly, the choice based on k -fold validation and pivot calibration works reasonably well in all cases, despite the fact that the ML objective function does not focus on the tail. The smaller α , the more the DRM zooms in on the tail risk due to rare events. As expected, the variance reduction becomes better the smaller α is. Similarly, variance reduction is also better the smaller γ , since DRMs with smaller γ put more emphasis on tail risk.

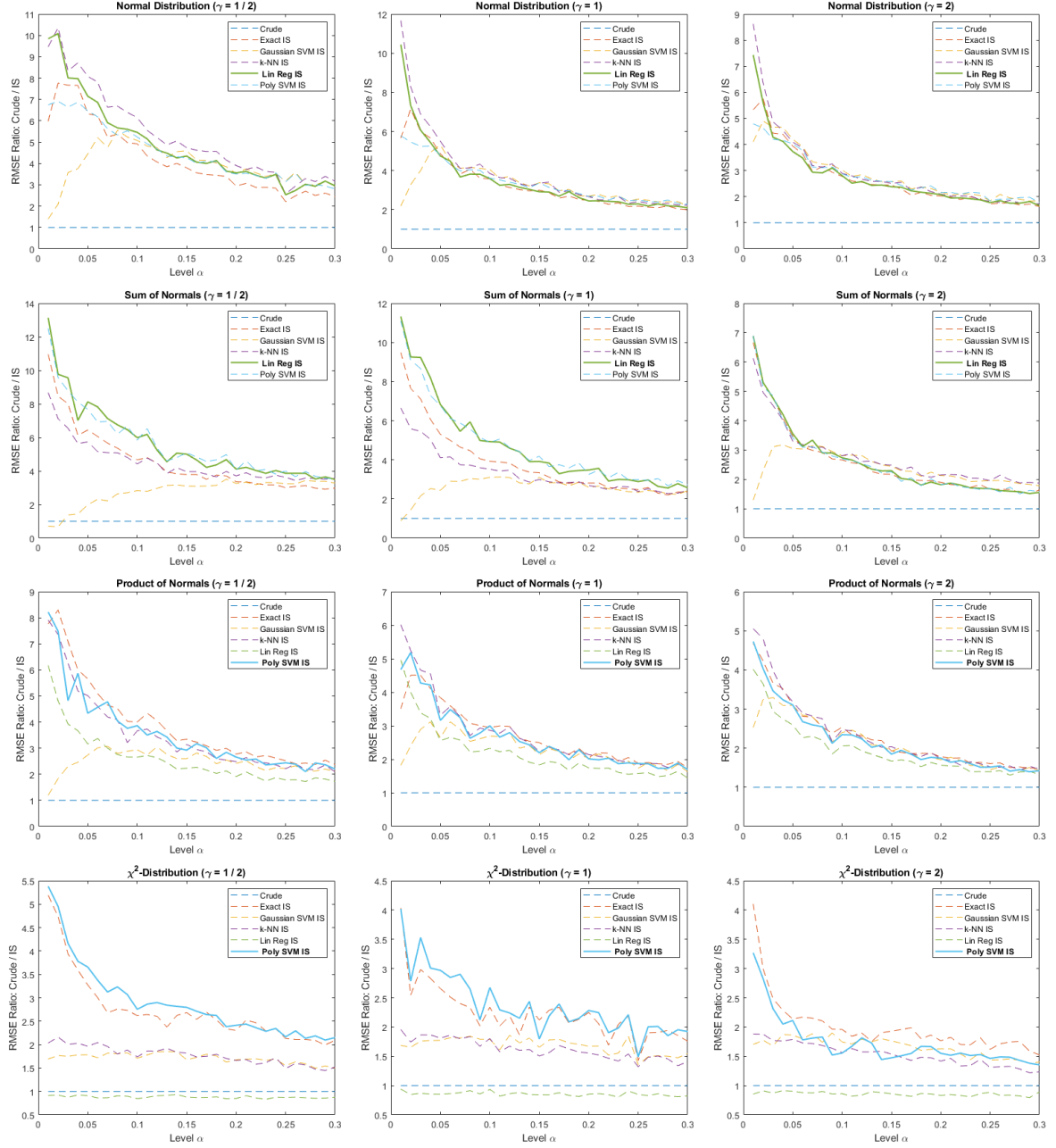


Figure 3: Ratio of the RMSE arising between the crude method and the IS method when estimating $\rho_{g_{\alpha,\gamma}}$, with $\gamma \in \{1/2, 1, 2\}$, $\alpha \in [0.01, 0.3]$, for the models (1) to (6). The comparison is made between the crude method and the proposed IS method using various approximations for the black box considered in the paper.

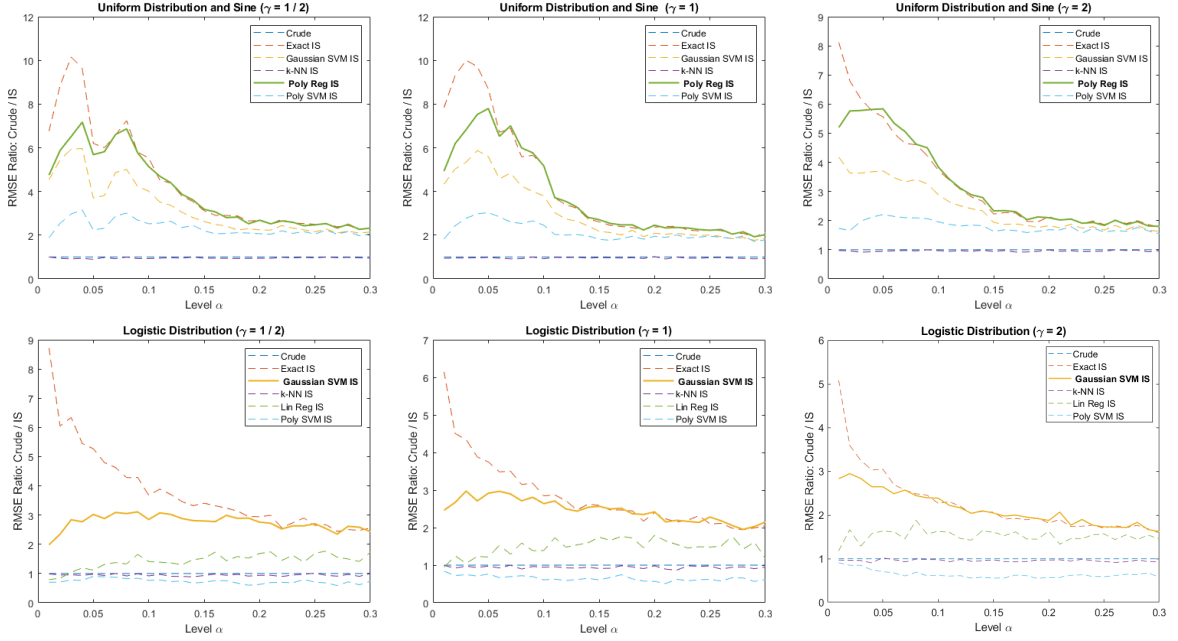


Figure 4: Continuation of Figure 3.

3.3 Iterative Exploration of the Extreme Tail

The discussed algorithm consists of two simulation steps. First, pivot samples are drawn that are used for both the choice of the ML approximation and the determination of an IS measure change. Second, samples are generated under the IS distribution and used for the estimation of the DRMs. However, if DRMs are considered that focus on particularly extreme tail events, this approach might not yet be sufficient. A possible extension to the suggested approach is the following: The samples from the IS distribution are not directly used for DRM estimation, but serve as additional data for calibrating the ML approximation a second time and the construction of a further measure change on this basis. By using samples from the IS distribution for the ML approximation, our goal is to obtain more samples in the right tail, thereby enhancing the ML approximation's accuracy in that region. This results in a more suitable IS distribution for the subsequent measure change in DRM estimation. In this section, we provide a case study that takes this approach – the iterative exploration of the extreme tail.

Simulation Design

We consider again the case studies (1) - (4) outlined in Section 3.1 with the same distortion functions $g_{\alpha,\gamma}(u)$, $\gamma \in \{1/2, 1, 2\}$ and corresponding DRMs. However, we focus on more extreme tail events by choosing $\alpha = 0.002$. In addition, we choose a finer partition by setting $m = 50$. In the experiment, we repeat all simulation runs $R = 2,000$ times to estimate the RMSE. We

γ	(1) Id. of Normals			(2) Sum of Normals			(3) Prod. of Normals			(4) Sum of Sq. Normals		
	1/2	1	2	1/2	1	2	1/2	1	2	1/2	1	2
Exact	3.35	3.16	3.02	5.40	5.09	4.88	4.08	3.60	3.30	20.58	19.01	17.99
Mean CRUDE	3.33	3.15	3.03	5.38	5.08	4.88	4.02	3.59	3.30	20.45	18.99	18.00
Mean IS	3.35	3.16	3.02	5.40	5.09	4.88	4.07	3.60	3.30	20.48	18.91	17.89
Mean ITER IS	3.35	3.16	3.02	5.40	5.09	4.88	4.06	3.59	3.29	20.55	18.98	17.97
RMSE $\frac{\text{CRUDE}}{\text{IS}}$	35.61	20.63	14.87	16.73	13.09	10.34	10.90	9.40	7.48	5.34	3.58	2.97
RMSE $\frac{\text{CRUDE}}{\text{ITER IS}}$	35.42	20.81	14.41	35.02	21.70	15.04	9.61	8.39	6.71	13.55	8.89	7.17

Table 1: Results of the DRM estimation in the extreme tail.

compare three simulation approaches for the estimation of the DRMs. In all cases, 27,500 samples are used, respectively.

The first approach is a crude simulation with 27,500 samples. The second approach is the algorithm suggested in the previous sections with 7,500 pivot samples and 20,000 samples from the IS distribution. The third approach is an iterative exploration: 5,000 pivot samples are used to calculate the IS distribution for level $\alpha' = 0.01$. Then we draw from this IS distribution 2,500 additional pivot samples. The IS distribution for $\alpha = 0.002$ is computed from the total 7,500 pivot samples. In the last step, 20,000 samples are drawn from this distribution to estimate the DRMs.

Results

The results of the case study are displayed in Table 1. The exact values of the DRMs, the means over $R = 2,000$ simulation runs and the corresponding ratios of the RMSE of the two IS methods and the crude method are documented. Overall, the iterative method typically provides the most substantial RMSE reduction, outperforming the direct IS approach substantially in experiments (2) and (4). The direct IS approach is still more efficient than the crude method. Especially in (1) and (2), the reduction of the RMSE is significant in contrast to the crude method, while in (3) and (4) the IS methods are not as efficient. When considering the mean over all simulation runs, we observe that the IS methods also reduce estimation bias.

4 Application to ALM

We apply Algorithm 1 to the estimation of solvency capital in a simple asset-liability management (ALM) model of an insurance firm. Instead of the risk measure $V@R$ which forms the basis of Solvency II, we use the same DRMs that we considered in Section 3. The suggested

method could also be applied in highly complex ALM models such as those applied by major insurance groups.

4.1 Model Description

Our ALM model, inspired by Weber et al. [2014] and Hamm et al. [2019], describes a snapshot in time of an ongoing insurance business. The focus is on a one-year time horizon with dates $t = 0, 1$, as in Solvency II. The values of assets and liabilities are denoted by A_t, L_t , $t = 0, 1$, respectively. At each point in time, their difference is the book value of equity $E_t = A_t - L_t$, $t = 0, 1$, which is used for the solvency capital calculation.

The evolution of the balance sheet is driven by market and insurance risks. For simplicity, we assume that reserves are constant, i.e., $L_t = v$, $\forall t$. Any changes in value are thus seen on the asset side. We assume that insurance claims are modeled by a collective model where the number of claims is given by the counting process N and their severities by independent, identically distributed losses ξ_k . Annual total premium payments π are received at the beginning of the year. We set

$$C = \sum_{k=1}^N \xi_k.$$

The random annual return of assets between dates $t = 0$ and $t = 1$ is denoted by R_A , i.e., we obtain that

$$A_1 = R_A \cdot A_0 - C + \pi.$$

In order to model the random return of assets we assume that

$$A_0 = \eta^S S_0 + \eta^B B_0,$$

where $B = (B_t)_{t \in \{0,1\}}$ and $S = (S_t)_{t \in \{0,1\}}$ are the prices of a bond and a stock and η^S and η^B the respective holdings. This implies that

$$R_A = \frac{\eta^S S_1 + \eta^B B_1}{A_0} = \frac{\eta^S S_0}{A_0} \cdot \frac{S_1}{S_0} + \frac{\eta^B B_0}{A_0} \cdot \frac{B_1}{B_0} = b \cdot \frac{S_1}{S_0} + (1-b) \cdot \frac{B_1}{B_0},$$

where b is the fraction of initial wealth invested in the stock. Setting $B_0 = 1$, $B_1 = 1 + R_B$ for

some random interest rate $R_B > -1$, and $S_0 = 1$,

$$S_1 = \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t}Z\right)$$

for a standard normal Z , we derive that

$$R_A = b \cdot \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t}Z\right) + (1 - b) \cdot (1 + R_B).$$

Solvency capital is determined in terms of risk measure applied to the change in net asset value over the time period of one year, i.e.,

$$E_1 - E_0 = A_1 - \underbrace{L_1}_{=v} - A_0 + \underbrace{L_0}_{=v} = (R_A - 1) \cdot A_0 - C + \pi.$$

4.2 Simulation Overview

As in Section 3, we apply the proposed importance sampling method to the DRMs $\rho_{g_{\alpha,\gamma}}$ with distortion function $g_{\alpha,\gamma}(u) = \mathbf{1}_{[0,\alpha]}(u) \left(\frac{u}{\alpha}\right)^\gamma + \mathbf{1}_{(\alpha,1]}(u)$ and $\gamma \in \{1/2, 1, 2\}$. In terms of the DRMs, solvency capital is $\rho_{g_{\alpha,\gamma}}(E_0 - E_1)$. The underlying random factors are $R_B, Z, N, \xi_1, \xi_2, \dots$. We set $E_0 = 1000$ and $R_B = (V - 1/2)/10$, where V is beta distributed with parameters $(2, 2)$, i.e., $V \sim B(2, 2)$. The parameters of the stock are $\mu = 0.02$, $\sigma = 0.2$, $\Delta t = 1$, and we assume that half of the available capital is invested into the stock, i.e., $b = 0.5$. For the collective model, we assume that N is a Poisson random variable with parameter $\lambda = 5$, and $(\xi_k)_{k \geq 1}$ are independent exponentially distributed random variables with mean $\vartheta' = 10$. The premium and reserve are 103% and 105% of the expected claims, respectively, such that $\pi = 1.03\lambda\vartheta'$ and $v = 1.05\lambda\vartheta'$.

As in Section 3, the importance sampling estimates with the different ML approximations used in the measure changes are performed with $M = 2,000$ pivot samples for calibration of the approximation and determination of the importance sampling mixture distribution, and $N = 20,000$ samples of the mixture distribution for DRM estimation. For comparison, a crude estimation with $M + N$ samples is implemented. We always use a discretization with $m = 20$. The ‘‘exact value’’ of benchmarking is determined with a crude estimation with 1,000,000 samples and used to calculate the RMSE over $R = 1,000$ simulation runs.

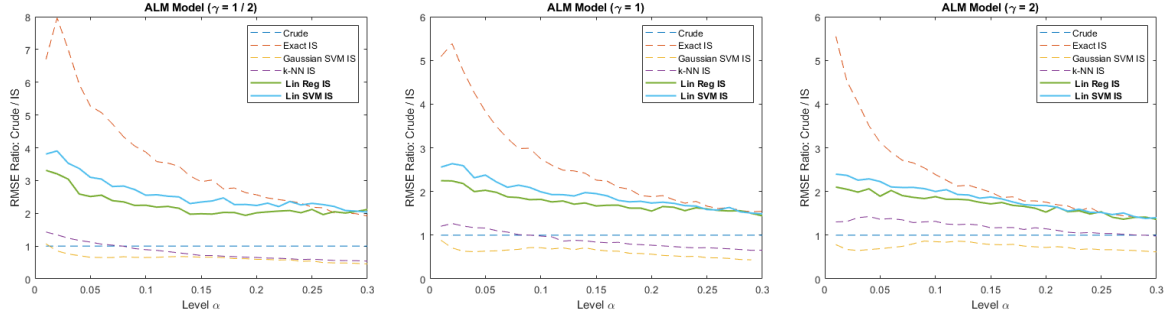


Figure 5: Ratio of the RMSE between the crude method and the importance sampling method for the estimation of the considered DRMs for the evolution of the net asset value in the ALM model. The importance sampling method is implemented with the different approximation techniques considered in the paper.

4.3 Results

The results of the simulations are shown in Figure 5, which presents the ratio of the RMSEs of the crude method and the studied importance sampling methods. For further reference, the absolute RMSEs of the estimates are provided in Figure 18 in Appendix A.14. For all DRMs, importance sampling with exact knowledge of the model leads to a significant reduction in the RMSE, especially for small α . The variance reduction becomes less pronounced as α increases. For $\gamma = 1$, the exact method achieves the highest observed RMSE ratio of 8.8 for $\alpha = 0.01$. Importance sampling with linear regression leads to a maximum reduction of 5.8 for $\alpha = 0.01$. For the DRMs with concave ($\gamma = 1/2$) and on $[0, \alpha]$ convex ($\gamma = 2$) distortion functions, the linear SVM gives the best reduction in RMSE. For $\gamma = 2$, the best ratio obtained with full knowledge of the model is 5.55 for $\alpha = 0.01$, and with linear SVM, the maximum reduction is 2.39 for $\alpha = 0.01$. For $\gamma = 1/2$, the exact importance sampling method has the best ratio of 7.96 for $\alpha = 0.02$ and the linear SVM approximation achieves 3.9 for $\alpha = 0.02$. Across all the different estimated DRMs, we see that the importance sampling methods with Gaussian SVM and k -NN regression can lead to a worse RMSE than the crude method. The worst ratio is observed in all cases with the Gaussian SVM with 0.51 as $\alpha = 0.28$ for $\gamma = 1$, 0.62 as $\alpha = 0.3$ for $\gamma = 2$ and 0.45 for $\gamma = 1/2$ with $\alpha = 0.3$.

In summary, the proposed method provides a good path to variance reduction. However, the ML approximation in the measure change needs to be chosen carefully, but k -fold validation seems to work quite well for this type of analysis. The variance reduction becomes better the more the risk measure depends on extreme tail events. In the ALM case study, the most extreme parameter α was 0.01. We expect that the iterative procedure outlined in Section 3.3 would also

lead to further improvements in variance reduction when the very extreme tail is considered.

References

- Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.
- Jae Youn Ahn and Nariankadu D. Shyamalkumar. Large sample behavior of the CTE and VaR estimators under importance sampling. *North American Actuarial Journal*, 15(3):393–416, 2011. doi: 10.1080/10920277.2011.10597627.
- Naomi S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. doi: 10.2307/2685209.
- Mansur Arief, Yuanlu Bai, Wenhao Ding, Shengyi He, Zhiyuan Huang, Henry Lam, and Ding Zhao. Certifiable deep importance sampling for rare-event simulation of black-box systems. In *Proceedings of the 24-th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 595–603, 2021.
- Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999. doi: 10.1111/1467-9965.00068.
- Søren Asmussen, Klemens Binswanger, and Bjarne Højgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6(2):303–322, 2000. doi: 10.2307/3318578.
- Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 1 edition, 2007.
- Karl F. Bannör and Matthias Scherer. On the calibration of distortion risk measures to bid-ask prices. *Quantitative Finance*, 14(7):1217–1228, 2014.
- Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007. doi: <https://doi.org/10.1111/j.1467-9965.2007.00311.x>.
- Sören Bettels, Sojung Kim, and Stefan Weber. Multinomial backtesting of distortion risk measures. *Insurance: Mathematics and Economics*, 119:130–145, 2024.
- Eric Beutner and Henryk Zähle. A modified functional delta method and its application to the estimation of risk functionals. *Journal of Multivariate Analysis*, 101(10):2452–2463, 2010. doi: 10.1016/j.jmva.2010.06.015.
- Valeria Bigozzi and Andreas Tsanakas. Parameter uncertainty and residual estimation risk. *Journal of Risk and Insurance*, 83(4):949–978, 2015. doi: 10.1111/jori.12075.
- Jose Blanchet and Peter Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *The Annals of Applied Probability*, 18(4):1351–1378, 2008. doi: 10.1214/07-aap485.

- Vytaras Brazauskas, Bruce L. Jones, Madan L. Puri, and Ričardas Zitikis. Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference*, 138(11):3590–3604, 2008. doi: 10.1016/j.jspi.2005.11.011.
- James A. Bucklew. *Introduction to Rare Event Simulation*. Springer New York, 5 edition, 2004.
- Alexander Cherny and Dilip Madan. New measures for performance evaluation. *Review of Financial Studies*, 22(7):2571–2606, 2008. doi: 10.1093/rfs/hhn081.
- Gustave Choquet. Theory of capacities. *Annales de l’institut Fourier*, 5:131–295, 1954. doi: 10.5802/aif.53.
- Dieter Denneberg. *Non-Additive Measure and Integral*. Springer Netherlands, 1 edition, 1994. doi: 10.1007/978-94-017-2434-0.
- Jan Dhaene, Steven Vanduffel, Marc J. Goovaerts, Rob Kaas, Qihe Tang, and David Vyncke. Risk measures and comonotonicity: A review. *Stochastic Models*, 22(4):573–606, 2006. doi: 10.1080/15326340600878016.
- Jan Dhaene, Alexander Kukush, Daniël Linders, and Qihe Tang. Remarks on quantiles and distortion risk measures. *European Actuarial Journal*, 2(2):319–328, 2012. doi: 10.1007/s13385-012-0058-0.
- Kevin Dowd, John Cotter, and Ghulam Sorwar. Spectral risk measures: Properties and limitations. *Journal of Financial Services Research*, 34(1):61–75, 2008. doi: 10.1007/s10693-008-0035-6.
- Harris Drucker, Christopher J. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9: 155 – 161, 1996.
- Jörn Dunkel and Stefan Weber. Efficient monte carlo methods for convex risk measures in portfolio credit risk models. *Operations Research*, 58(5):1505 – 1521, 2007.
- Paul Dupuis and Hui Wang. Importance sampling, large deviations, and differential games. *Stochastics: An International Journal of Probability and Stochastic Processes*, 76(6):481–508, 2004.
- Jonathan El Methni and Gilles Stupfler. Extreme versions of wang risk measures and their estimation for heavy-tailed distributions. *Statistica Sinica*, 27(2):907–930, 2017.
- Rong-En Fan, Pai-Hsuen Chen, Chih-Jen Lin, and Thorsten Joachims. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6(12):1889–1918, 2005.
- Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2 edition, 1999.
- Marco Frittelli and Emanuela Rosazza Gianin. Putting order in risk measures. *Journal of Banking & Finance*, 26(7):1473–1486, 2002. doi: 10.1016/s0378-4266(02)00270-4.

- Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002. doi: 10.1007/s007800200072.
- Hans Föllmer and Alexander Schied. *Stochastic Finance*. De Gruyter, 4 edition, 2016. doi: 10.1515/9783110463453.
- Hans Föllmer and Stefan Weber. The axiomatic approach to risk measures for capital determination. *Annual Review of Financial Economics*, 7(1):301–337, 2015. doi: 10.1146/annurev-financial-111914-042031.
- Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer New York, 1 edition, 2003. doi: 10.1007/978-0-387-21617-1.
- Paul Glasserman, Philip Heidelberger, and Perwez Shahabuddin. Portfolio value-at-risk with heavy-tailed risk factors. *Mathematical Finance*, 12(3):239–269, 2002. doi: 10.1111/1467-9965.00141.
- Peter W. Glynn. Importance sampling for monte carlo estimation of quantiles. In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, pages 180–185, 1996.
- Dominique Guégan and Bertrand Hassani. Distortion risk measure or the transformation of unimodal distributions into multimodal functions. In *Future Perspectives in Risk Models and Finance*, pages 71–88. Springer International Publishing, 2014.
- Montserrat Guillen, Jose Maria Sarabia, Jaume Belles-Sampera, and Faustino Prieto. Distortion risk measures for nonnegative multivariate risks. *Journal of Operational Risk*, 13(2):35–57, 2018. doi: 10.21314/jop.2018.206.
- Anna-Maria Hamm, Thomas Knispel, and Stefan Weber. Optimal risk sharing in insurance networks. *European Actuarial Journal*, 10(1):203–234, 2019.
- Jian Huang, Junyi Chai, and Stella Cho. Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1):1–24, 2020.
- Te-Ming Huang, Vojislav Kecman, and Ivica Kopriva. *Kernel Based Algorithms for Mining Huge Data Sets*. Springer Berlin, 1 edition, 2006.
- Henrik Hult and Pierre Nyquist. Large deviations for weighted empirical measures arising in importance sampling. *Stochastic Processes and their Applications*, 126(1):138–170, 2016. doi: 10.1016/j.spa.2015.08.002.
- Sandeep Juneja and Perwez Shahabuddin. Rare-event simulation techniques: An introduction and recent advances. In *Handbooks in Operations Research and Management Science*, volume 13, pages 291–350. Elsevier, 2006. doi: 10.1016/s0927-0507(06)13011-x.
- Sojung Kim and Stefan Weber. Simulation methods for robust risk assessment and the distorted mix approach. *European Journal of Operational Research*, 298(1):380–398, 2022. doi: 10.1016/j.ejor.2021.07.005.

- Shigeo Kusuoka. On law invariant coherent risk measures. In *Advances in Mathematical Economics*, volume 3, pages 83–95. Springer Japan, 2001.
- Lujun Li, Hui Shao, Ruodu Wang, and Jingping Yang. Worst-case range value-at-risk with partial information. *SIAM Journal on Financial Mathematics*, 9(1):190–218, 2018. doi: 10.1137/17m1126138.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2 edition, 2018.
- Ajay Kumar Pandey, Prashanth L.A., and Sanjay P. Bhat. Estimation of spectral risk measures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12166–12173, 2021. doi: 10.1609/aaai.v35i13.17444.
- John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Microsoft Research*, 1998.
- R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.
- R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26:1443–1471, 2002.
- Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.
- Gerardo Rubino and Bruno Tuffin. *Rare Event Simulation using Monte Carlo Methods*. John Wiley & Sons, 1 edition, 2009. ISBN 9780470772690.
- Ranadeera G.M. Samanthi and Jungsywan Sepanski. Methods for generating coherent distortion risk measures. *Annals of Actuarial Science*, 13(2):400–416, 2018. doi: 10.1017/s1748499518000258.
- Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., 1 edition, 1980. doi: 10.1002/9780470316481.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 1 edition, 2014.
- Lawrence F. Shampine. Matlab program for quadrature in 2d. *Applied Mathematics and Computation*, 202(1):266–274, 2008.
- Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. doi: 10.1023/B:STCO.0000035301.49549.88. URL <https://doi.org/10.1023/B:STCO.0000035301.49549.88>.
- Yongsheng Song and Jia-An Yan. Risk measures with comonotonic subadditivity or convexity and respecting stochastic orders. *Insurance: Mathematics and Economics*, 45(3):459–465, 2009a. doi: 10.1016/j.insmatheco.2009.09.011.

- Yongsheng Song and Jia-An Yan. An overview of representation theorems for static risk measures. *Science in China Series A: Mathematics*, 52(7):1412–1422, 2009b. doi: 10.1007/s11425-009-0122-7.
- Yongsheng Song and Jia’an Yan. The representation of two types functionals on $l^\infty(\omega, \mathcal{F})$ and $l^\infty(\omega, \mathcal{F}, p)$. *Science in China Series A: Mathematics*, 49(10):1376–1382, 2006. doi: 10.1007/s11425-006-2010-8.
- Stephen M. Stigler. Linear functions of order statistics with smooth weight functions. *The Annals of Statistics*, 2(4), 1974. doi: 10.1214/aos/1176342756.
- Lihua Sun and L. Jeff Hong. A general framework of importance sampling for value-at-risk and conditional value-at-risk. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 415–422, 2009.
- Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2 edition, 1999.
- Shaun Wang. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, 17(1):43–54, 1995. doi: 10.1016/0167-6687(95)00010-p.
- Shaun Wang. Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, 26(1):71–92, 1996.
- Shaun Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, pages 15–36, 2000.
- Shaun Wang. A risk measure that goes beyond coherence. *12th AFIR International Colloquium*, 2001.
- Stefan Weber. Solvency II, or how to sweep the downside risk under the carpet. *Insurance: Mathematics and Economics*, 82:191–200, 2018. doi: 10.1016/j.insmatheco.2017.11.010.
- Stefan Weber, Anna-Maria Hamm, Torsten Becker, Claudia Cottin, Matthias Fahrenwaldt, and Stefan Nörtemann. Market consistent embedded value – eine praxisorientierte Einführung. *Der Aktuar*, 1:4–8, 2014.
- Julia Lynn Wirch and Mary R. Hardy. A synthesis of risk measures for capital adequacy. *Insurance: Mathematics and Economics*, 25(3):337–347, 1999. doi: 10.1016/s0167-6687(99)00036-0.
- David Wozabal. Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research*, 62(6):1302–1315, 2014.
- Shlomo Yitzhaki. Stochastic dominance, mean variance, and gini’s mean difference. *The American Economic Review*, 72(1):178–185, 1982.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

This paper, including its appendix, is original research conducted by the authors. Generative AI tools were employed solely for language refinement and clarity enhancement. The substantive content, analyses, and conclusions remain entirely the work of the authors.

Funding Declaration

This research received no external funding.

Conflict of Interest

The authors declare that they have no conflict of interest.

A Online Appendix

This is an online appendix that is provided as an electronic supplement to the paper.

A.1 Distortion Risk Measures

We review some facts related to risk measures and the special case of DRMs. Let \mathcal{X} denote a set of suitable (e.g., all bounded) measurable functions on the measurable space (Ω, \mathcal{F}) . Elements in \mathcal{X} model financial positions or insurance losses. We use the sign convention to interpret positive values as losses and negative values as gains. Precise technical conditions for the results below are stated in the references that we mention.

The axiomatic definition of risk measures goes back to Artzner et al. [1999]; the notion of DRMs was developed by Wang [1996] and Acerbi [2002]. DRMs are a subclass of comonotonic risk measures. The link of comonotonic risk measures and DRMs is briefly discussed in Appendix A.2. For an excellent overview of risk measures and DRMs, we refer to Föllmer and Schied [2016]. A risk measure $\rho : \mathcal{X} \rightarrow \mathbb{R}$ is a functional that quantifies the risk of elements of \mathcal{X} :

Definition A.1. *A mapping $\rho : \mathcal{X} \rightarrow \mathbb{R}$ is called a monetary risk measure if the following properties hold:*

(i) *Monotonicity:* If $X \leq Y$, $X, Y \in \mathcal{X}$, then $\rho(X) \leq \rho(Y)$.

(ii) *Cash-Invariance:* If $X \in \mathcal{X}$ and $m \in \mathbb{R}$, then $\rho(X + m) = \rho(X) + m$.

Risk measures may exhibit additional properties such as quasi-convexity, which in economic terms means that diversification of positions does not increase the measured risk. This property can be shown to be equivalent to convexity:

Definition A.2. *A risk measure $\rho : \mathcal{X} \rightarrow \mathbb{R}$ is called convex, if for $X, Y \in \mathcal{X}$, $\lambda \in [0, 1]$*

$$\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y).$$

A DRM is defined as follows:

Definition A.3. (i) *A non-decreasing function $g : [0, 1] \rightarrow [0, 1]$ with $g(0) = 0$ and $g(1) = 1$ is called distortion function.*

(ii) Let \mathbf{P} be a probability measure on (Ω, \mathcal{F}) and g be a distortion function. The monetary risk measure $\rho_g : \mathcal{X} \rightarrow \mathbb{R}$ defined by

$$\rho_g(X) = \int_{-\infty}^0 [g(\mathbf{P}(\{X > x\})) - 1] dx + \int_0^{\infty} g(\mathbf{P}(\{X > x\})) dx$$

is called a DRM with respect to g .

If the distortion function g is concave we obtain a convex risk measure (see Föllmer and Schied [2016]):

Theorem A.4. *Consider the distortion function g and the corresponding DRM ρ_g . If g is concave, then the DRM ρ_g is a convex risk measure. If the underlying probability space is atomless, the converse implication is also true.*

DRMs are expressible as mixtures of quantiles. One must focus on the details of the continuity properties of the distortion function to obtain the correct representation, as shown in Dhaene et al. [2012].

Theorem A.5. (i) *Let g be a right continuous distortion function. Then the DRM $\rho_g(X)$ is given by*

$$\rho_g(X) = \int_{[0,1]} q_X^+(1-u) dg(u),$$

where $q_X^+(u) = \sup\{x | F_X(x) \leq u\}$.

(ii) *Let g be a left continuous distortion function. Then the DRM $\rho_g(X)$ is given by*

$$\rho_g(X) = \int_{[0,1]} q_X(1-u) dg(u) = \int_{[0,1]} q_X(u) d\bar{g}(u),$$

where $q_X(u) = \inf\{x | F(x) \geq u\}$ and $\bar{g}(u) = 1 - g(1-u)$, $u \in [0, 1]$.

Many important risk measures fall into the class of DRMs; for various examples, see Cherny and Madan [2008], Föllmer and Schied [2016], Weber [2018], and the Appendix A.3. Particularly important examples will be discussed here:

Example A.6. (i) *Let $g(u) = \mathbf{1}_{(\alpha,1]}(u)$, then ρ_g is the Value at Risk at level α , so that*

$$\rho_g(X) = V@R_\alpha(X) = \inf\{x | F(x) \geq 1 - \alpha\}.$$

(ii) The distortion function $g(u) = \frac{u}{\alpha} \mathbf{1}_{[0,\alpha]}(u) + \mathbf{1}_{(\alpha,1]}(u)$ yields the Average Value at Risk at level α , i.e.,

$$\rho_g(X) = AV@R_\alpha(X) = \frac{1}{\alpha} \int_0^\alpha V@R_\lambda(X) d\lambda.$$

(iii) The distortion function $g(u) = \left(\frac{u}{\alpha}\right)^\gamma \mathbf{1}_{[0,\alpha]}(u) + \mathbf{1}_{(\alpha,1]}(u)$ with $\alpha \in (0, 1]$ and $\gamma \in \mathbb{R}_{>0}$ generalizes the distortion function of the AV@R ($\gamma = 1$) and other special cases such as the hazard transform ($\gamma \geq 1, \alpha = 1$) and MAXV@R ($\gamma \in \mathbb{N}, \alpha = 1$).

If $\gamma \leq 1$, the distortion function $g_{\alpha,\gamma}$ is concave such that the corresponding DRM is convex. If $\gamma > 1$ the distortion function is convex on the interval $[0, \alpha]$. In this case, the resulting DRM is not convex.

Remark A.7. Every distortion function g can be decomposed in the convex combination of a left and right continuous distortion function (see Dhaene et al. [2012]), such that $g(u) = c_1 g_1(u) + c_2 g_2(u)$ with $c_1 + c_2 = 1$ and $c_1, c_2 \geq 0$. As a consequence, any distortion risk measure ρ_g with general distortion function can be expressed as a convex combination $\rho_g(X) = c_1 \rho_{g_1}(X) + c_2 \rho_{g_2}(X)$. The decompositions of g and ρ_g are not unique, unless g is a step function. Bettels et al. [2024] point out that a decomposition of g into a left and a right continuous step function and a continuous function is unique.

A.2 Comonotonic Risk Measures

We review the connections between comonotonic risk measures, Choquet integrals and DRMs. More details can be found in Föllmer and Schied [2016]. (Ω, \mathcal{F}) is a measurable space on which the financial positions in \mathcal{X} are defined.

Definition A.8. (i) Two measurable functions X, Y on (Ω, \mathcal{F}) are called comonotonic if

$$(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \geq 0 \quad \forall (\omega, \omega') \in \Omega \times \Omega.$$

(ii) A risk measure $\rho : \mathcal{X} \rightarrow \mathbb{R}$ is called comonotonic if

$$\rho(X + Y) = \rho(X) + \rho(Y),$$

for comonotonic $X, Y \in \mathcal{X}$.

Comonotonic risk measures are expressible as Choquet integrals with respect to capacities.

Definition A.9. (i) A map $c : \mathcal{F} \rightarrow [0, \infty)$ is called *monotonic set function* if it satisfies the following properties:

a) $c(\emptyset) = 0$.

b) $A, B \in \mathcal{F}, A \subseteq B \Rightarrow c(A) \leq c(B)$.

If, in addition, $c(\Omega) = 1$, i.e., c is normalized, then c is called a *capacity*.

(ii) For $X \in \mathcal{X}$ the Choquet integral of X with respect to the monotone set function c is defined by

$$\int X dc = \int_{-\infty}^0 [c(\{X > x\}) - c(\Omega)] dx + \int_0^{\infty} c(\{X > x\}) dx.$$

The Choquet integral coincides with the Lebesgue integral if c is a σ -additive probability measure. The following characterization theorem can, for example, be found in Chapter 4 of Föllmer and Schied [2016].

Theorem A.10. A monetary risk measure $\rho : \mathcal{X} \rightarrow \mathbb{R}$ is comonotonic, if and only if there exists a capacity c on (Ω, \mathcal{F}) such that

$$\rho(X) = \int X dc.$$

DRMs are an important special case of the comonotonic risk measures. In this case, the capacity is defined in terms of a distorted probability measure \mathbb{P} . The resulting capacity is absolutely continuous with respect to \mathbb{P} , but typically not additive.

Definition A.11. (i) If \mathbb{P} is a probability measure on (Ω, \mathcal{F}) and g is a distortion function, then

$$c^g(A) := g(\mathbb{P}(A)), \quad A \in \mathcal{F},$$

is called a *distorted probability*.

(ii) A comonotonic risk measure $\rho(X) = \int X dc$ is called a *DRM*, if the capacity c can be expressed as a distorted probability.

A.3 Examples of DRMs

For further reference, we include a list of examples of distortion risk measures in Table 2 which was compiled in the online appendix of El Methni and Stupfler [2017].

Name	Distortion	Closed form	Reference
MINV@R	$1 - (1 - u)^n$	$-\mathbb{E}[\min\{-X_1, \dots, -X_n\}]$ $= \mathbb{E}[\max\{X_1, \dots, X_n\}]$	Cherny and Madan [2008] Föllmer and Schied [2016] Bannör and Scherer [2014]
MAXV@R	$u^{1/n}$	$-\mathbb{E}[Y_1]$ such that $\max\{Y_1, \dots, Y_n\} \sim -X$	Cherny and Madan [2008] Föllmer and Schied [2016] Bannör and Scherer [2014]
MINMAXV@R	$1 - (1 - u^{1/n})^n$	$-\mathbb{E}[\min\{Y_1, \dots, Y_n\}]$ such that $\max\{Y_1, \dots, Y_n\} \sim -X$	Cherny and Madan [2008] Föllmer and Schied [2016] Bannör and Scherer [2014]
MAXMINV@R	$(1 - (1 - u)^n)^{1/n}$	$-\mathbb{E}[Y_1]$ such that $\max\{Y_1, \dots, Y_n\}$ $\sim \min\{-X_1, \dots, -X_n\}$	Cherny and Madan [2008] Föllmer and Schied [2016] Bannör and Scherer [2014]
<i>RV@R</i> (Range <i>V@R</i>)	$\frac{u-\beta}{\alpha-\beta} \mathbb{1}_{\{\beta < u \leq \alpha\}} + \mathbb{1}_{\{u > \alpha\}}$ $0 < \beta < \alpha < 1$	$\frac{1}{\alpha-\beta} \int_{\beta}^{\alpha} V@R_{\lambda}(X) d\lambda$	Bignozzi and Tsanakas [2015] Weber [2018], Li et al. [2018]
Proportional hazard transform	$u^{1/\gamma}$ $\gamma > 1$	$\int_0^{\infty} (1 - F(x))^{1/\gamma} dx,$ if $X \geq 0$ a.s.	Wang [1995, 1996] Guillen et al. [2018]
Dual power transform	$1 - (1 - u)^{\gamma}$ $\gamma > 1$	$\int_0^{\infty} 1 - F(x)^{\gamma} dx,$ if $X \geq 0$ a.s.	Wirch and Hardy [1999] Guillen et al. [2018]
Gini's principle	$(1 - \vartheta)u + \vartheta u^2$ $0 < \vartheta < 1$	$\mathbb{E}[X] + \frac{\vartheta}{2} \mathbb{E}[X - X_1]$	Yitzhaki [1982], Wozabal [2014] Guillen et al. [2018]
Exponential transform	$\frac{1 - \exp(-ru)}{1 - \exp(-r)}$ if $r > 0$ u if $r = 0$	-	El Methni and Stupfler [2017] Dowd et al. [2008]
Inverse S-shaped polynomial of degree 3	$a \left[\frac{u^3}{6} - \frac{\delta u^2}{2} + \left(\frac{\delta^2}{2} + \beta \right) u \right]$ $a = \left(\frac{1}{6} - \frac{\delta}{2} + \frac{\delta^2}{2} + \beta \right)^{-1}$ $0 < \delta < 1, \beta \in \mathbb{R}$	-	Guégan and Hassani [2014] El Methni and Stupfler [2017]
Beta family	$\int_0^u \frac{t^{a-1}(1-t)^{b-1}}{B(a,b)} dt$ $a, b > 0$	-	Samanthi and Sepanski [2018] Wirch and Hardy [1999]
Wang transform	$\Phi(\Phi^{-1}(u) - \Phi^{-1}(q))$ $0 < q < 1$	-	Wang [2000, 2001] Wozabal [2014]

Table 2: Further examples of distortion risk measures of a random variable X . Table 1 of the online appendix of El Methni and Stupfler [2017] provides these examples of distortion functions; we include this table of examples as a convenient reference for the reader. In the third column, X_1, \dots, X_n denote independent copies of X , $n \in \mathbb{N}$; Y_1, \dots, Y_n are suitable iid random variables satisfying the conditions given in the third column of the table. B denotes the beta function, Φ, Φ^{-1} the distribution and quantile function of the standard normal distribution, respectively.

A.4 Asymptotics of Quantile Estimators in Importance Sampling

The importance sampling estimator in Section 2.1, eq. (2) is studied, along with other alternatives, in Glynn [1996]. We can rewrite the estimator in (2) as

$$\hat{q}_{F^*,N}(u) = \inf \left\{ x \in \mathbb{R} \mid \frac{1}{N} \sum_{i=1}^N \frac{dF}{dF^*}(X_i) \mathbf{1}_{\{h(X_i) \leq x\}} \geq u \right\}.$$

Setting $F^* = F$, the estimator coincides with the crude Monte Carlo estimator of quantiles, the empirical quantile. We analyze under which conditions the estimator in eq. (2) is finite. For this purpose, we first consider a deterministic problem. Let $\xi_1, \dots, \xi_N \in \mathbb{R}$, $\gamma_i \geq 0$ for $i = 1, 2, \dots, N$ and $z > 0$. Then $q := \inf \left\{ x \in \mathbb{R} \mid \sum_{\xi_i > x} \gamma_i \leq z \right\} \in \mathbb{R} \iff \sum \gamma_i > z$. To see this, we observe that $q = -\infty$ is equivalent to $\sum_{\xi_i > x} \gamma_i \leq z$ for all x , which simply means that $\sum \gamma_i \leq z$. This proves the claim, since $q = \infty$ is equivalent to $\sum_{\xi_i > x} \gamma_i > z$ for all x , but for large enough x the sum is empty and equal to 0, contradicting $z > 0$.

The simple characterization implies for $u \in (0, 1)$ that

$$\hat{q}_{F^*,N}(u) \in \mathbb{R} \iff \frac{1}{N} \sum_{i=1}^N \frac{dF}{dF^*}(X_i) > 1 - u \quad (8)$$

Assuming the samples $(X_1, h(X_1)), \dots, (X_N, h(X_N))$ from F^* are independent and identically distributed, we obtain by a law of large numbers that $\frac{1}{N} \sum_{i=1}^N \frac{dF}{dF^*}(X_i) \longrightarrow \mathbf{E}_{F^*} \left[\frac{dF}{dF^*}(X) \right] = 1$, thus eq. (8) is satisfied for N large enough.

The asymptotic normality of the estimator $\hat{q}_{F^*,N}(u)$ can be shown, if the following assumptions hold. This is stated in Theorem 2.1 of Ahn and Shyamalkumar [2011], generalizing Glynn [1996].

Assumption A.12. *Let G, G^* be the distribution functions of $h(X)$, when X is distributed according to F and F^* , respectively.*

Assume that for $u \in (0, 1)$ the following properties hold:

(A1) *G is absolutely continuous with respect to G^* .*

(A2) *G^* is continuous at $q_Y(u)$.*

(A3) *G has a strictly positive first derivative at $q_Y(u)$.*

(A4) *$\frac{dG}{dG^*}(\cdot)$ is a function of finite variation on compacts and has finite negative variation on (y, ∞) for all $y \in \mathbb{R}$.*

(A5) $\frac{dG}{dG^*}(\cdot)$ is right continuous.

(A6) There exists a $\lambda \in (0, 1/2]$ such that

$$\int_y^\infty (1 - G^*(x-))^{1/2-\lambda} d \left| \frac{dG}{dG^*}(x) \right| < \infty \quad \forall y \in \mathbb{R}.$$

Remark A.13. These assumptions of Ahn and Shyamalkumar [2011] are weaker than the assumptions of Glynn [1996] to obtain the implications in Theorem 2.1. Glynn [1996] assumes that (A1) to (A3) hold and $\mathbf{E}_{G^*} \left[\frac{dG}{dG^*}(X)^3 \right] < \infty$. The latter is replaced by assumption (A6), together with the technical conditions (A4) and (A5); here, $|\cdot|$ denotes total variation.¹ Ahn and Shyamalkumar [2011] show that if $\mathbf{E}_{G^*} \left[\frac{dG}{dG^*}^{2+\delta} \right] < \infty$ holds for some $\delta > 0$, then (A6) is satisfied for $\lambda \in (0, \delta/(4 + 2\delta))$.

Proposition A.14. If Assumption A.12 holds, we obtain for $u \in (0, 1)$:

$$\mathbf{E}_{F^*} \left[\frac{dF}{dF^*}(X)^2 \mathbf{1}_{\{h(X) \in (q_Y(u), \infty)\}} \right] \geq (1 - u)^2$$

Proof. By Jensen's inequality,

$$\mathbf{E}_{F^*} \left[\frac{dF}{dF^*}(X)^2 \mathbf{1}_{\{h(X) > q_Y(u)\}} \right] \geq \left(\mathbf{E}_{F^*} \left[\frac{dF}{dF^*}(X) \mathbf{1}_{\{h(X) > q_Y(u)\}} \right] \right)^2.$$

By the Radon–Nikodym theorem, the right-hand side equals

$$\left(\mathbf{E}_F \left[\mathbf{1}_{\{h(X) > q_Y(u)\}} \right] \right)^2 = \mathbf{P}(h(X) > q_Y(u))^2.$$

Since Assumption A.12 implies that G is differentiable, hence continuous, at $q_Y(u)$, we have $G(q_Y(u)) = u$, and therefore $\mathbf{P}(h(X) > q_Y(u)) = 1 - u$. This proves the claim. \square

We now consider the estimation of the quantile $q_Y(1 - \alpha)$, $\alpha \in (0, 1)$, by $\hat{q}_{F_\vartheta, N}(1 - \alpha)$, the estimator defined in Section 2.1, eq. (2). According to Theorem 2.1 we should choose F_ϑ such that

$$\mathbf{E}_{F_\vartheta} \left[\frac{dF}{dF_\vartheta}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha)\}} \right] \tag{9}$$

¹If $f : \mathbb{R} \rightarrow \mathbb{R}$ is a function of bounded variation, there exist two increasing functions $f^+, f^- : \mathbb{R} \rightarrow \mathbb{R}$ with $f = f^+ - f^-$, and $|f| := f^+ + f^-$. The former is called the Jordan decomposition of f and is closely related to the Hahn decomposition of signed measures. For details we refer to Folland [1999].

is small. We consider exponential twists given in eq. (3). The following standard result for the cumulant generating function is useful:

Lemma A.15. *Let $F : \mathbb{R}^d \rightarrow [0, 1]$ be the distribution function of X , $h : \mathbb{R}^d \rightarrow \mathbb{R}$ a measurable function. If $\psi(\vartheta + t) = \log(\mathbb{E}[\exp((\vartheta + t)h(X))]) < \infty$ for all t in some neighborhood of 0, then $\psi'(\vartheta) = \mathbb{E}_{F_\vartheta}[h(X)]$, where F_ϑ is the family of distributions defined in (3).*

To find an appropriate parameter ϑ to make (9) small, we take an approach like in Sun and Hong [2009]. By the definition of F_ϑ samples in the right tail are more likely to occur when $\vartheta > 0$, which also indicates that to minimize (9) we should choose $\vartheta > 0$. We observe that

$$\begin{aligned} \mathbb{E}_{F_\vartheta} \left[\frac{dF}{dF_\vartheta}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha)\}} \right] &= \mathbb{E} \left[\frac{dF}{dF_\vartheta}(X) \mathbf{1}_{\{h(X) > q_Y(1-\alpha)\}} \right] = \mathbb{E} \left[\exp(\psi(\vartheta) - \vartheta h(X)) \mathbf{1}_{\{h(X) > q_Y(1-\alpha)\}} \right] \\ &\leq \exp(\psi(\vartheta) - \vartheta q_Y(1-\alpha)) \cdot \mathbb{P}(h(X) > q_Y(1-\alpha)). \end{aligned}$$

Minimizing the upper bound is then equivalent to minimizing $\psi(\vartheta) - \vartheta q_Y(1-\alpha)$, from which we obtain a first order condition using Lemma A.15

$$q_Y(1-\alpha) = \mathbb{E}_{F_\vartheta}[h(X)]. \quad (10)$$

In their paper, Sun and Hong [2009] show that this approach yields a strict reduction of the objective function (9):

Theorem A.16. *Consider the situation as described above. Assume there exists $\varepsilon > 0$ such that G is differentiable with strictly positive derivative on $(q_Y(1-\alpha) - \varepsilon, q_Y(1-\alpha) + \varepsilon)$. Further suppose that $q_Y(1-\alpha) > \mathbb{E}[h(X)]$, $\frac{dF}{dF_{\vartheta^*}}(x) = \exp(\psi(\vartheta^*) - \vartheta^* h(x))$, and let ϑ^* be chosen such that $q_Y(1-\alpha) = \mathbb{E}_{F_{\vartheta^*}}[h(X)]$. Then $\mathbb{E}_{F_{\vartheta^*}} \left[\frac{dF}{dF_{\vartheta^*}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha)\}} \right] < \mathbb{E} \left[\mathbf{1}_{\{h(X) > q_Y(1-\alpha)\}} \right]$.*

A.5 Refinements, Alternatives, and Clarifications of Algorithm 1

Here, we aim to discuss various details, potential refinements, and alternatives applicable to Algorithm 1.

Remark A.17.

- *Line 12:* The quantity $G'(x)$ needed to define c_i is typically unknown. In the algorithm, we estimate it by a kernel density approximation based on the pivot samples and then evaluate the resulting estimator at the relevant quantile levels. Other estimation methods,

such as finite-difference approximations based on the empirical distribution function, could be used as well.

- Line 16: The normalizing factor $\hat{\psi}(\vartheta_i)$ is estimated from the pivot samples. Due to estimation error, the resulting dF_i is therefore not exactly normalized. In implementations of the Metropolis–Hastings algorithm, it is often more efficient to work with an approximated density than with an unnormalized function. The normalization step in line 16 is discussed in Section 2.4.
- Line 17: Instead of selecting the mixture component by randomly drawing ϑ_i , one could stratify the sampling according to p_i by drawing $p_i N$ samples from each F_i . In the algorithm, we adopt the mixture sampling scheme described above.
- Line 18: The quantile estimation could also be based on the approximation \hat{h} rather than on further evaluations of h . This would reduce computational cost, but at the price of additional approximation error. In this paper, we do not use \hat{h} in the final estimation step, so that the performance of the IS method can be compared directly with that of the crude estimator.
- Line 21: The variance estimation used for the comparison relies on sample averages of the two second moments discussed in Section 2.3.2. Performing this comparison at the end of the algorithm makes it possible to use both the N importance sampling observations and the M pivot samples, which improves accuracy. In our algorithm, the estimation of $q_Y(1 - \alpha_i)$ reuses the estimate from line 7. Alternatively, one could base the quantile estimation on samples drawn under the importance sampling distribution. Both choices lead to viable algorithms; the difference is a design choice balancing computational effort against estimation accuracy.

A.6 Tools from Machine Learning

For convenience, we briefly review the considered ML regression techniques and the methodology of k -fold validation. An excellent introduction to ML is provided by Shalev-Shwartz and Ben-David [2014]. In our simulation algorithm, pivot samples

$$S = (X_1, h(X_1)), \dots, (X_M, h(X_M)) = (X_i, h(X_i))_{i=1, \dots, M}$$

are used as training data.

A.6.1 Linear Predictors

We briefly review linear prediction; this is based on Section 9.2 of Shalev-Shwartz and Ben-David [2014]. For regressions, we consider the hypothesis class

$$\mathcal{H}_{lin} = \left\{ x \mapsto \langle w, x \rangle + b \mid w \in \mathbb{R}^d, b \in \mathbb{R} \right\} \quad (11)$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. One approach is to determine w^*, b^* by empirical risk minimization (ERM) with quadratic loss function. The empirical risk is thus given by $L_S(h) = \sum_{i=1}^M (\langle w, X_i \rangle + b - h(X_i))^2$, where $h \in \mathcal{H}_{lin}$ is the predictor corresponding to w, b . Optimal w^*, b^* are determined by $w^*, b^* = \arg \min_{w, b} \sum_{i=1}^M (\langle w, X_i \rangle + b - h(X_i))^2$. As is well known, the first order conditions lead to a linear problem. Shalev-Shwartz and Ben-David [2014] discuss the application of linear programming and perceptrons (cf. Rosenblatt [1958]).

A.6.2 Polynomial Predictors

Again, for more details, we refer to Shalev-Shwartz and Ben-David [2014], Section 9.2.2. To illustrate the main idea, we assume that the dimension of the training patterns is $d = 1$. The hypothesis class of polynomial predictors with degree k is given as

$$\mathcal{H}_{poly}^k = \left\{ x \mapsto p(x) \mid w \in \mathbb{R}^{k+1} \right\}$$

where $p(x) = w_0 + w_1x + w_2x^2 + \dots + w_kx^k$. Obviously, polynomial predictors can be seen as the application of linear hypotheses to features which are obtained as transformations of the original input patterns, in this case leading to monomials as features. Namely, setting $\psi(x) = (1, x, x^2, \dots, x^k)$, we have $p(x) = \langle w, \psi(x) \rangle$. We can thus apply the same methods on the transformed sample $S' = (\psi(X_i), h(X_i))_{i=1, \dots, M}$ as in the case of linear predictors with ERM specified by $w^* = \arg \min_w \sum_{i=1}^M (\langle w, \psi(X_i) \rangle - h(X_i))^2$.

A.6.3 Support Vector Machines

Support vector machines can be used for classification and regression purposes. A good overview of classification is Chapter 15 of Shalev-Shwartz and Ben-David [2014]. An early extension to regression tasks is Drucker et al. [1996]. For more details see Chapter 6 in Vapnik [1999],

Chapter 2 in Huang et al. [2006], or the tutorial article Smola and Schölkopf [2004] which form the basis for our brief review.

Linear Support Vector Machine Regression First, we consider again the linear predictor hypothesis class (11). A support vector machine regression considers the optimization problem

$$(w^*, b^*) = \arg \min_{w, b} \frac{1}{2} \|w\|^2$$

subject to $|h(X_i) - \langle w, X_i \rangle - b| \leq \varepsilon,$

where $\varepsilon > 0$ is a parameter controlling the tolerated distance of the samples to the predictor; within the tolerance bound, the flatness of the solution is minimized. As the solution to the optimization problem above may not exist, the soft margin concept introduces the slack variables $\xi, \bar{\xi} \in \mathbb{R}^M$ and considers instead

$$(w^*, b^*, \xi^*, \bar{\xi}^*) = \arg \min_{w, b, \xi, \bar{\xi}} \left\{ \frac{1}{2} \|w\|^2 + C \cdot \sum_{i=1}^M (\xi_i + \bar{\xi}_i) \right\},$$

subject to $h(X_i) - (\langle w, X_i \rangle + b) \leq \varepsilon + \xi_i,$

$(\langle w, X_i \rangle + b) - h(X_i) \leq \varepsilon + \bar{\xi}_i,$

$\xi_i, \bar{\xi}_i \geq 0.$

To approximate the solution of the soft margin optimization we use in this paper a sequential optimization described in Platt [1998] and Fan et al. [2005].

The Kernel Trick The summary in this section is based on Chapter 16 of Shalev-Shwartz and Ben-David [2014], Section 6.3 of Vapnik [1999], Section 2.2 of Huang et al. [2006], and Smola and Schölkopf [2004]. When generalizing support vector machines to nonlinear predictors, the same approach as outlined for polynomial predictors can be taken. Instead of considering linear hypotheses on the input space, one considers instead the concatenation of an unknown linear function and a known mapping from the input space to a feature space. Machine learning then determines a suitable linear predictor on the feature space. Good feature spaces can be very high-dimensional and the algorithm might become infeasible.

In the case of support vector machines, a computationally cheaper way is available which relies on the following observation. If linear predictors are learned on a Euclidean space using

support vector optimization, the solution can be determined if scalar products of all elements of the domain of the linear predictors can be computed. Consider, for example, the input space \mathbb{R} and the transformation to features $\psi : \mathbb{R} \rightarrow \mathbb{R}^m$. Replacing the original training samples $S = (X_i, h(X_i))_{i=1, \dots, M}$ by $\hat{S} = (\psi(X_i), h(X_i))_{i=1, \dots, M}$, we seek a support vector linear predictor computed from \hat{S} . Since the solution can be computed from the knowledge of scalar products of features $\langle \psi(x), \psi(x') \rangle = K(x, x')$ which are labeled by inputs, it suffices to specify the corresponding kernel $K : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, but explicit knowledge of ψ is not required. In this article, we again use the sequential optimization described, e.g., in Fan et al. [2005] with two commonly used kernel functions:

Example A.18. (i) *Polynomial kernel:* The kernel $K(x, x') = (1 + \langle x, x' \rangle)^k$ corresponds to

$$K(x, x') = \langle \psi(x), \psi(x') \rangle = \sum_{J \in \{0,1\}^k} \prod_{i=1}^k x_{J_i} \prod_{i=1}^k x'_{J_i},$$

where we define $x_0 = x'_0 = 1$. Then $\psi(x)$ has as components monomials up to degree k , and the SVM will learn a polynomial predictor.

(ii) *Gaussian kernels:* The kernel

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma}\right),$$

for $\sigma > 0$, is called *Gaussian kernel*. The Gaussian kernel corresponds to the embedding $\psi(x)$ with the components

$$\psi(x)_i = \frac{1}{\sqrt{i!}} \exp\left(-\frac{x^2}{2}\right) x^i.$$

A.6.4 k -fold Cross Validation

Based on Section 11.2 of Shalev-Shwartz and Ben-David [2014], we briefly describe k -fold cross validation. The training of different methods was already discussed, now we need a strategy to select among these methods. For this purpose, the training sample S is partitioned into subsets S_1, \dots, S_k , each of size M/k (where k divides M which can easily be realized in the implementation), such that $S_j := (X_i, h(X_i))_{i=j \cdot \frac{M}{k} + 1, \dots, (j+1) \cdot \frac{M}{k}}$. Assume that $r \in \{1, \dots, R\}$ enumerates the different methods considered and/or parameters of these methods, and let $A_r(S)$ be the output of the algorithm trained on the training data S resulting in the predictor h_r . For each r , the algorithm can alternatively be trained on the training data $S \setminus S_j$, $j \in \{1, \dots, k\}$

Hypothesis Class	Hyperparameter	Stop Criterion
Linear Predictors	-	-
Polynomial Predictors of degree q_1	ordered increasing in $q_1 \in \{2, 3, \dots\}$	Overfitting observed
Linear SVM	-	-
Polynomial SVM of degree q_2	ordered increasing in $q_2 \in \{2, 3, \dots\}$	Overfitting observed Fitting computational unfeasible
Gaussian SVM	-	-
k -NN Regression	ordered increasing in $k \in \{1, 2, 3, \dots\}$	Overfitting observed

Table 3: Overview of the hypothesis classes and order of the hypothesis classes considered in the k -fold validation. The stop criterion determines the largest hyperparameter considered for the hypothesis classes.

with output hypothesis $h_{r,j}$. The individual predictors $h_{r,j}$ are validated on the remaining fold of training data, i.e.,

$$\text{error}(r) = \frac{1}{k} \sum_{i=1}^k L_{S_i}(h_{r,i}) = \frac{1}{k} \sum_{i=1}^k \sum_{(x,h(x)) \in S_i} l(h(x), h_{r,i}(x)),$$

where $l(\cdot, \cdot)$ is the considered loss function. In our implementation, we use $l(x, y) = (x - y)^2$ for the purpose of error measurement, although this loss function is not used in SVMs or k -NN. From the estimated errors of the predictors h_r we can then choose the one which is performing best. The hypothesis classes from Sections 3 & 4 are displayed in Table 3.

A.7 Proofs and Calculations

A.7.1 Appendix to Section 2.3.1

Auxiliary computations. Suppose that Assumption A.12 holds. For large enough N_i , we use the approximation from Theorem 2.1 for all i , i.e.,

$$\hat{q}_{F_i, N_i}(1 - \alpha_i) \sim \mathcal{N} \left(q_Y(1 - \alpha_i), \frac{\mathbb{E}_{F_i} \left[\frac{dF}{dF_i}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1 - \alpha_i))^2} \right).$$

For $u \in [\alpha_i, \alpha_{i+1})$ we have

$$\begin{aligned} \mathbb{E} [(q_Y(1 - u) - \hat{q}_Y(1 - u))^2] &= \mathbb{E} [(q_Y(1 - u) - \hat{q}_{F_i, N_i}(1 - \alpha_i))^2] \\ &\approx \mathbb{E} \left[\left(q_Y(1 - u) - q_Y(1 - \alpha_i) - \sqrt{\frac{\mathbb{E}_{F_i} \left[\frac{dF}{dF_i}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1 - \alpha_i))^2}} Z_i \right)^2 \right] \end{aligned}$$

$$\begin{aligned}
&= (q_Y(1-u) - q_Y(1-\alpha_i))^2 - 2(q_Y(1-u) - q_Y(1-\alpha_i)) \sqrt{\frac{\mathbb{E}_{F_i} \left[\frac{dF}{dF_i}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2}} \mathbb{E}[Z_i] \\
&+ \frac{\mathbb{E}_{F_i} \left[\frac{dF}{dF_i}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2} \mathbb{E}[Z_i^2] \\
&= (q_Y(1-u) - q_Y(1-\alpha_i))^2 + \frac{\mathbb{E}_{F_i} \left[\frac{dF}{dF_i}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2}
\end{aligned}$$

where $Z_i, i \in \{0, 1, \dots, m\}$ are i.i.d. standard normals. With this we obtain

$$\begin{aligned}
&\int_0^1 \mathbb{E}[(q_Y(1-u) - \hat{q}_Y(1-u))^2] dg(u) = \sum_{i=0}^m \int_{\alpha_i}^{\alpha_{i+1}} \mathbb{E}[(q_Y(1-u) - \hat{q}_{F_i, N_i}(1-\alpha_i))^2] dg(u) \\
&\approx \sum_{i=0}^m \int_{\alpha_i}^{\alpha_{i+1}} (q_Y(1-u) - q_Y(1-\alpha_i))^2 dg(u) + \frac{\mathbb{E}_{F_i} \left[\frac{dF}{dF_i}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2} (g(\alpha_{i+1}) - g(\alpha_i)).
\end{aligned}$$

□

Proof of Equation (7). Let

$$c_i := \frac{\mathbb{E}_{F_i} \left[\frac{dF}{dF_i}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{G'(q_Y(1-\alpha_i))^2} (g(\alpha_{i+1}) - g(\alpha_i)).$$

The optimization problem becomes to minimize $\sum_{i=0}^m \frac{c_i}{N_i}$ under the constraint $(\sum_{i=0}^m N_i) - N = 0$ with $c_i \geq 0$, since g is increasing and according to Proposition A.14. The Lagrangian for this optimization problem is $\mathcal{L}(N_0, N_1, \dots, N_m; \lambda) = \sum_{i=0}^m \frac{c_i}{N_i} + \lambda (\sum_{i=0}^m N_i - N)$ with gradient

$$\nabla \mathcal{L}(N_0, N_1, \dots, N_m; \lambda) = \left(-\frac{c_0}{N_0^2} + \lambda \quad -\frac{c_1}{N_1^2} + \lambda \quad \dots \quad -\frac{c_m}{N_m^2} + \lambda \quad \sum_{i=0}^m N_i - N \right)^T \stackrel{!}{=} 0.$$

We rewrite the first $m+1$ equations as $\sqrt{\frac{c_i}{\lambda}} = N_i, i = 0, 1, \dots, m$, and plug this into the last equation to obtain $\sum_{i=0}^m \sqrt{\frac{c_i}{\lambda}} = N$ which is equivalent to $\lambda = \left(\frac{1}{N} \sum_{i=0}^m \sqrt{c_i} \right)^2$. This yields the critical point $N_i = N \frac{\sqrt{c_i}}{\sum_{i=0}^m \sqrt{c_i}}, i = 0, 1, \dots, m$. To show that this is the minimum under the constraint it suffices to verify that $\mathcal{L}(N_0, N_1, \dots, N_m; \lambda)$ is a convex function in (N_0, N_1, \dots, N_m) . Rewriting $\mathcal{L}(N_0, N_1, \dots, N_m; \lambda) = \sum_{i=0}^m \left(\frac{c_i}{N_i} + \lambda \left(N_i - \frac{N}{m+1} \right) \right)$, we observe that the functions $\mathcal{L}_i(N') = \frac{c_i}{N'} + \lambda N' - \frac{\lambda N}{m+1}$ are each the sum of convex functions and are therefore convex for $N' \in \mathbb{R}_+$. It follows that $\mathcal{L}(N_0, N_1, \dots, N_m; \lambda)$ is a convex function, implying that the critical point is a minimum. □

A.8 Conditional Sampling

In view of the error bound in (6), note that the second moments are affected only by the change of measure in the tail. This raises the question whether the estimation error can be reduced by conditioning the sampling distributions on the tail. To address this question, we define the conditional sampling distribution as follows:

$$dF_{\vartheta,v}(x) = \exp(\vartheta h(x) - \psi(\vartheta, v)) \mathbb{1}_{\{h(x) \geq q_Y(v)\}} dF(x),$$

where $\psi(\vartheta, v) = \log(\mathbf{E}_F[\exp(\vartheta h(X)) \mathbb{1}_{\{h(X) \geq q_Y(v)\}}])$ and $v \in [0, 1]$.

To define a conditional mixture sampling distribution, let $\alpha_{m'}$ be the largest element of the partition such that $g(\alpha_{m'+1}) - g(\alpha_{m'}) > 0$. Assume further that $\sum p_i = 1$, with $p_i = 0$ for $i < m'$ and $p_i > 0$ for $i \geq m'$. For $v \leq 1 - \alpha_{m'}$, the conditional mixture sampling distribution is given by

$$dF_v^*(x) = \sum_{i=m'}^m p_i dF_{\vartheta_i,v}(x).$$

For both individual and mixture IS estimators, conditioning on the tail reduces the estimation error, as shown in the following proposition.

Proposition A.19. (i) Let $v_i \leq w_i \leq 1 - \alpha_i$, for all $i \in \{0, 1, \dots, m\}$. Then:

$$\begin{aligned} & \mathcal{E}(F_{\vartheta_0,v_0}, F_{\vartheta_1,v_1}, \dots, F_{\vartheta_m,v_m}, \tilde{N}_{Ind}) - \mathcal{E}(F_{\vartheta_0,w_0}, F_{\vartheta_1,w_1}, \dots, F_{\vartheta_m,w_m}, \tilde{N}_{Ind}) \\ &= \sum_{i=0}^m \frac{g(\alpha_{i+1}) - g(\alpha_i)}{N_i G'(q_Y(1 - \alpha_i))^2} \mathbf{E}_F [\exp(\vartheta_i h(X)) \mathbb{1}_{\{h(X) \in (q_Y(v_i), q_Y(w_i)]\}}] \\ & \cdot \mathbf{E}_F [\exp(-\vartheta_i h(X)) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}}] \geq 0, \end{aligned}$$

where $\tilde{N}_{Ind} = (N_0, N_1, \dots, N_m)^T$ and $\sum N_i = N$.

(ii) Let $v \leq w \leq 1 - \alpha_{m'}$, then

$$\begin{aligned} & \mathcal{E}(F_v^*, F_v^*, \dots, F_v^*, \tilde{N}_{Mix}) - \mathcal{E}(F_w^*, F_w^*, \dots, F_w^*, \tilde{N}_{Mix}) \\ &= \sum_{i=m'}^m \frac{g(\alpha_{i+1}) - g(\alpha_i)}{N G'(q_Y(1 - \alpha_i))^2} \\ & \mathbf{E}_F \left[\left(\frac{1}{\sum_{j=m'}^m p_j \exp(\vartheta_j h(X) - \psi(\vartheta_j, v))} - \frac{1}{\sum_{j=m'}^m p_j \exp(\vartheta_j h(X) - \psi(\vartheta_j, w))} \right) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] \\ & \geq 0, \end{aligned}$$

where $\tilde{N}_{Mix} = (N, \dots, N)^T$.

Proof. (i) For $v_i \leq w_i \leq 1 - \alpha_i$, $i = 0, \dots, m$, we compute

$$\begin{aligned} & \mathcal{E}(F_{\vartheta_0, v_0}, F_{\vartheta_1, v_1}, \dots, F_{\vartheta_m, v_m}, \tilde{N}) - \mathcal{E}(F_{\vartheta_0, w_0}, F_{\vartheta_1, w_1}, \dots, F_{\vartheta_m, w_m}, \tilde{N}) = \sum_{i=0}^m \frac{g(\alpha_{i+1}) - g(\alpha_i)}{N_i G'(q_Y(1 - \alpha_i))^2} \\ & \cdot \left(\underbrace{\mathbb{E}_F \left[\frac{dF}{dF_{\vartheta_i, v_i}}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right]}_{(1)} - \mathbb{E}_F \left[\frac{dF}{dF_{\vartheta_i, w_i}}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right]} \right). \end{aligned}$$

By Lemma A.20, (1) ≥ 0 , and since the prefactor is nonnegative, the inequality follows.

Moreover, with

$$\psi(\vartheta_i, v) = \log(\mathbb{E}_F[\exp(\vartheta_i h(X)) \mathbb{1}_{\{h(X) \geq q_Y(v)\}}]),$$

we can rewrite

$$\begin{aligned} (1) &= \mathbb{E}_F \left[\exp(-\vartheta_i h(X)) (\exp(\psi(\vartheta_i, v_i)) - \exp(\psi(\vartheta_i, w_i))) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] \\ &= (\exp(\psi(\vartheta_i, v_i)) - \exp(\psi(\vartheta_i, w_i))) \mathbb{E}_F \left[\exp(-\vartheta_i h(X)) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] \\ &= \mathbb{E}_F \left[\exp(\vartheta_i h(X)) \mathbb{1}_{\{h(X) \in (q_Y(v_i), q_Y(w_i)]\}} \right] \mathbb{E}_F \left[\exp(-\vartheta_i h(X)) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right]. \end{aligned}$$

(ii) The argument for part (ii) is analogous. □

Lemma A.20. *Let $w \leq v \leq 1 - \alpha$. Then:*

$$\mathbb{E}_F \left[\frac{dF}{dF_{\vartheta, v}}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha)\}} \right] \leq \mathbb{E}_F \left[\frac{dF}{dF_{\vartheta, w}}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha)\}} \right]$$

and

$$\mathbb{E}_F \left[\frac{dF}{dF_v^*}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha)\}} \right] \leq \mathbb{E}_F \left[\frac{dF}{dF_w^*}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha)\}} \right].$$

Proof. To show the first inequality, note that since $v \leq 1 - \alpha$, we have

$$\frac{dF}{dF_{\vartheta, v}}(X) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha)\}} = \exp(\psi(\vartheta, v) - \vartheta h(X)) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha)\}},$$

where we define $0/0 = 0$. Given that $\psi(\vartheta, v) = \log(\mathbb{E}_F[\exp(\vartheta h(X)) \mathbb{1}_{\{h(X) \geq q_Y(v)\}}])$, it follows for $w \leq v$ that

$$\psi(\vartheta, v) \leq \psi(\vartheta, w).$$

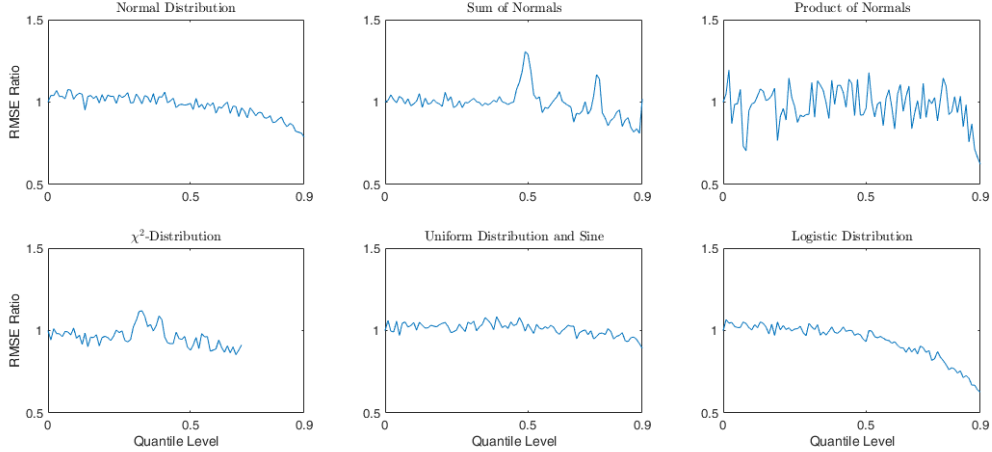


Figure 6: RMSE ratio between unconditional and conditional importance sampling methods for estimating the distortion risk measure $\rho_{g_{\alpha,\gamma}}(Y)$ with $\gamma = 1$ and $\alpha = 0.05$. A total of $N = 20,000$ samples are drawn from the unconditional importance sampling distribution based on the exact function h , and only those samples exceeding the estimated quantiles are considered in the DRM estimation. The plots depict the RMSE ratio as a function of conditioning levels across the six case studies from Section 3. Some curves are truncated at higher levels due to instability in the estimations.

Utilizing the monotonicity of the integral, the first inequality of the lemma follows directly. The second inequality can be shown following the same reasoning. \square

Remark A.21. (i) *The effectiveness of an IS estimation of a DRM with conditioned sampling distributions depends critically on the feasibility of efficient simulation methods. In particular, acceptance–rejection based on $F_{\mathcal{Y}_i}$ as proposal may entail a substantial increase in computational cost, since it typically requires the generation of many candidate draws. Hence, any practical advantage of conditioning hinges on whether the induced sampling overhead is offset by the corresponding variance reduction.*

(ii) *In practice, the quantiles of the conditioned sampling distributions are unknown and must be estimated from the pivot samples produced by the algorithm. Since these estimators are themselves random, instabilities may occur if $\hat{q}_{F,M}(v_i) > q_Y(1 - \alpha_i)$ for some $v_i \leq 1 - \alpha_i$. To mitigate this risk, the choice of v_i should avoid values too close to $1 - \alpha_i$.*

We now turn to numerical case studies of conditional importance sampling, adopting the framework of Section 3 for estimating the distortion risk measure $\rho_{g_{\alpha,\gamma}}(Y)$ with parameters $\gamma = 1$ and $\alpha = 0.05$ across the six case studies. For each importance sampling method, we generate $M = 5,000$ pivot samples. These samples serve to estimate the conditioning quantile $q_Y(v)$ by $\hat{q}_{F,M}(v)$, which is then used for conditioning the sampling distribution. Here v denotes

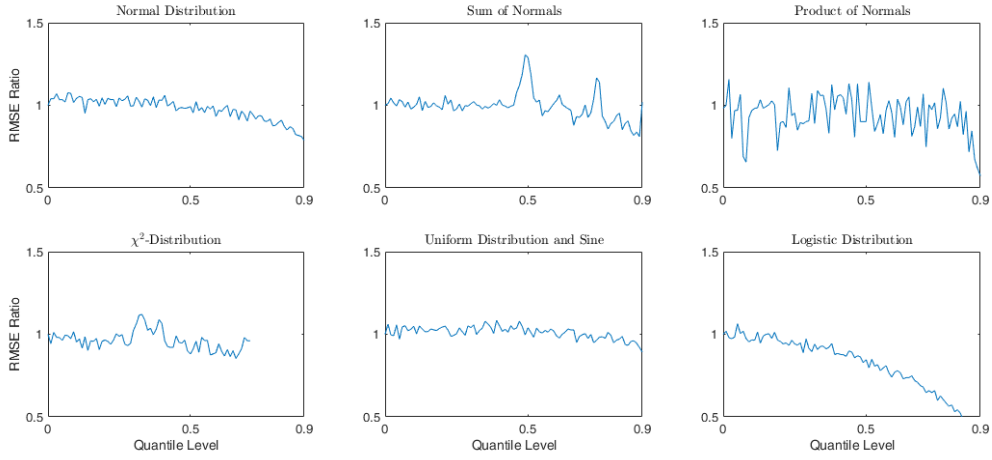


Figure 7: RMSE ratio between unconditional and conditional importance sampling methods for estimating the distortion risk measure $\rho_{g_{\alpha,\gamma}}(Y)$ with $\gamma = 1$ and $\alpha = 0.05$. Exactly $N = 20,000$ samples are drawn from the conditional importance sampling distribution based on the exact function h . The plots depict the RMSE ratio as a function of conditioning levels across the six case studies from Section 3. Some curves are truncated at higher levels due to instability in the estimations.

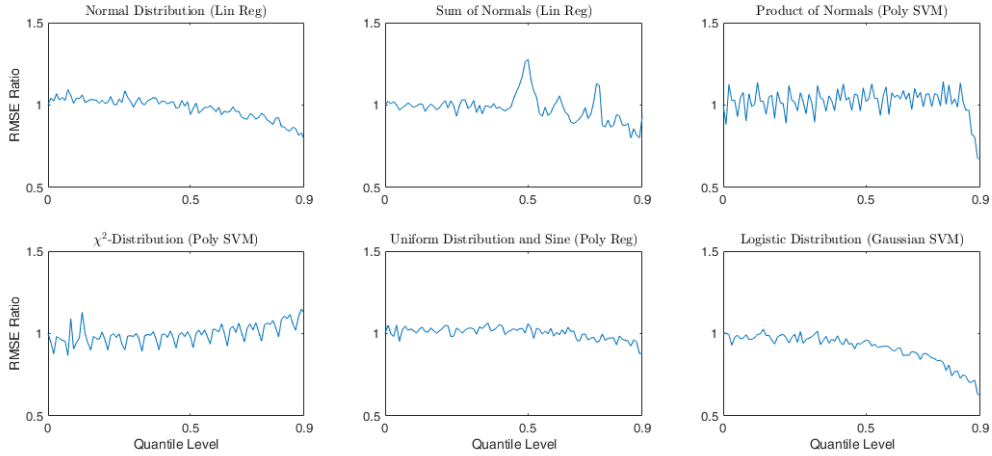


Figure 8: RMSE ratio between unconditional and conditional importance sampling methods for estimating the distortion risk measure $\rho_{g_{\alpha,\gamma}}(Y)$ with $\gamma = 1$ and $\alpha = 0.05$. A total of $N = 20,000$ samples are drawn from the unconditional importance sampling distribution based on the best performing approximation \hat{h} of h from the case studies in Section 3, and only those samples exceeding the estimated quantiles are considered in the DRM estimation. The plots depict the RMSE ratio as a function of conditioning levels across the six case studies from Section 3. Some curves are truncated at higher levels due to instability in the estimations.

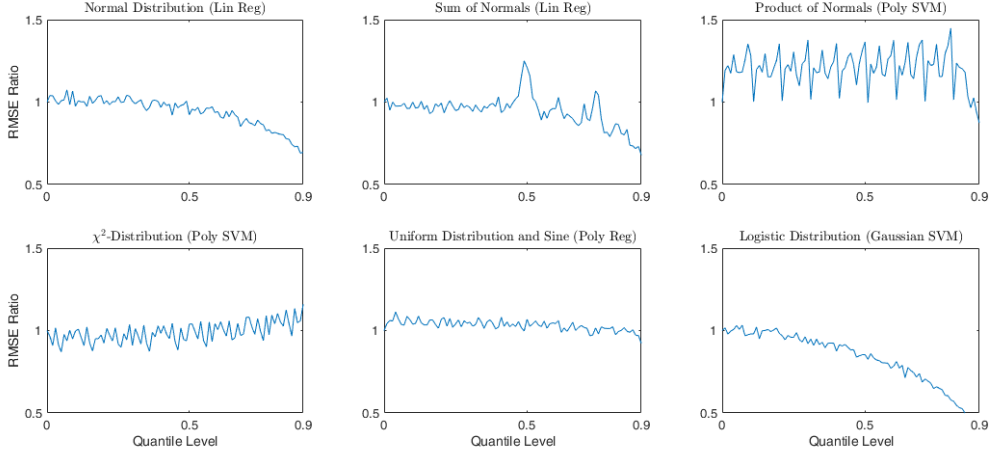


Figure 9: RMSE ratio between unconditional and conditional importance sampling methods for estimating the distortion risk measure $\rho_{g_{\alpha,\gamma}}(Y)$ with $\gamma = 1$ and $\alpha = 0.05$. Exactly $N = 20,000$ samples are drawn from the conditional importance sampling distribution based on the best performing approximation \hat{h} of h from the case studies in Section 3. The plots depict the RMSE ratio as a function of conditioning levels across the six case studies from Section 3. Some curves are truncated at higher levels due to instability in the estimations.

a quantile level, so larger values of v correspond to more restrictive conditioning on the upper tail.

- In the indirect conditional sampling method, the algorithm draws $N = 20,000$ samples from the unconditional importance sampling distribution and retains only those exceeding $\hat{q}_{F,M}(v)$ for the DRM estimation.
- In contrast, the direct conditional sampling method generates $N = 20,000$ samples from the conditional importance sampling distribution corresponding to the threshold $\hat{q}_{F,M}(v)$.

The remaining algorithms are implemented in accordance with Section 3.

A.8.1 Indirect Conditional Importance Sampling: Results

The results of the case studies are summarized in Figures 6 and 8. Figure 6 presents the ratio of the RMSE for the indirect conditional importance sampling estimation compared to the unconditional approach, based on the exact knowledge of h . In the normal distribution, sum of normals, uniform distribution and sine, and logistic distribution case studies, the RMSE ratio generally decreases with v , indicating that the RMSE reduction improves as the conditioning quantile increases. Notably, in the logistic distribution case, the RMSE is reduced by nearly 40% due to conditioning. In contrast, for the product of normals, the RMSE ratio exhibits significant fluctuations, preventing a clear trend to be observed. For the χ^2 -distribution, the

estimation becomes unstable at higher values of v , making it impossible to determine a reliable ratio. Figure 8 displays results where the importance sampling distributions are based on the best-performing approximations of h from the case studies in Section 3. Here, the RMSE ratio decreases with v in all but one case study, as the conditioning quantile increases. The sole exception is the χ^2 -distribution, where the RMSE ratio appears to increase with increasing v .

A.8.2 Direct Conditional Importance Sampling: Results

The results of the case studies are summarized in Figures 7 and 9. Figure 7 displays the RMSE ratio for the direct conditional importance sampling approach, based on the exact knowledge of h . Similar to the indirect conditional importance sampling method, in the normal distribution, sum of normals, uniform distribution and sine, and logistic distribution case studies, the RMSE ratio decreases with v . This indicates a reduction in RMSE as the conditioning quantile becomes larger. For the product of normals, the ratio fluctuates without showing a clear trend. At higher v values, estimating the DRM for the χ^2 -distribution becomes unstable. Figure 9 illustrates the RMSE ratio using importance sampling distributions based on the best-performing approximation of h from the case studies in Section 3. Once again, the RMSE ratio decreases with v in all but one case study, corresponding to an increase in the conditioning quantile. The χ^2 -distribution is the exception, with an increasing RMSE.

A.9 Choosing the Size of the Partition

The partition $(\alpha_i)_{i \in \{0,1,\dots,m\}}$ plays an important role in the estimation of the DRM. A finer partition reduces the discretization error, but increasing the partition size may also increase the estimation error of the individual quantiles. The choice of m therefore reflects a trade-off between these two sources of error.

This trade-off is illustrated in the following remark:

Remark A.22. Consider the partition $0 = \alpha_0 < \alpha_1 < \dots < \alpha_m < \alpha_{m+1} = 1$, with $\beta \in (\alpha_j, \alpha_{j+1})$, and sampling distributions $F_{\vartheta_0}, F_{\vartheta_1}, \dots, F_{\vartheta_m}, F_{\vartheta_\beta}$ satisfying Assumption A.12. Define

$$\tilde{N} = (N_0, N_1, \dots, N_m), \quad \tilde{N}' = (N_0, \dots, N_{j-1}, N'_j, N'_{j+1}, N_{j+1}, \dots, N_m),$$

where $N'_j + N'_{j+1} = N_j$. Then

$$\begin{aligned}
& \mathcal{E}(F_{\vartheta_0}, F_{\vartheta_1}, \dots, F_{\vartheta_m}, \tilde{N}) - \mathcal{E}(F_{\vartheta_0}, F_{\vartheta_1}, \dots, F_{\vartheta_j}, F_{\vartheta_\beta}, F_{\vartheta_{j+1}}, \dots, F_{\vartheta_m}, \tilde{N}') \\
&= \int_{\beta}^{\alpha_{j+1}} \left((q_Y(1-u) - q_Y(1-\alpha_j))^2 - (q_Y(1-u) - q_Y(1-\beta))^2 \right) dg(u) \\
&+ \frac{\mathbb{E}_{F_{\vartheta_j}} \left[\frac{dF}{dF_{\vartheta_j}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_j)\}} \right] - \alpha_j^2}{G'(q_Y(1-\alpha_j))^2} \left(\frac{g(\alpha_{j+1}) - g(\alpha_j)}{N_j} - \frac{g(\beta) - g(\alpha_j)}{N'_j} \right) \\
&- \frac{\mathbb{E}_{F_{\vartheta_\beta}} \left[\frac{dF}{dF_{\vartheta_\beta}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\beta)\}} \right] - \beta^2}{G'(q_Y(1-\beta))^2} \frac{g(\alpha_{j+1}) - g(\beta)}{N'_{j+1}}.
\end{aligned}$$

In particular,

$$\mathcal{E}(F_{\vartheta_0}, F_{\vartheta_1}, \dots, F_{\vartheta_m}, \tilde{N}) \geq \mathcal{E}(F_{\vartheta_0}, F_{\vartheta_1}, \dots, F_{\vartheta_j}, F_{\vartheta_\beta}, F_{\vartheta_{j+1}}, \dots, F_{\vartheta_m}, \tilde{N}')$$

whenever

$$\begin{aligned}
& \int_{\beta}^{\alpha_{j+1}} \left((q_Y(1-u) - q_Y(1-\beta))^2 - (q_Y(1-u) - q_Y(1-\alpha_j))^2 \right) dg(u) \\
&\geq \frac{\mathbb{E}_{F_{\vartheta_j}} \left[\frac{dF}{dF_{\vartheta_j}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_j)\}} \right] - \alpha_j^2}{G'(q_Y(1-\alpha_j))^2} \left(\frac{g(\alpha_{j+1}) - g(\alpha_j)}{N_j} - \frac{g(\beta) - g(\alpha_j)}{N'_j} \right) \\
&- \frac{\mathbb{E}_{F_{\vartheta_\beta}} \left[\frac{dF}{dF_{\vartheta_\beta}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\beta)\}} \right] - \beta^2}{G'(q_Y(1-\beta))^2} \frac{g(\alpha_{j+1}) - g(\beta)}{N'_{j+1}}.
\end{aligned}$$

To study how the MSE depends on the partition size and the sample size, we establish the following result.

Proposition A.23. *Assume $Y \in L^\infty$ with a continuous quantile function $q_Y(\cdot)$ and a continuous distortion function g . Define the partition α_i , $i \in \{0, 1, \dots, m+1\}$, such that $g(\alpha_{i+1}) - g(\alpha_i) = \frac{1}{m+1}$. Let $\vartheta(\cdot) : [0, 1] \rightarrow \mathbb{R}$ and $\tilde{N} = (N_0, N_1, \dots, N_m)$. Then*

$$\mathcal{E}(F_{\vartheta(\alpha_0)}, F_{\vartheta(\alpha_1)}, \dots, F_{\vartheta(\alpha_m)}, \tilde{N}) \leq \frac{C_m}{m+1} + \frac{D_m}{m+1} \sum_{i=0}^m \frac{1}{N_i},$$

where $C_m \geq 0$ depends only on the partition and the distribution of Y , with $\lim_{m \rightarrow \infty} C_m = 0$, and

$$D_m := \max_{i=0, \dots, m} \frac{\mathbb{E}_{F_{\vartheta(\alpha_i)}} \left[\frac{dF}{dF_{\vartheta(\alpha_i)}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{G'(q_Y(1-\alpha_i))^2}.$$

In particular, if $N_i = \frac{N}{m+1}$ for all i , then

$$\mathcal{E}(F_{\vartheta(\alpha_0)}, F_{\vartheta(\alpha_1)}, \dots, F_{\vartheta(\alpha_m)}, \tilde{N}) \leq \frac{C_m}{m+1} + \frac{(m+1)D_m}{N}.$$

Proof. Consider the estimation error:

$$\begin{aligned} & \mathcal{E}(F_{\vartheta(\alpha_0)}, F_{\vartheta(\alpha_1)}, \dots, F_{\vartheta(\alpha_m)}, \tilde{N}) \\ &= \underbrace{\sum_{i=0}^m \int_{\alpha_i}^{\alpha_{i+1}} (q_Y(1-u) - q_Y(1-\alpha_i))^2 dg(u)}_{(i)} + \underbrace{\sum_{i=0}^m \frac{\mathbb{E}_{F_{\vartheta(\alpha_i)}} \left[\frac{dF}{dF_{\vartheta(\alpha_i)}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2} (g(\alpha_{i+1}) - g(\alpha_i))}_{(ii)}. \end{aligned}$$

We analyze the discretization error (i) and estimation error (ii) as follows:

- (i) For the discretization error, since $q_Y(1-u)$ is decreasing in u and $g(\alpha_{i+1}) - g(\alpha_i) = \frac{1}{m+1}$, we find:

$$\begin{aligned} & \sum_{i=0}^m \int_{\alpha_i}^{\alpha_{i+1}} (q_Y(1-u) - q_Y(1-\alpha_i))^2 dg(u) \\ & \leq \sum_{i=0}^m \int_{\alpha_i}^{\alpha_{i+1}} (q_Y(1-\alpha_i) - q_Y(1-\alpha_{i+1}))^2 dg(u) \\ & = \sum_{i=0}^m (q_Y(1-\alpha_i) - q_Y(1-\alpha_{i+1}))^2 (g(\alpha_{i+1}) - g(\alpha_i)) \\ & = \frac{1}{m+1} \sum_{i=0}^m (q_Y(1-\alpha_i) - q_Y(1-\alpha_{i+1}))^2 \\ & \leq \frac{1}{m+1} \max_{i \in \{0,1,\dots,m\}} (q_Y(1-\alpha_i) - q_Y(1-\alpha_{i+1})) \sum_{i=0}^m (q_Y(1-\alpha_i) - q_Y(1-\alpha_{i+1})) \\ & \leq \frac{1}{m+1} C'_m \sum_{i=0}^m q_Y(1-\alpha_i) - q_Y(1-\alpha_{i+1}) \\ & \leq \frac{1}{m+1} C'_m 2 \|Y\|_\infty =: \frac{C_m}{m+1}. \end{aligned}$$

Here, C'_m is defined by $C'_m := \sup_{\substack{0 \leq v \leq u \leq 1 \\ u-v \leq \max_i \{\alpha_{i+1} - \alpha_i\}}} q_Y(1-v) - q_Y(1-u)$. The continuity of $q_Y(1-u)$ implies $\lim_{m \rightarrow \infty} C'_m = 0$ and hence $\lim_{m \rightarrow \infty} C_m = 0$.

- (ii) Given $g(\alpha_{i+1}) - g(\alpha_i) = \frac{1}{m+1}$, we get:

$$\begin{aligned} & \sum_{i=0}^m \frac{\mathbb{E}_{F_{\vartheta(\alpha_i)}} \left[\frac{dF}{dF_{\vartheta(\alpha_i)}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2} (g(\alpha_{i+1}) - g(\alpha_i)) \\ &= \frac{1}{m+1} \sum_{i=0}^m \frac{\mathbb{E}_{F_{\vartheta(\alpha_i)}} \left[\frac{dF}{dF_{\vartheta(\alpha_i)}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2} \end{aligned}$$

$$\leq \frac{D_m}{m+1} \sum_{i=0}^m \frac{1}{N_i}.$$

Assuming $N_i = \frac{N}{m+1}$, we find $\frac{D_m}{m+1} \sum_{i=0}^m \frac{1}{N_i} = \frac{(m+1)D_m}{N}$.

Combining results (i) and (ii), we deduce bounds for $\mathcal{E}(F_{\vartheta(\alpha_0)}, F_{\vartheta(\alpha_1)}, \dots, F_{\vartheta(\alpha_m)}, \tilde{N})$ as stated in the proposition. \square

Remark A.24. (i) *The comparison in Remark A.22 is influenced by several factors such as $G'(\cdot)$, the density of the distribution of $h(X)$, the chosen distortion function $g(\cdot)$, and the reduction in the second moment of the individual sampling distributions. Generally, refining the partition improves estimation performance only if the reduction in discretization error outweighs the additional estimation error from individual quantiles.*

(ii) *The inequality in Proposition A.23 highlights a trade-off between discretization and estimation errors in choosing the partition size m and the sample sizes N_i . Increasing m refines the partition and thereby reduces the discretization term governed by C_m , but it may also enlarge the estimation term through both the larger number of summands and the dependence of D_m on the partition points. Enlarging the sample sizes N_i decreases the estimation term, but leaves the discretization term unaffected. Thus, the bound reflects a balance between refinement of the partition and the available sampling budget.*

(iii) *The convergence behavior in Proposition A.23 depends on two quantities. First, it depends on how quickly C_m tends to zero, which is determined by the modulus of continuity of the quantile function. Second, the usefulness of the bound also depends on the behavior of D_m as m increases. Under the present assumptions, no general rate can be inferred for C_m , and no uniform boundedness of D_m is guaranteed.*

A.10 Estimation Error under Black Box Approximations

Using a machine learning approximation \hat{h} of h perturbs the likelihood ratios in the importance sampling scheme and thereby affects the corresponding DRM estimation error. The next result makes this effect explicit.

Proposition A.25. *Let \hat{h} be an approximation of h . Define*

$$d\hat{F}_{\vartheta_i}(x) = \exp(\vartheta_i \hat{h}(x) - \hat{\psi}(\vartheta_i)) dF(x),$$

$$d\hat{F}^*(x) = \sum_{i=0}^m p_i d\hat{F}_{\vartheta_i}(x),$$

where $i \in \{0, 1, \dots, m\}$ and

$$\hat{\psi}(\vartheta) = \log\left(\mathbb{E}_F[\exp(\vartheta \hat{h}(X))]\right).$$

Suppose Assumption A.12 holds. Then:

(i)

$$\begin{aligned} & \mathcal{E}(F_{\vartheta_0}, F_{\vartheta_1}, \dots, F_{\vartheta_m}, \bar{N}_{Ind}) - \mathcal{E}(\hat{F}_{\vartheta_0}, \hat{F}_{\vartheta_1}, \dots, \hat{F}_{\vartheta_m}, \bar{N}_{Ind}) \\ &= \sum_{i=0}^m \frac{g(\alpha_{i+1}) - g(\alpha_i)}{N_i G'(q_Y(1 - \alpha_i))^2} \\ & \quad \cdot \mathbb{E}_F \left[\exp(\psi(\vartheta_i) - \vartheta_i h(X)) \left(1 - \exp(\hat{\psi}(\vartheta_i) - \psi(\vartheta_i) + \vartheta_i(h(X) - \hat{h}(X))) \right) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right], \end{aligned}$$

where $\bar{N}_{Ind} = (N_0, N_1, \dots, N_m)^T$ and $\sum_{i=0}^m N_i = N$.

(ii)

$$\begin{aligned} & \mathcal{E}(F^*, F^*, \dots, F^*, \bar{N}_{Mix}) - \mathcal{E}(\hat{F}^*, \hat{F}^*, \dots, \hat{F}^*, \bar{N}_{Mix}) \\ &= \sum_{i=0}^m \frac{g(\alpha_{i+1}) - g(\alpha_i)}{N G'(q_Y(1 - \alpha_i))^2} \\ & \quad \cdot \mathbb{E}_F \left[\left(\frac{1}{\sum_{j=0}^m p_j \exp(\vartheta_j h(X) - \psi(\vartheta_j))} - \frac{1}{\sum_{j=0}^m p_j \exp(\vartheta_j \hat{h}(X) - \hat{\psi}(\vartheta_j))} \right) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right], \end{aligned}$$

where $\bar{N}_{Mix} = (N, N, \dots, N)^T$.

Proof. For arbitrary sampling distributions $F_i, F'_i, i \in \{0, 1, \dots, m\}$, satisfying Assumption A.12, and a sample allocation $\tilde{N} = (N_0, N_1, \dots, N_m)^T$, we have

$$\begin{aligned} & \mathcal{E}(F_0, F_1, \dots, F_m, \tilde{N}) - \mathcal{E}(F'_0, F'_1, \dots, F'_m, \tilde{N}) \\ &= \sum_{i=0}^m \frac{g(\alpha_{i+1}) - g(\alpha_i)}{N_i G'(q_Y(1 - \alpha_i))^2} \mathbb{E}_F \left[\left(\frac{dF}{dF_i}(X) - \frac{dF}{dF'_i}(X) \right) \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right]. \end{aligned}$$

Applying this identity with $F_i = F_{\vartheta_i}$ and $F'_i = \hat{F}_{\vartheta_i}$ yields part (i), since

$$\frac{dF}{dF_{\vartheta}}(x) - \frac{dF}{d\hat{F}_{\vartheta}}(x) = \exp(\psi(\vartheta) - \vartheta h(x)) - \exp(\hat{\psi}(\vartheta) - \vartheta \hat{h}(x))$$

$$= \exp(\psi(\vartheta) - \vartheta h(x)) \left(1 - \exp(\hat{\psi}(\vartheta) - \psi(\vartheta) + \vartheta(h(x) - \hat{h}(x))) \right).$$

Likewise, applying the same identity with $F_i = F^*$ and $F'_i = \hat{F}^*$ for all i yields part (ii), since

$$\frac{dF}{dF^*}(x) - \frac{dF}{d\hat{F}^*}(x) = \frac{1}{\sum_{j=0}^m p_j \exp(\vartheta_j h(x) - \psi(\vartheta_j))} - \frac{1}{\sum_{j=0}^m p_j \exp(\vartheta_j \hat{h}(x) - \hat{\psi}(\vartheta_j))}.$$

This proves the proposition. \square

A.11 Time Efficiency for Individual Quantile Estimations

In Section 2.5, we analyzed the time efficiency of the crude and mixture IS estimators. We now turn to the time efficiency of individual importance sampling, as quantified by the error bound (12). We denote by $T_{CR}(N_{CR}, m)$ and $T_{IS}(M, N_{IS}, m)$ the computation times of the crude and the importance sampling estimators, respectively. The parameters have the following roles: N_{CR} is the crude sample size, M is the number of pivot samples, N_{IS} the importance sampling size, and m the partition size. We determine the sample sizes under the condition that the crude and importance sampling estimators share the same approximate upper error bound according to (12).

Lemma A.26. *Let F_0, F_1, \dots, F_m be distribution functions. Assume that for all $i \in \{0, 1, \dots, m\}$ F is absolutely continuous with respect to F_i and Assumptions A.12 hold. Then, if the samples are drawn i.i.d. from the F_i according to the allocation $\tilde{N}_{IS} = (N_0, N_1, \dots, N_m)$ we have:*

$$\tilde{\mathcal{E}}(F, \dots, F, \tilde{N}_{CR}) = \tilde{\mathcal{E}}(F_0, F_1, \dots, F_m, \tilde{N}_{IS}) \iff N_{CR} = \frac{\sum_{i=0}^m \tilde{V}(1 - \alpha_i, F)(g(\alpha_{i+1}) - g(\alpha_i))^2}{\sum_{i=0}^m \frac{\tilde{V}(1 - \alpha_i, F_i)}{N_i}(g(\alpha_{i+1}) - g(\alpha_i))^2}.$$

Proof. We have:

$$\begin{aligned} & \tilde{\mathcal{E}}(F, \dots, F, \tilde{N}_{CR}) = \tilde{\mathcal{E}}(F_0, F_1, \dots, F_m, \tilde{N}_{IS}) \\ \iff & \sqrt{\frac{1}{N_{CR}} \sum_{i=0}^m \tilde{V}(1 - \alpha_i, F)(g(\alpha_{i+1}) - g(\alpha_i))^2} = \sqrt{\sum_{i=0}^m \frac{\tilde{V}(1 - \alpha_i, F_i)}{N_i}(g(\alpha_{i+1}) - g(\alpha_i))^2} \\ \iff & N_{CR} = \frac{\sum_{i=0}^m \tilde{V}(1 - \alpha_i, F)(g(\alpha_{i+1}) - g(\alpha_i))^2}{\sum_{i=0}^m \frac{\tilde{V}(1 - \alpha_i, F_i)}{N_i}(g(\alpha_{i+1}) - g(\alpha_i))^2}. \end{aligned}$$

\square

By accounting for the same computational components as in Section 2.5, we obtain the

following result:

Proposition A.27. *Assume $N_{CR} > M + N_{IS}$, and let N_{CR} be chosen according to Lemma A.26.*

Then

$$T_{CR}(N_{CR}, m) - T_{IS}(M, N_{IS}, m) > 0$$

whenever

$$t_h(N_{CR} - (N_{IS} + M)) > t_{Mix}(M, m) + t_{kFold}(M) + t_{MH}(N_{IS}) + t_{Norm}(m).$$

Proof. This is analogous to the proof of Proposition 2.3. □

A.12 Computational Resources

All case studies were implemented in MATLAB and executed on the cluster system of Leibniz Universität Hannover, using nodes of varying specifications. In every case study, the dominant contribution to computation time arose from the evaluation of the normalizing constant.

For the case studies of Section 3 using the quadrature methods of Section 2.4, the corresponding ranges of computation times are reported in Table 4. The estimation times are sensitive to the

Method	Fastest Calculation	Slowest Calculation
Exact IS	1s - Normal Distribution	15s - Unif. Dist. and Sine
Gaussian SVM IS	67s - Normal Distribution	1463s - χ^2 -Distribution
k -NN IS	7s - Logistic Distribution	540s - Product of Normals
Lin Reg IS	3s - Logistic Distribution	30s - χ^2 -Distribution
Poly SVM IS	2s - Normal Distribution	750s - χ^2 -Distribution

Table 4: Slowest and fastest calculation times for the case studies in Section 3.

choice of hyperparameters in the ML components, in particular for polynomial SVMs and k -NN regression.

Using kernel-smoothing density estimates at 100 randomly chosen points for the normalizing constant reduces computation times to below 90s in all cases.

The estimates for the iterative exploration of the extreme tail in Section 3.3 were implemented with the quadrature formulas outlined in Section 2.4. For the identity of normals, the DRM estimation required about 5s for both the non-iterative and iterative IS methods. Estimating the sum of normals took 5 seconds with the non-iterative IS and 15 seconds with the iterative IS. For the product of normals, computation times were 48 seconds for the non-iterative IS and 120 seconds for the iterative IS. Finally, estimating the sum of squared normals required 493

seconds for the non-iterative IS and 1107 seconds for the iterative IS.

In the ALM case studies discussed in Section 4, the estimation of the DRMs with exact IS required 116s, with the Gaussian SVM IS 245s, with the k -NN IS 99s, with the linear regression IS 131s and with the linear SVM 284s.

A.13 A Sharper Error Bound for Individual Importance Sampling

The error bound (6) does not take advantage of the independence of the quantile estimates when individual IS is applied. Consequently, the error bound in (6) remains applicable to both individual and pooled sample allocation, as discussed in Section 2.3.2. However, by concentrating on IS sampling and leveraging the independence of the quantile estimates, we can derive a more precise inequality for the MSE in the DRM estimation.

Proposition A.28. *Consider the estimator $\hat{\rho}_g(Y)$ defined in (5) and the individual sample allocation. Then:*

$$\begin{aligned} \sqrt{\mathbb{E}[(\rho_g(Y) - \hat{\rho}_g(Y))^2]} &\lesssim DE(m, \tilde{N}) + \sqrt{\sum_{i=0}^m \frac{\tilde{V}(1 - \alpha_i, F_{\vartheta_i^*})}{N_i} (g(\alpha_{i+1}) - g(\alpha_i))^2}, \\ &=: \tilde{\mathcal{E}}(F_{\vartheta_0^*}, F_{\vartheta_1^*}, \dots, F_{\vartheta_m^*}, \tilde{N}, m), \end{aligned} \quad (12)$$

where:

$$\begin{aligned} DE(m, \tilde{N}) &= \sum_{i=0}^m q_Y(1 - \alpha_i) (g(\alpha_{i+1}) - g(\alpha_i)) - \rho_g(Y), \\ \tilde{V}(1 - \alpha_i, F_{\vartheta_i^*}) &= \frac{\mathbb{E}_{F_{\vartheta_i^*}} \left[\frac{dF}{dF_{\vartheta_i^*}}(X)^2 \mathbb{1}_{\{h(X) > q_Y(1 - \alpha_i)\}} \right] - \alpha_i^2}{G'(q_Y(1 - \alpha_i))^2}. \end{aligned}$$

Proof. Let $\bar{q}_Y(1 - u) = \sum_{i=0}^m \mathbb{1}_{\{u \in [\alpha_i, \alpha_{i+1})\}} q_Y(1 - \alpha_i)$. Applying the triangle inequality, we obtain:

$$\begin{aligned} &\sqrt{\mathbb{E}[(\rho_g(Y) - \hat{\rho}_g(Y))^2]} \\ &\leq \underbrace{\sqrt{\mathbb{E} \left[\left(\int_0^1 q_Y(1 - u) - \bar{q}_Y(1 - u) dg(u) \right)^2 \right]}}_{=: DE(m, \tilde{N})} + \underbrace{\sqrt{\mathbb{E} \left[\left(\int_0^1 \bar{q}_Y(1 - u) - \hat{q}_Y(1 - u) dg(u) \right)^2 \right]}}_{(*)}. \end{aligned}$$

For the discretization error, we express:

$$\begin{aligned}
DE(m, \tilde{N}) &= \left(\mathbb{E} \left[\left(\sum_{i=0}^m \int_{[\alpha_i, \alpha_{i+1})} q_Y(1-u) - q_Y(1-\alpha_i) dg(u) \right)^2 \right] \right)^{\frac{1}{2}} \\
&= \left| \sum_{i=0}^m \int_{[\alpha_i, \alpha_{i+1})} q_Y(1-u) - q_Y(1-\alpha_i) dg(u) \right| \\
&= \sum_{i=0}^m q_Y(1-\alpha_i)(g(\alpha_{i+1}) - g(\alpha_i)) - \int_0^1 q_Y(1-u) dg(u) \\
&= \sum_{i=0}^m q_Y(1-\alpha_i)(g(\alpha_{i+1}) - g(\alpha_i)) - \rho_g(Y).
\end{aligned}$$

For the estimation error, using Theorem 2.1 and the independence of the quantile estimators, we have:

$$\begin{aligned}
(*)^2 &= \mathbb{E} \left[\left(\sum_{i=0}^m q_Y(1-\alpha_i) - \hat{q}_{F_{\vartheta_i^*}, N_i}(1-\alpha_i) \right) (g(\alpha_{i+1}) - g(\alpha_i)) \right]^2 \\
&= \sum_{i=0}^m \frac{\mathbb{E}_{F_{\vartheta_i^*}} \left[\frac{dF}{dF_{\vartheta_i^*}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}{N_i G'(q_Y(1-\alpha_i))^2} (g(\alpha_{i+1}) - g(\alpha_i))^2.
\end{aligned}$$

□

Following the approach to derive (7), we obtain the optimal allocation by minimizing $\sum_{i=0}^m \frac{\tilde{V}(1-\alpha_i, F_{\vartheta_i^*})}{N_i} (g(\alpha_{i+1}) - g(\alpha_i))^2$ under the constraint $\sum_{i=0}^m N_i = N$. Using calculations similar to those in Appendix A.7.1, we find:

$$\tilde{N}_i^* = N \frac{\tilde{c}_i}{\sum_{j=0}^m \tilde{c}_j}, \quad i = 0, 1, \dots, m,$$

where

$$\tilde{c}_i = \frac{\sqrt{\mathbb{E}_{F_{\vartheta_i^*}} \left[\frac{dF}{dF_{\vartheta_i^*}}(X)^2 \mathbf{1}_{\{h(X) > q_Y(1-\alpha_i)\}} \right] - \alpha_i^2}}{G'(q_Y(1-\alpha_i))} (g(\alpha_{i+1}) - g(\alpha_i)).$$

A.14 Additional Plots

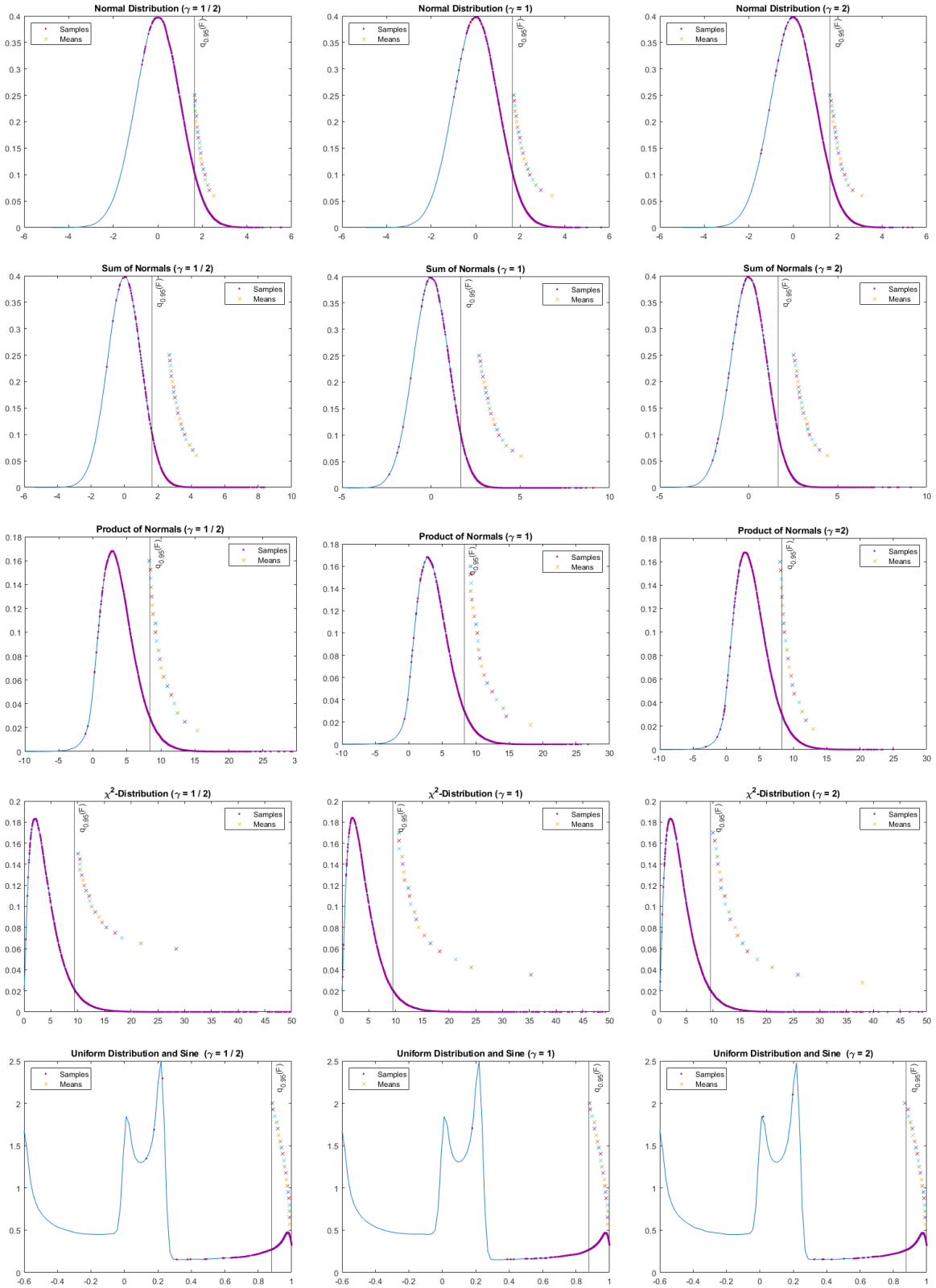


Figure 10: 200 samples drawn from the mixture distribution plotted on the underlying distribution of the model Y for the case studies (1) to (6). To approximate the mixture weights and optimal mixture components $M = 20,000$ pivot samples were drawn. For the estimation of the quantile and DRM $N = 100,000$ samples are drawn from the mixture distribution. For further details, see Section 3.2.1.

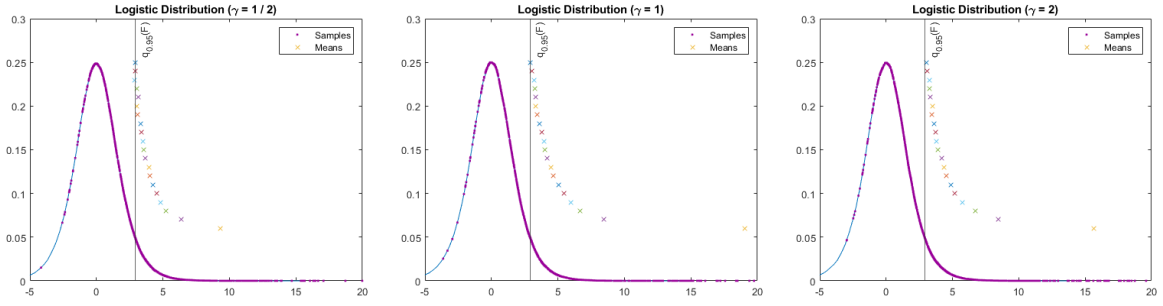


Figure 11: Continuation of Figure 10.

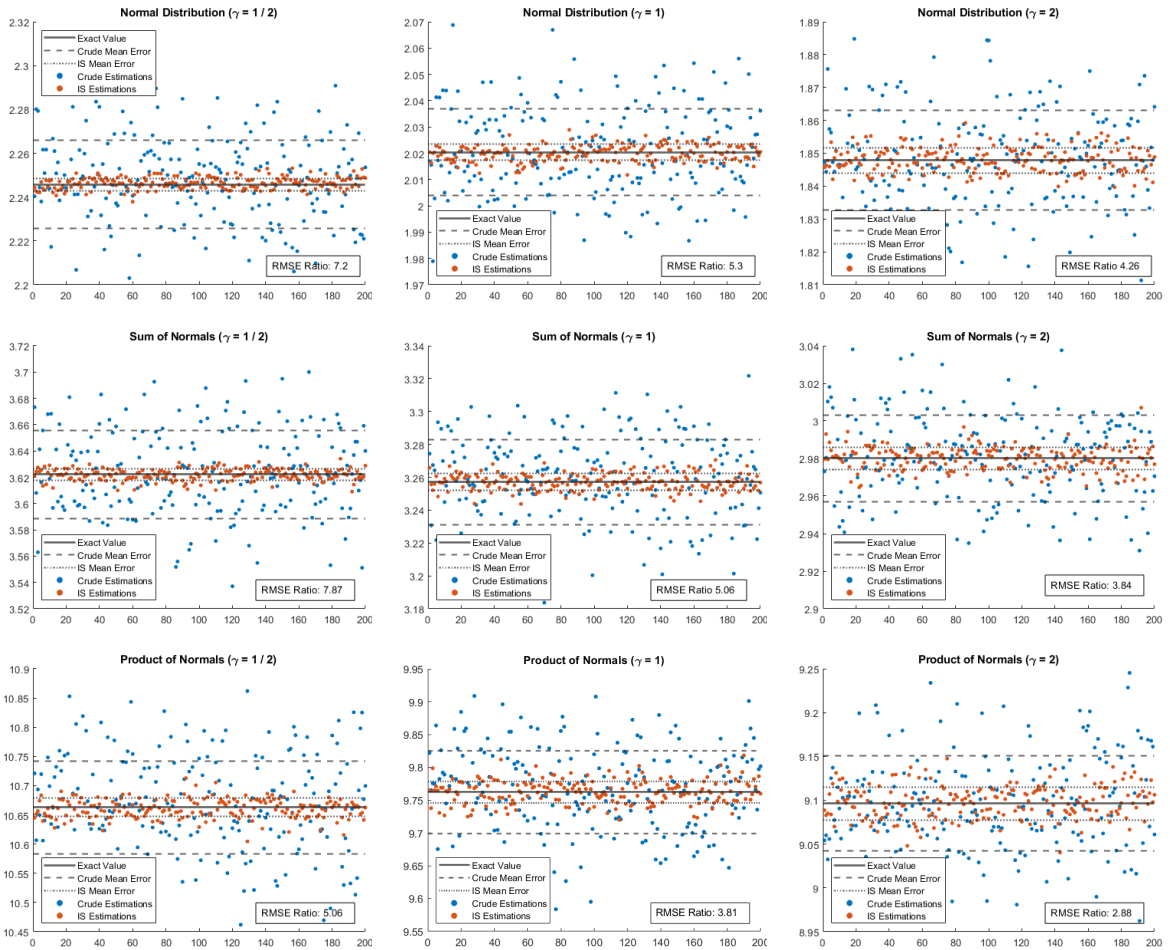


Figure 12: 200 estimations with a crude Monte Carlo estimation and the proposed importance sampling method for the models (1) to (6) and the considered DRMs $\rho_{g_{\alpha, \gamma}}$, $\gamma \in \{1/2, 1, 2\}$, $\alpha = 0.05$. Also shown is the “exact value”, which is calculated with a crude Monte Carlo estimation over 10,000,000 samples, the estimated root mean square error of the estimates around the exact value and the ratio of the root mean square error of the crude method and importance sampling method.

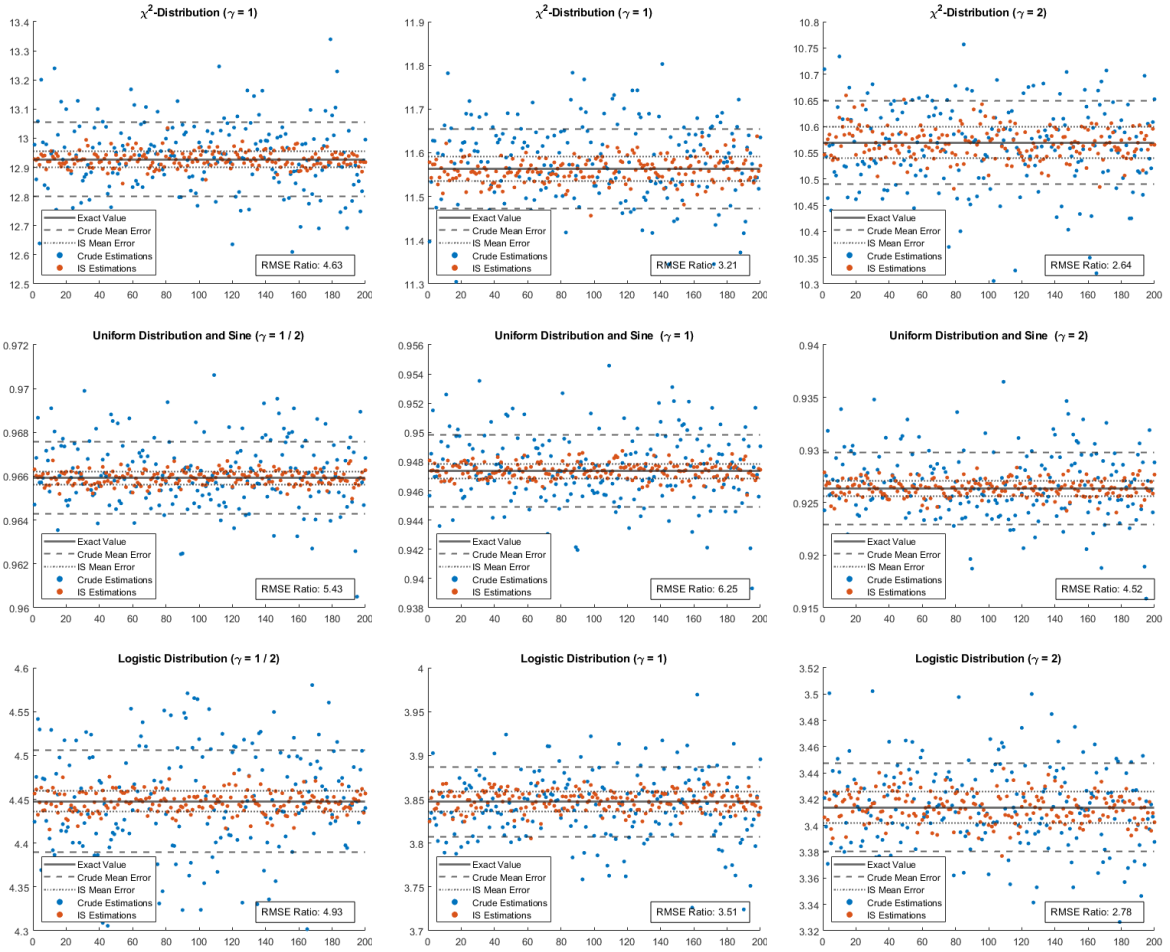


Figure 13: Continuation of Figure 12.

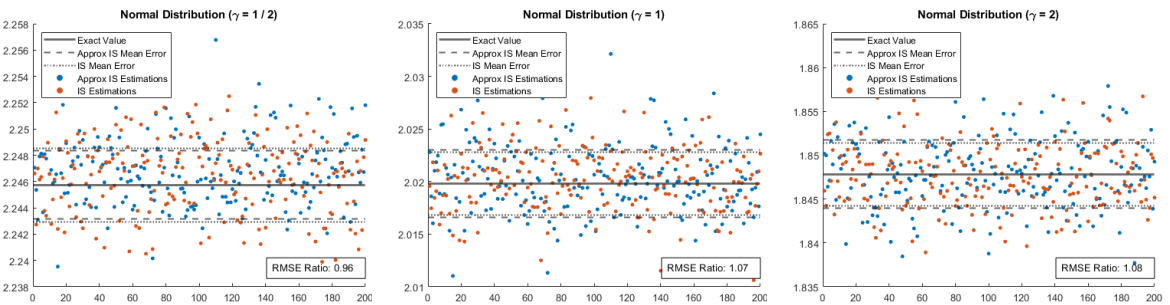


Figure 14: 200 estimations with the importance sampling method with exact knowledge of the model and the importance sampling method with an approximation of the model, chosen through k -fold validation, for the case studies (1) to (6) and the considered DRMs $\rho_{g_{\alpha}, \gamma}$, $\gamma \in \{1/2, 1, 2\}$, $\alpha = 0.05$. Also shown is the “exact value”, which is calculated with a crude Monte Carlo estimation over 10,000,000 samples, the estimated root mean square error of the estimation around the exact value and the RMSE ratio between the two importance sampling methods.

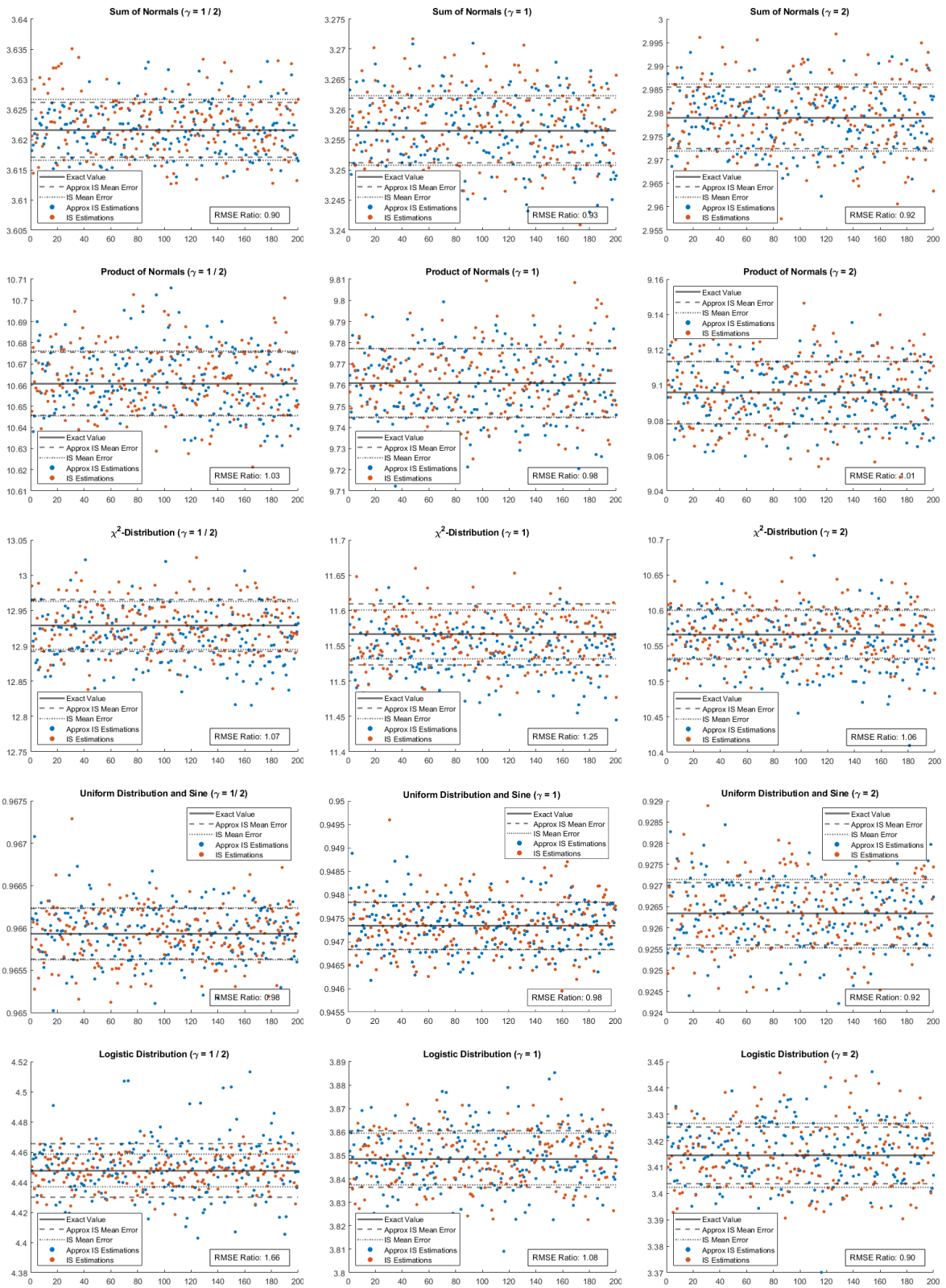


Figure 15: Continuation of Figure 14.

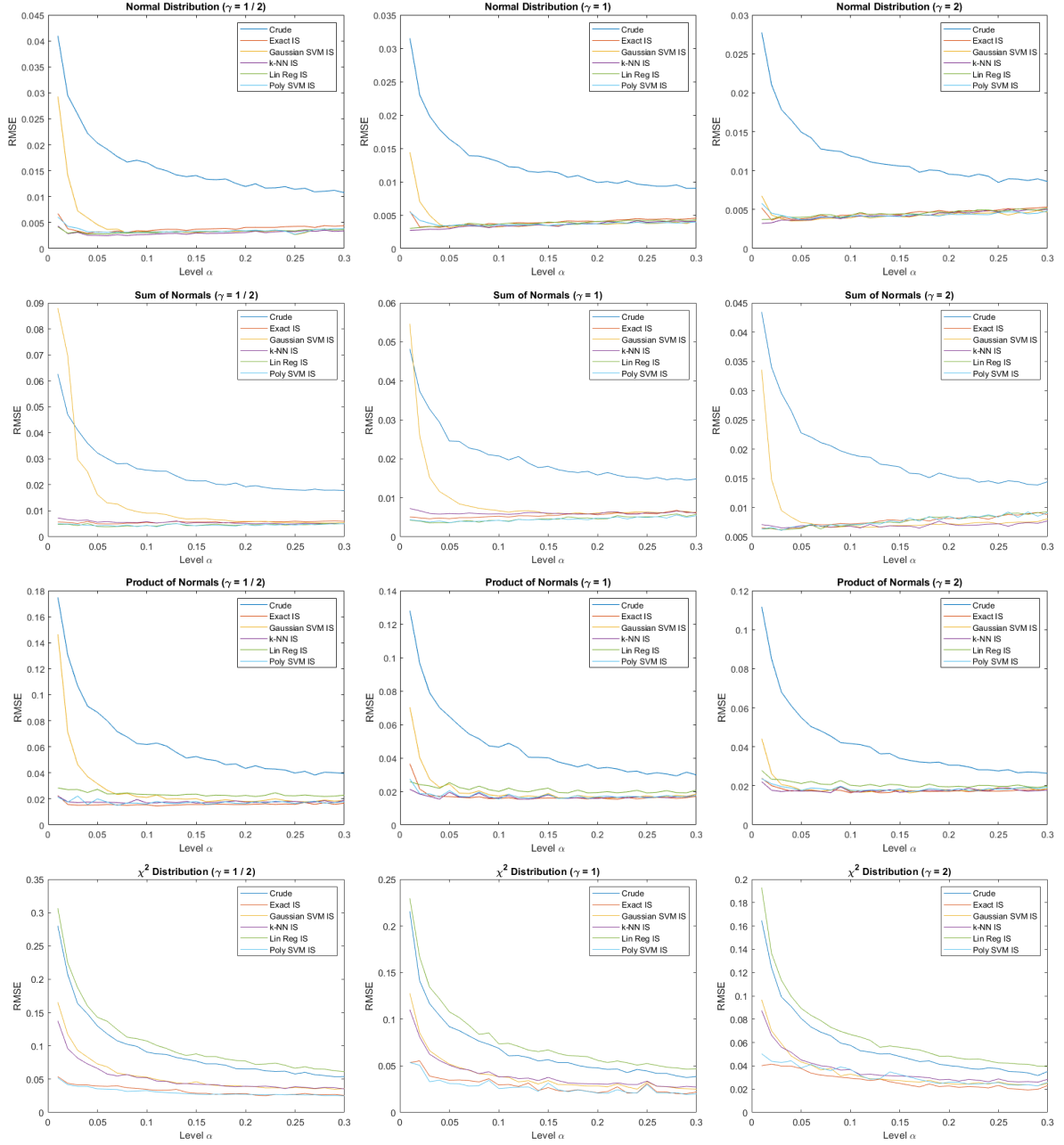


Figure 16: Root Mean Square Error (RMSE) for estimating the DRMs $\rho_{g_{\alpha,\gamma}}$, with $\gamma \in \{1/2, 1, 2\}$, $\alpha \in [0.01, 0.3]$, for the models (1) to (6). The DRMs are estimated with a crude Monte Carlo method and the proposed importance sampling method using different approximations of the black box models used in the paper.

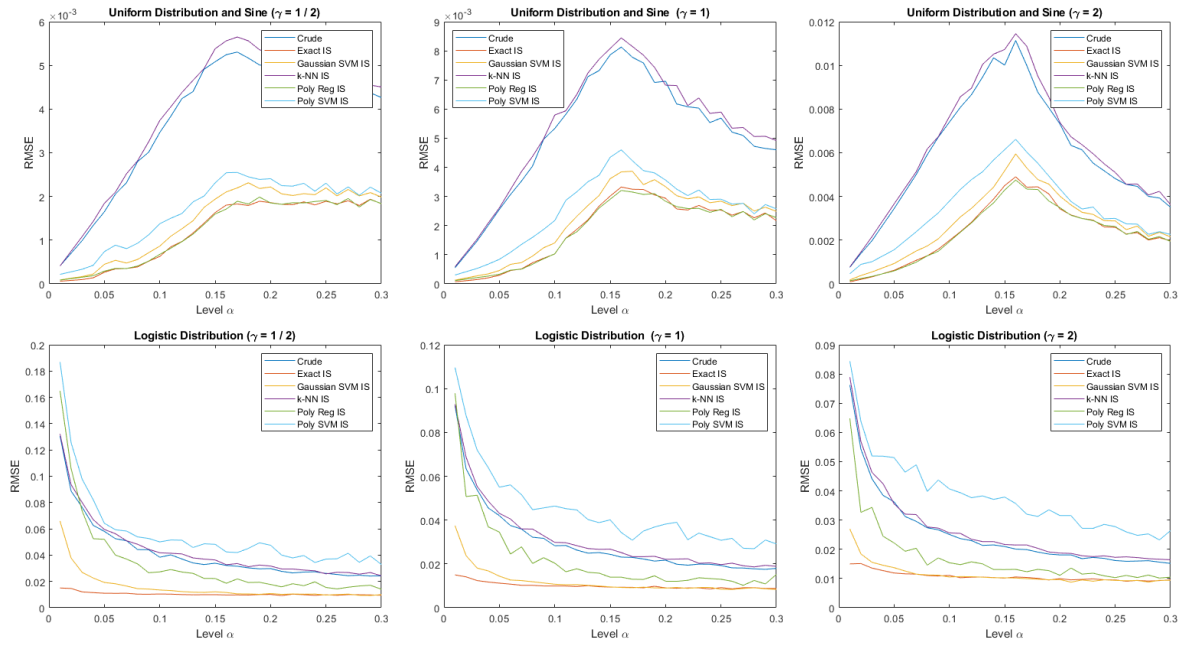


Figure 17: Continuation of Figure 16.

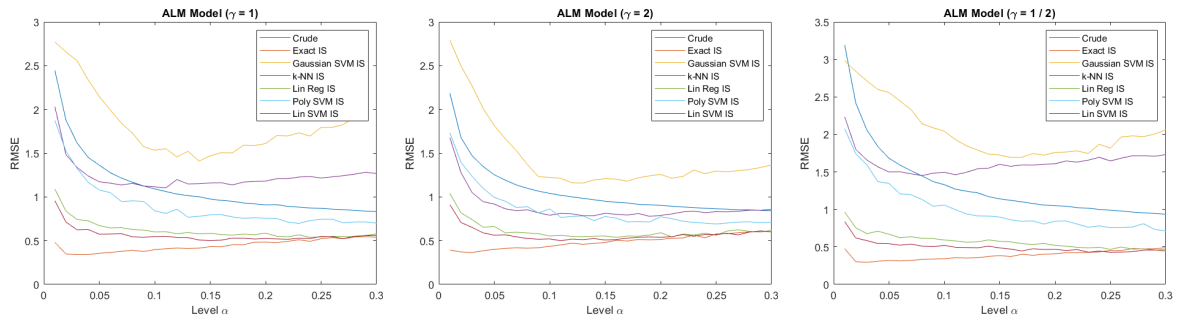


Figure 18: RMSE of the crude method and various importance sampling methods of the considered DRMs for the evolution of the net asset value in the ALM model. The importance sampling methods are implemented with the different approximation techniques considered in the paper.