*Proceedings of the 2021 Winter Simulation Conference*
*S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.*

# EFFICIENT BLACK-BOX IMPORTANCE SAMPLING FOR VaR AND CVaR ESTIMATION

Anand Deo

Singapore University of Technology and Design
8 Somapah Rd, SINGAPORE 487372

Karthyek Murthy

Singapore University of Technology and Design
8 Somapah Rd, SINGAPORE 487372

## ABSTRACT

This paper considers efficient Importance Sampling (IS) for the estimation of tail risks of a loss defined in terms of a sophisticated object such as a machine learning predictor or a mixed-integer linear optimization formulation. Assuming only black-box access to the loss and the distribution of the underlying random vector, the paper presents an efficient IS algorithm for estimating the Value at Risk and Conditional Value at Risk. The key challenge in any IS procedure, namely, identifying an appropriate change-of-measure, is automated with a self-structuring IS transformation that learns and replicates the concentration properties of the conditional excess from less rare samples. The resulting estimators are enjoy asymptotically optimal variance reduction when viewed in the logarithmic scale. Simulation experiments, both on synthetic and real datasets, highlight the efficacy and practicality of the proposed IS scheme.

## 1 INTRODUCTION

Value at Risk (VaR) and Conditional Value at Risk (CVaR) constitute two widely used measures of tail risk in quantitative risk management (McNeil, Frey, and Embrechts 2015). Desirable properties such as subaddivity, convexity, etc., have allowed CVaR to further flourish as a vehicle for introducing risk aversion in planning problems in operations and machine learning (see, for example, Rockafellar and Uryasev 2000; Bienstock et al. 2014; Ban et al. 2018; Tamar et al. 2016). The value at risk at a tail probability level $\beta$ is the $(1-\beta)$-th quantile of the loss distribution. CVaR at level $\beta$ is the expected value of the loss over its largest $\beta$ fraction of outcomes and is relatively more challenging to estimate than VaR (Lim, Shanthikumar, and Vahn 2011). With a limited fraction of data representing the loss distribution tail, estimation of VaR/CVaR via simulation is executed typically with a rare event simulation technique such as importance sampling, splitting, conditional Monte Carlo or control variates for the purposes of variance reduction and accelerated estimation (Glasserman 2013).

As we explain imminently in the context of importance sampling (IS), efficient use of these simulation techniques rely often on leveraging the structure of the loss at hand and the distribution of the underlying random vector. In terms of general methodology, Glynn 1996 demonstrates how variance reduction using IS in tail estimation can be translated to efficient estimation of VaR. Sun and Hong 2010 develop asymptotic representations for VaR/CVaR which yield conveniently applicable characterizations of asymptotic variances for VaR and CVaR. Bardou et al. 2009; Egloff and Leippold 2010 and, more recently, He et al. 2021 develop adaptive algorithms which incorporate generic IS changes of measure in estimation of VaR/CVaR. While generically applicable, it is not within the scope of these works to provide specific prescriptions of IS changes of measure that offer variance reduction guarantees. In this regard, Glasserman et al. 2000; Glasserman et al. 2002; Bassamboo et al. 2005 demonstrate how the properties of multivariate normal and $t$ distributions can be exploited to reap substantial variance reduction in portfolio risk estimation contexts. These algorithms critically utilize the specific structural properties of the loss, such as the linear-quadratic or the sum of indicators structure, and are restricted to settings involving multivariate normal and $t$ distributions.

In a number of operations and risk management contexts, the underlying loss often involves a sophisticated structure. Planning problems typically specify a loss in terms of an optimization formulation involving

**978-1-6654-3311-2/21/$31.00 ©2021 IEEE**

numerous constraints. In the rapidly growing instances of operations and risk management models which use machine learning tools, a suitable loss is written in terms of a feature-map (or) a feature-based decision rule specified, for example, in terms of representation-learning devices such as kernels or deep neural networks (see Ban and Rudin 2019, Elmachtoub and Grigas 2021 and references therein). Given the rich modelling power of these loss instances, it is impractical to explicitly tailor the IS change of measure to the problem considered. Adaptive IS methods, which utilize the estimator variance (or) cross-entropy criterion (Rubinstein and Kroese 2013) to search for the best parameter choice within a chosen IS distribution family, remain the most common approach to address this challenge. The performance of the adaptive approaches is however determined crucially by the IS family distribution family initially chosen and may additionally involve systematic underestimation (Arief et al. 2021).

This paper aims to tackle the challenges in marrying efficiency with black-box IS for VaR/CVaR estimation. Restricting to multivariate normal distributions, Bai et al. 2020; Arief et al. 2021 utilise the machinery of dominating points to algorithmically arrive at efficient IS mixture distributions for estimation of distribution tails of losses that can be either directly written or approximated with a piece-wise linear structure. Assuming only a black box access to the evaluations of loss $L(\cdot)$ and the distribution of the underlying random vector $\boldsymbol{X}$, we present here an efficient IS algorithm (Algorithm 1) to jointly estimate VaR/CVaR of $L(\boldsymbol{X})$. The IS scheme in this paper builds upon a generically applicable large deviations framework and the IS scheme developed in Deo and Murthy 2021 for the estimation of distribution tails. Exploiting the self-similarity in conditional excess distributions at different thresholds, the novel approach informs a suitable IS measure by extrapolating excess loss samples observed at less rare thresholds. We show that the proposed IS scheme offers asymptotically optimal variance reduction, when viewed at a logarithmic scale, for a broad class of useful losses and multivariate distributions. Specifically, given any $\varepsilon > 0$, we show that the sample complexity for estimating CVaR at a tail probability level $\beta$ scales as $O(\beta^{-\varepsilon})$ with the proposed IS scheme. It is instructive to contrast this with the scaling of $O(\beta^{-1-\varepsilon})$ obtained for the case of naive estimation without IS. We complement the variance reduction guarantees with numerical experiments that validate the efficacy and generic applicability of the proposed scheme.

We note that an attempt at black box CVaR estimation is made by Deo and Murthy 2020 for the case where $\boldsymbol{X}$ has regularly varying tails (that is, when $P(X_i > x) \sim x^{-\alpha_i}$, for $\alpha_i > 0$). While their scheme bears some similarity to Algorithm 1, it relies heavily on the weak convergence properties of regularly varying densities, and does not result in asymptotically optimal variance reduction.

**Notation:** We use $\xrightarrow{\mathscr{D}}$ to denote convergence in distribution. Boldface letters denote vectors. Likewise for a function $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^k$, $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))$. We let $N(\mu, \sigma^2)$ denote a normal variable with mean $\mu$ and variance $\sigma^2$. Let $\|\boldsymbol{x}\|_p$ denote the $\ell_p$ norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$ and $B_r(\boldsymbol{x})$ denote the $l_\infty$-metric ball of radius $r$ centred at $\boldsymbol{x}$. For an increasing function $f : \mathbb{R} \to \mathbb{R}$, we let $f^{-1}$ denote its left-inverse. For real valued functions $f$ and $g$, we say that $f(x) = O(g(x))$ as $x \to \infty$ if there exist positive constants $M, x_0$ such that for all $x > x_0$, $|f(x)| \leq M|g(x)|$. We say that $f(x) = \tilde{O}(g(x))$ if $f(x) = O(g(x)\log^k(x))$, for some $k > 0$.

## 2 PROBLEM DESCRIPTION

Suppose $L(\boldsymbol{x})$ denotes the loss incurred when the underlying random vector $\boldsymbol{X}$ realizes the value $\boldsymbol{x}$. Let $F_L$ denote the distribution function of $L(\boldsymbol{X})$, that is, $F_L(u) = P(L(\boldsymbol{X}) \leq u)$, and let $f_L$ be its density. Given a confidence level $\beta \in (0,1)$, denote the Value at Risk (VaR) and Conditional Value at Risk (CVaR) of the loss $L$ at level $\beta$ as,

$$v_\beta = F_L^{-1}(u) := \inf\{u \in \mathbb{R} : F_L(u) \geq 1-\beta\}, \quad \text{and} \quad C_\beta := v_\beta + \beta^{-1}\mathbb{E}\left(L(\boldsymbol{X}) - v_\beta\right)^+$$

respectively. Our objective is to enable efficient estimation of the VaR $v_\beta$ and CVaR $C_\beta$ for values of $\beta$ close to 0. Assumption 1 below imposes a mild regularity condition on the function $L(\cdot)$, whose evaluation may be available only via a black-box.

**Assumption 1.** The function $L : \mathbb{R}^d \to \mathbb{R}$ satisfies the following conditions:

(i) the set $\{\boldsymbol{x} \in \text{supp}(\boldsymbol{X}) : L(\boldsymbol{x}) > u\}$ is contained in $\mathbb{R}^d_+$ for all sufficiently large $u$; and
(ii) for any sequence $\{\boldsymbol{x}_n\}_{n\geq 1}$ of $\mathbb{R}^d_+$ satisfying $\boldsymbol{x}_n \to \boldsymbol{x}$, we have

$$\lim_{n\to\infty} \frac{L(n\boldsymbol{x}_n)}{n^\rho} = L^*(\boldsymbol{x}),$$

where $\rho$ is a positive constant and the limiting function $L^* : \mathbb{R}^d_+ \to \mathbb{R}$ is such that the cone $\{\boldsymbol{x} \in \mathbb{R}^d_+ : L^*(\boldsymbol{x}) > 0\}$ is nonempty.

Assumption 1 stipulates that the loss incurred, denoted by $L(\boldsymbol{X})$, is large when at least one of components of $\boldsymbol{X}$ takes large values. Besides commonly considered examples such as piecewise affine and linear-quadratic losses, Assumption 1 is satisfied for a wide-class of operations and quantitative risk management models that motivate our study. These include cases where $L(\cdot)$ is written as the value of a suitable mixed integer linear program or a quadratic program, and instances in prescriptive analytics where a suitable $L(\cdot)$ is written in terms of feature maps or decision-rules specified by a neural network with ReLU activation units. We refer the reader to Deo and Murthy 2021, Section 2 for a precise description of these examples for which Assumption 1 is readily satisfied. Notably, Assumption 1 does not require the loss to be convex or possess specific combinatorial structure.

**Monte-Carlo estimation without any change of measure.** Given $n$ independent samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $\boldsymbol{X}$, let $\hat{F}_{n,L}$ denote the empirical cumulative distribution function (c.d.f.) formed from the samples $L(\boldsymbol{X}_1), \ldots, L(\boldsymbol{X}_n)$. Then the VaR and CVaR at level $\beta$ can be estimated as,

$$\hat{v}_{\beta,n} = \hat{F}^{-1}_{n,L}(1-\beta) \quad \text{and} \quad \hat{C}_{\beta,n} = \hat{v}_{\beta,n} + [n\beta]^{-1} \sum_{i=1}^{n} \left[L(\boldsymbol{X}_i) - \hat{v}_{\beta,n}\right]^+, \tag{1}$$

respectively. These estimators satisfy asymptotic normality under nominal assumptions (see, for example, Serfling 2009, pg. 75 and Trindade et al. 2007, Theorem 2):

$$\sqrt{n}(v_\beta - \hat{v}_{\beta,n}) \xrightarrow{\mathscr{D}} \sigma_v(\beta)N(0,1) \quad \text{and} \quad \sqrt{n}(C_\beta - \hat{C}_{\beta,n}) \xrightarrow{\mathscr{D}} \sigma_c(\beta)N(0,1) \tag{2}$$

where

$$\sigma_v^2(\beta) = \beta(1-\beta)[f_L(v_\beta)]^{-2} \quad \text{and} \quad \sigma_c^2(\beta) = \beta^{-2}\text{Var}\left[\left(L(\boldsymbol{X}) - v_\beta\right)^+\right]. \tag{3}$$

The asymptotic variances indicate the price paid in terms of sample complexity when $\beta \searrow 0$. Observe that (2) and (3) imply that with the error in CVaR estimation with $n$ samples is roughly $N(0, n^{-1}\sigma_c^2(\beta))$. It can be seen that $\sigma_c^2(\beta) = \tilde{O}(\beta^{-1})$ (see for example, (23)). Therefore, estimating $C_\beta$ within a relative error of $\varepsilon$ with $(1-\delta)$ confidence necessarily requires $\tilde{O}(\beta^{-1}\delta^{-1}\varepsilon^{-2})$ samples of $\boldsymbol{X}$ when using the above sample-average based estimators; see also Sun and Hong 2010. Since this sample requirement is impractically large when $\beta$ is small, importance sampling is typically considered in order to reduce mean square error (MSE) to a lower order than $\tilde{O}(\beta^{-1})$.

## 3 THE PROPOSED IS ALGORITHM

We begin by describing the IS scheme presented in Algorithm 1 below. To circumvent the issue of limited relevant observations in tail exceedance events of the form $\{L(\boldsymbol{X}) > u\}$, IS typically involves obtaining samples from an alternate distribution under which these exceedance events are less rare. To accomplish this in our context, define the $\mathbb{R}^d$-valued function $\boldsymbol{T}(\boldsymbol{x}) := \boldsymbol{x}[r_\beta]^{\boldsymbol{\kappa}(\boldsymbol{x})}$, where $r_\beta : [0,1) \to \mathbb{R}_+$ is a decreasing function of $\beta$ explicitly identified in Algorithm 1 and

$$\boldsymbol{\kappa}(\boldsymbol{x}) := \frac{\log(1+|\boldsymbol{x}|)}{\rho\|\log(1+|\boldsymbol{x}|)\|_\infty}. \tag{4}$$

Exponentiation is done component-wise in the above expression for $\boldsymbol{T}(\boldsymbol{x})$ as in, $\boldsymbol{T}(\boldsymbol{x}) = (x_1 r_\beta^{\kappa_1(\boldsymbol{x})}, \ldots, x_d r_\beta^{\kappa_d(\boldsymbol{x})})$. In Algorithm 1, we use independent samples of $\boldsymbol{Z} := \boldsymbol{T}(\boldsymbol{X})$ as the samples from IS distribution specified implicitly via $\boldsymbol{T}$. The map $\boldsymbol{T} : \mathbb{R}^d \to \mathbb{R}^d$ can be shown to be invertible almost everywhere on $\mathbb{R}^d$ (see Deo and Murthy 2021, Proposition 1) and the resulting vector $\boldsymbol{Z}$ has a probability density if $\boldsymbol{X}$ has a density. Letting $f_{\boldsymbol{X}}$ and $f_{\boldsymbol{Z}}$ denote the respective densities of $\boldsymbol{X}$ and $\boldsymbol{Z}$, the likelihood ratio resulting from this change-of measure is given by,

$$\mathscr{L}_R = \frac{f_{\boldsymbol{X}}(\boldsymbol{Z})}{f_{\boldsymbol{Z}}(\boldsymbol{Z})} = \frac{f_{\boldsymbol{X}}(\boldsymbol{Z})}{f_{\boldsymbol{X}}(\boldsymbol{X})} J(\boldsymbol{X}) \tag{5}$$

An explicit expression of the Jacobian, $J(\boldsymbol{x}) = \partial \boldsymbol{T}(\boldsymbol{x}) / \partial \boldsymbol{x}$ in the above expression, is given in Algorithm 1. With this change-of-measure, we have the following unbiased estimator for the c.d.f. $F_L(u)$ :

$$\hat{F}^{\mathrm{IS}}_{n,L}(u) = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{f_{\boldsymbol{X}}(\boldsymbol{Z}_i)}{f_{\boldsymbol{X}}(\boldsymbol{X}_i)} J(\boldsymbol{X}_i)\ \mathbf{I}(L(\boldsymbol{Z}_i) > u), \tag{6}$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are drawn i.i.d. from $\boldsymbol{X}$, and $\boldsymbol{Z}_i = \boldsymbol{T}(\boldsymbol{X}_i)$. Subsequent IS estimation of $v_\beta, C_\beta$ involves a routine computation of VaR and CVaR from the given IS estimator $\hat{F}^{\mathrm{IS}}_{n,L}(u)$ for the c.d.f. and is described precisely in Algorithm 1 below.

A key feature of Algorithm 1 is that it is agnostic to the specific forms of both the loss $L(\cdot)$ and the distribution of $\boldsymbol{X}$ and it requires only a black-box access to evaluations of $L(\cdot)$ and $f_{\boldsymbol{X}}(\cdot)$. This is in sharp contrast to most existing literature requiring careful tailoring of the IS density to the underlying distribution and the loss considered. Building on the self-structuring IS procedure introduced in Deo and Murthy 2021 for estimating tail probabilities of the form $P(L(\boldsymbol{X}) > u)$, Algorithm 1 below offers a suitable adaptation to the root-finding task required to estimate VaR. In contrast to estimating $P(L(\boldsymbol{X}) > u)$ for a fixed large $u$, VaR/CVaR estimation requires that the extrapolation parameter $r_\beta$ is chosen carefully as a function of $\beta$ such that variance reduction is pronounced even if the precise range of $u$ over which root-finding has to be conducted for quantile estimation is not known apriori. The choice of hyperparameter $h$ can be made either with a cross-validation based approach we demonstrate in numerical experiments, or with recursive schemes such as those considered in Bardou, Frikha, and Pagès 2009 or He et al. 2021.

## 4 VARIANCE REDUCTION GUARANTEES FOR ALGORITHM 1

Let $\boldsymbol{\Lambda}(\boldsymbol{x}) = (\Lambda_1(x_1), \ldots, \Lambda_d(x_d))$, where $\Lambda_i(x) = -\log P(X_i \geq x)$ denotes the hazard function of component $X_i$. We say that $f : \mathbb{R} \to \mathbb{R}$ is regularly varying if for all $x \in \mathbb{R}_+$,

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^p,$$

for some $p \in \mathbb{R}$ (see de Haan and Ferreira 2007, Definition B.1.1). In this case, we write $f \in \mathcal{RV}(p)$. Letting $\boldsymbol{Y} := \boldsymbol{\Lambda}(\boldsymbol{X})$, we see that vector $\boldsymbol{Y}$ has standard exponential distribution as marginals. Just as in the use of copula models, standardization of marginals allows to state the main result without getting distracted by the differing marginal distributions.

**Assumption 2.** The marginal distribution of $\boldsymbol{X} = (X_1, \ldots, X_d)$ is such that each of $\{\Lambda_i : i = 1 \ldots, d\}$ are eventually strictly increasing and $\Lambda_i \in \mathcal{RV}(\alpha_i)$ for some $\alpha_i > 0$. The joint distribution, when written in terms of the probability density $f_{\boldsymbol{Y}}(\cdot)$ of $\boldsymbol{Y} = \boldsymbol{\Lambda}(\boldsymbol{X})$, admits the form,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = p(\boldsymbol{y}) \exp(-\varphi(\boldsymbol{y})), \tag{10}$$

where the functions $\varphi(\cdot), p(\cdot)$ satisfy the following: There exists a limiting function $I : \mathbb{R}^d_+ \to \mathbb{R}_+$ such that,

$$n^{-1} \varphi(n \boldsymbol{y}_n) \to I(\boldsymbol{y}) \quad \text{and} \quad n^{-\varepsilon} \log p(n \boldsymbol{y}_n) \to 0, \tag{11}$$

for any sequence $\{\boldsymbol{y}_n\}_{n \geq 1}$ of $\mathbb{R}^d_+$ satisfying $\boldsymbol{y}_n \to \boldsymbol{y} \neq \boldsymbol{0}$, and $\varepsilon > 0$.

**Algorithm 1:** Importance Sampling Algorithm for joint computation of VaR and CVaR

**Input:** Target tail probability level $\beta$, hyper-parameter $h > 0$, $n$ i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $f_{\boldsymbol{X}}(\cdot)$.

**1. Transform the samples:** For each sample $i = 1, \ldots, N$, compute the transformation,

$$\boldsymbol{Z}_i = \boldsymbol{T}(\boldsymbol{X}_i) := \boldsymbol{X}_i [r_\beta]^{\boldsymbol{\kappa}(\boldsymbol{X}_i)},$$

where $r_\beta = h \log\log(1/\beta)$ and $\boldsymbol{\kappa}(\boldsymbol{x})$ is given as in (4).

**2. Compute the associated likelihood:** For each transformed sample $\boldsymbol{Z}_i$, compute the respective likelihood ratio as,

$$\mathscr{L}_{R,i} := \frac{f_{\boldsymbol{X}}(\boldsymbol{Z}_i)}{f_{\boldsymbol{X}}(\boldsymbol{X}_i)} J(\boldsymbol{X}_i) \qquad i = 1, \ldots, N, \tag{7}$$

where $f_{\boldsymbol{X}}(\cdot)$ is the density of $\boldsymbol{X}$ and $J(\cdot)$ is the Jacobian of the transformation $\boldsymbol{T}(\cdot)$ given by,

$$J(\boldsymbol{x}) := \left[\prod_{i=1}^{d} \tilde{J}_i(\boldsymbol{x})\right] \times \frac{r_\beta^{\mathbf{1}^\top \boldsymbol{\kappa}(\boldsymbol{x})}}{\max_{i=1,\ldots,d} \tilde{J}_i(\boldsymbol{x})}, \tag{8}$$

$$\text{where } \tilde{J}_i(\boldsymbol{x}) := 1 + \frac{\rho^{-1} \log(r_\beta)}{\|\log(1+|\boldsymbol{x}|)\|_\infty} \frac{|x_i|}{1+|x_i|}, \quad i = 1, \ldots, d.$$

**3 Compute the IS based VaR and CVaR:**

$$\hat{C}^{\mathrm{IS}}_{\beta,n} := \hat{v}^{\mathrm{IS}}_{\beta,n} + \frac{1}{n\beta} \sum_{i=1}^{n} \left(L(\boldsymbol{Z}_i) - \hat{v}^{\mathrm{IS}}_{\beta,n}\right)^+ \mathscr{L}_{R,i}, \tag{9}$$

where IS based VaR, $\hat{v}^{\mathrm{IS}}_{\beta,n} := \inf\{u : \hat{F}^{\mathrm{IS}}_{n,L}(u) \geq 1-\beta\}$, is estimated from the c.d.f. estimate $\hat{F}^{\mathrm{IS}}_{n,L}(\cdot)$ in (6).

A wide variety of parametric and nonparametric multivariate distributions, including normal, exponential family, elliptical, log-concave distributions and Archimedian copula models satisfy Assumption 2. Marginal distributions which satisfy $\Lambda_i \in \mathcal{RV}(\alpha_i)$ include all distributions that are either Weibull-type heavy-tailed or possess lighter tails (such as exponential, normal, etc.). See Deo and Murthy 2021, Appendix B for further details and sufficient conditions directly in terms of the distribution of $\boldsymbol{X}$.

**Choice of the IS density.** A cornerstone of VaR/CVaR estimation is the accurate estimation of the loss tail distribution, $1 - F_L(u)$, for large values of $u$. In elementary examples, this is typically achieved by choosing an IS density with features suitably mirroring the conditional distribution of $\boldsymbol{X}$ over $L(\boldsymbol{x}) > u$ (see Bucklew 2013, Section 4.2). A central component in this endeavour is to utilize large deviations to identify the most likely way in which the loss $L(\boldsymbol{X})$ becomes large. For the broad family of losses and distributions specified by Assumptions 1 and 2 above, Deo and Murthy 2021 show that the sequence of random vectors $\{t^{-1}\boldsymbol{Y} : t \geq 1\}$ satisfy (i) the following large deviations principle (LDP),

$$\lim_{t\to\infty} \frac{1}{t} \log P\left(t^{-1}\boldsymbol{Y} \in A\right) = -\inf_{\boldsymbol{y}\in A} I(\boldsymbol{y}), \text{ for any Borel set } A, \tag{12}$$

and (ii) consequently, satisfy the tail asymptotic,

$$\lim_{u\to\infty} \frac{1}{\Lambda(u^{1/\rho})} \log P(L(\boldsymbol{X}) > u) = -I^*, \tag{13}$$

for some positive constant $I^*$ (see Deo and Murthy 2021, Theorems 3.3 and 4.1).

The lack of explicit dependence on parameter $u$ in the right-hand side of (13) suggests that the concentration of the target conditional distribution of $(\boldsymbol{X} \mid L(\boldsymbol{X}) > v_\beta)$ may be approximated from the conditional samples of $(\boldsymbol{X} \mid L(\boldsymbol{X}) > l_\beta)$, where $l_\beta \ll v_\beta$ and $l_\beta \to \infty$ as the tail probability level $\beta \to 0$. The requirements on $l_\beta$ ensure that the event $\{L(\boldsymbol{X}) > l_\beta\}$, though rare by itself, is significantly less rare than the target event $\{L(\boldsymbol{X}) > v_\beta\}$ and is observed more often in the samples. Letting $l_\beta = v_\beta / r_\beta$, the map $\boldsymbol{T}$ in Algorithm 1 suitably replicates these more frequent samples from the less rare region $\{L(\boldsymbol{x}) > l_\beta\}$ onto the target set $\{L(\boldsymbol{x}) > v_\beta\}$. Specifically, the distribution of $\boldsymbol{Z} = \boldsymbol{T}(\boldsymbol{X})$ can be roughly written as approximating the conditional distribution of $\boldsymbol{X}$ given $L(\boldsymbol{X}) > v_\beta$ as in,

$$\frac{\log f_{\boldsymbol{X}}(\boldsymbol{x})}{\log P(L(\boldsymbol{X}) > v_\beta)} \approx \frac{\log f_{\boldsymbol{Z}}(\boldsymbol{x})}{\log P(L(\boldsymbol{Z}) > v_\beta)}, \text{ over } \boldsymbol{x} \in \{L(\boldsymbol{x}) > v_\beta\}. \tag{14}$$

**Example 1.** To see (14) by means of an example, suppose $\boldsymbol{X}$ has a multivariate exponential distribution with density $f_{\boldsymbol{X}}(\boldsymbol{x}) = g(\boldsymbol{x}) \exp(-\|\boldsymbol{x}\|_m)$, $\boldsymbol{x} \in \mathbb{R}^d_+$, for some $g : \mathbb{R}^d_+ \to \mathbb{R}_+$ and $m \in [1, \infty)$ (see Lu and Bhattacharyya 1990, Section 4). Changing variable to $\boldsymbol{p} = v_\beta^{-1/\rho} \boldsymbol{x}$ and letting $I^* = \inf_{L^*(\boldsymbol{p}) \geq 1} \|\boldsymbol{p}\|_m$, we obtain

$$\log f_{\boldsymbol{X}}(v_\beta^{1/\rho} \boldsymbol{p}) = -v_\beta^{1/\rho} \|\boldsymbol{p}\|_m (1 + o(1)) \quad \text{and} \quad \log f_{\boldsymbol{Z}}(v_\beta^{1/\rho} \boldsymbol{p}) = -l_\beta^{1/\rho} \|\boldsymbol{p}\|_m (1 + o(1)),$$
$$\log P(L(\boldsymbol{X}) > v_\beta) = -v_\beta^{1/\rho} I^* (1 + o(1)) \quad \text{and} \quad \log P(L(\boldsymbol{Z}) > v_\beta) = -l_\beta^{1/\rho} I^* (1 + o(1)),$$

as $\beta \to 0$, and over $\boldsymbol{x} = v_\beta^{1/\rho} \boldsymbol{p}$ in the region $\{\boldsymbol{x} : L(\boldsymbol{x}) > v_\beta\}$.

Indeed the approximating feature of $\boldsymbol{T}$ demonstrated in Example 1 can be shown to hold more generally for any $\boldsymbol{X}$ satisfying Assumption 2; see Deo and Murthy 2021, Proposition 5.1 for a precise statement of this self-structuring feature of the map $\boldsymbol{T}$ and the accompanying figures. The following asymptotic variance reduction guarantees for the proposed VaR/CVaR estimation are obtained as a consequence.

**Theorem 1** Under Assumptions 1 and 2, the IS estimators for VaR and CVaR returned by Algorithm 1 are asymptotically normal and offer the following variance reduction:

$$\sqrt{n}(v_\beta - \hat{v}^{\text{IS}}_{n,\beta}) \xrightarrow{\mathscr{D}} \sigma_{is,v}(\beta) N(0,1) \quad \text{and} \quad \sqrt{n}(C_\beta - \hat{C}^{\text{IS}}_{n,\beta}) \xrightarrow{\mathscr{D}} \sigma_{is,c}(\beta) N(0,1),$$

where the limiting variances, $\sigma^2_{is,v}(\beta)$ and $\sigma^2_{is,c}(\beta)$, satisfy,

$$\frac{\sigma^2_{is,v}(\beta)}{\sigma^2_v(\beta)} = o(\beta^{1-\varepsilon}) \quad \text{and} \quad \frac{\sigma^2_{is,c}(\beta)}{\sigma^2_c(\beta)} = o(\beta^{1-\varepsilon}),$$

as $\beta \to 0$, when compared to the naive estimation variances $\sigma^2_v$ and $\sigma^2_c(\beta)$ in (3).

Considering the proposed change of measure for the example of CVaR estimation, Theorem 1 guarantees a sample complexity of $o(\beta^{-\varepsilon})$ as $\beta \searrow 0$, where $\varepsilon > 0$ can be made arbitrarily small. Thus the asymptotic variance reduction is optimal when viewed in the logarithmic scale (see Bassamboo et al. 2005). In contrast, naive estimation without any change of measure requires $\tilde{O}(\beta^{-1})$ samples. With the variance reduction guarantee holding for any choice of hyperparameter $h > 0$, an effective $h$ can be chosen via cross-validation without incurring a change of scaling in sample complexity. The numerical experiments below demonstrate this by illustrating the relative insensitivity of variance reduction to various choices of $h$.

## 5 NUMERICAL EXAMPLES

For a given loss $L(\cdot)$ and the random vector $\boldsymbol{X}$, we adopt the following procedure across all the experiments. Following Algorithm 1, we take $n$ independent samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ to arrive at the IS c.d.f. estimate $\hat{F}^{\text{IS}}_{n,L}(\cdot)$

in (6) and subsequently use it to arrive at the IS VaR estimate $\hat{v}^{\text{IS}}_{\beta,n} := \inf\{u : \hat{F}^{\text{IS}}_{n,L}(u) \geq 1-\beta\}$ and the IS CVaR estimate in (9). For every choice of $\beta$ considered, the hyper-parameter $h$ is chosen by performing cross-validation over the observed coefficient of variation. Each experiment involves computation of CVaR as above from $n$ independent samples of $\boldsymbol{X}$ and we report the relative root-mean square error = (root mean-square error of CVaR observed across 50 independent experiments)/(average of CVaR observed across 50 experiments). To enable comparison with naive estimation without IS, we also report its sample complexity for attaining the same precision offered by the IS algorithm. We observe the following across the experiments: 1) the proposed IS has a significantly smaller relative error and a lower sample complexity when compared to estimation without any change of measure, and 2) the errors obtained using IS do not increase as the problem is made increasingly difficult by considering smaller values of $\beta$. These observations align with the conclusions of Theorem 1. The specific details of the experiments are given below.

### 5.1 PERT Network:

We consider a PERT network where the project completion time $L(\cdot)$ is generally written as the value of a mixed integer linear program. We consider an example with $d = 7$ tasks and take $L(\boldsymbol{x}) = x_1 + x_7 + \max\{x_5 + \max\{x_2, x_3\}, x_6 + \max\{x_4, x_3\}\}$. Here $L(\boldsymbol{x})$ is taken to be completion time of the PERT network when the individual task completion times realise the values $\boldsymbol{x}$. To demonstrate performance for heavier than exponential delays, we assume that the marginal distribution of each delay is $F(x) = 1 - \mathrm{e}^{-x^{0.5}}$ and their joint dependence is through a Gaussian copula whose correlation matrix is given by

$$R_{i,j} = \begin{cases} 0.1 & \text{if } |i-j| = 1, \\ 1 & \text{if } i = j, \\ 0 & \text{other-wise.} \end{cases} \quad . \tag{15}$$

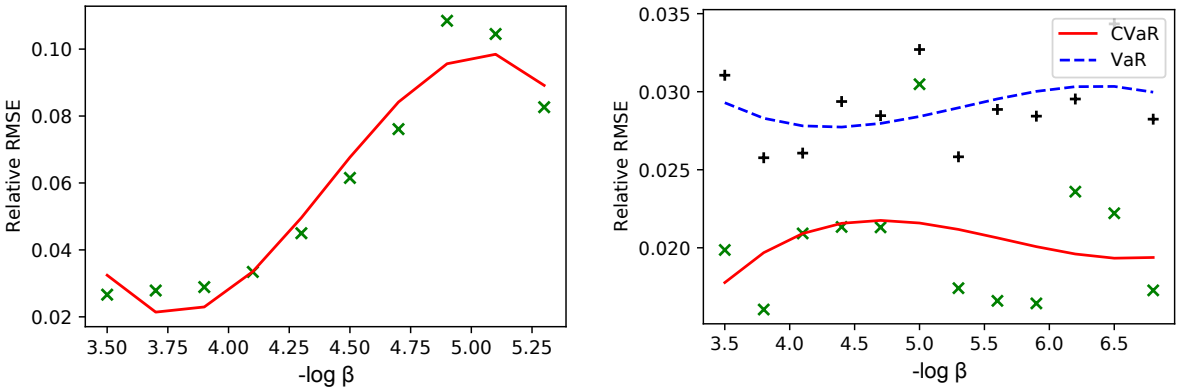
In each experiment, we take $n = 10^3$ samples to compute VaR/CVaR using the IS estimator. We plot the observed root mean square errors (observed across 50 independent experiments) in Figure 1 as a function of the tail probability level $\beta \in (10^{-7}, 10^{-3})$. The parameter $h$ is selected as $h(\beta) = 2 - 0.6\log\beta$. Figure 1(a) details the results. Contrast this to estimation without IS which requires $\approx 2 \times 10^5$ samples to attain a relative error similar to the IS scheme at $\beta = 10^{-3.5}$ (see Figure 1(b)).

### 5.2 Linear portfolios:

We consider the equally weighted linear portfolio loss $L(\boldsymbol{x}) = \mathbf{1}^{\intercal}\boldsymbol{x}$ in this example. To illustrate performance in a case where the marginal distributions of the components of $\boldsymbol{X} \in \mathbb{R}^{10}$ are different, we consider the marginal c.d.f.s $F_i(x) = P(X_i \leq x) = 1 - \mathrm{e}^{-x^{\alpha_i}}$ where $\alpha_i = 0.9$ for $1 \leq i \leq 5$, and $\alpha_i = 1.1$ for $6 \leq i \leq 10$. Dependence among the components of $\boldsymbol{X}$ is introduced through a Gaussian copula for which the correlation matrix $R$ is specified by the off-diagonal entries $[R]_{i,j} = 0.1$ for $i \neq j$, and diagonal entries $[R]_{i,i} = 1$ for $i \in \{1, \ldots, 10\}$. Figure 2(a) below presents details on cross-validation by plotting the relative error of estimation observed for different choices of hyper-parameter $h$ considered. With the relative RMSE staying less than 5% throughout the interval $h \in (1.5, 3.5)$, we note that the error reduction is robust and the estimator variance is relatively less sensitive to the choice of parameter $h$. Notice from Figure 2(b) that a relative error between $3\% - 4\%$ is obtained with only $n = 10^3$ samples upon use of the IS algorithm, and that this error is constant even as the target level $\beta$ is varied from $10^{-3.5}$ to $10^{-7}$. Note that to obtain a 3% relative error at level $\beta = 10^{-3.5}$, estimation without IS requires $n \approx 2 \times 10^5$ samples.

### 5.3 Forest fires data-set

We consider a loss trained from the forest fires dataset used in Cortez and Morais 2007 in this example. The input covariates $\boldsymbol{X}$ consist of climatic factors such as wind speed, daily rainfall, temperature, humidity etc. The output $L(\boldsymbol{X})$ is the area of forest fires, in hectares, corresponding to the respective climatic data.




(a) Relative error in CVaR estimation without IS

(b) Relative error in VaR/CVaR estimation using IS

Figure 1: Figure 1(a) displays the relative RMSE in CVaR estimation without IS. The solid red curve is fit to the estimated relative RMSE from the sample estimates indicated by green crosses. Figure 1(b) shows the relative errors in VaR (blue fit line to black marks) and CVaR (red fit line to green crosses) estimation using the IS scheme. The RMSE does not grow even as the tail probability level $\beta$ is made small.

We train a deep neural network (DNN) network to learn the function $L_{\boldsymbol{\theta}}(\cdot)$, which maps the covariates to the log of the area of the forest fire. The DNN has one hidden layer consisting of 12 neurons with ReLU link. The parameters $\boldsymbol{\theta}$ are learnt via stochastic gradient descent. We consider the following example distribution for covariates $\boldsymbol{X}$ for the sake of the experiment: The marginal distribution of the components are given by $F(x) = 1 - \mathrm{e}^{-x^{0.6}}$ and the dependence structure is informed via a Gaussian copula whose correlation matrix is given as in (15). For the purpose of this experiment, we choose $n = 517$ in (9) to match the size of the input data-set. As before, we cross validate over the parameter $h$ (see Figure 3(a)), and then using $h = 4.6$ in Algorithm 1, jointly estimate VaR / CVaR for $\beta \in (10^{-4.5}, 10^{-2})$. Figure 3(b) gives the result of our experiment. It is worthwhile to note that although the loss $L_{\boldsymbol{\theta}}(\cdot)$ is a black box, our algorithm still produces estimates of VaR/CVaR with a small relative error (4-6% for CVaR and 7-10% for VaR). Contrast this to MC estimation, which requires $n \approx 7 \times 10^3$ samples to give a relative error of 4% in CVaR estimation at $\beta = 10^{-2}$.

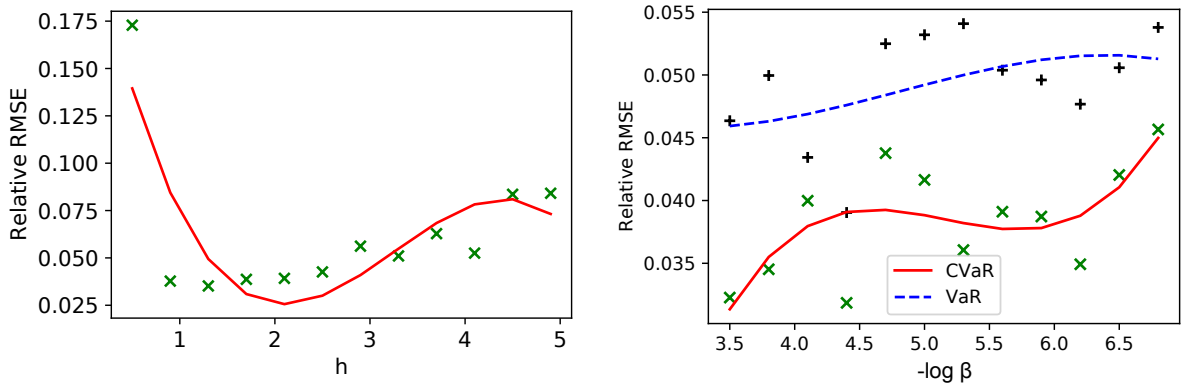
## 6 PROOF OF THEOREM 1

For ease of presentation, we focus on variance reduction in CVaR estimation and assume that $\boldsymbol{X}$ has identical marginals (that is, $\Lambda_i = \Lambda$ for all $i$). The proof for the case where $\boldsymbol{X}$ has heterogeneous marginals can be similarly accomplished by introducing a vector for capturing differing relative tail heaviness as in the results in Deo and Murthy 2021. To begin, we recall Sun and Hong 2010, Corollary 2, as applicable to our IS estimator:

$$\sqrt{n}(\hat{C}^{\mathrm{IS}}_{\beta,n} - C_\beta) \xrightarrow{\mathscr{D}} \sigma_{is,c}(\beta) N(0,1) \tag{16}$$

where $\beta^2 \sigma^2_{is,c}(\beta) = \mathrm{Var}\left[(L(\boldsymbol{Z}) - v_\beta)^+ \mathscr{L}_R \, \boldsymbol{I}(L(\boldsymbol{Z}) \geq v_\beta)\right]$ and the likelihood ratio $\mathscr{L}_R$ is defined as in (5). To present the main ideas in deconstructing the above variance term, we postpone the verification of the technical conditions required for applying Sun and Hong 2010, Corollary 2 towards the end of this section.

For any $\boldsymbol{a} \in \mathbb{R}^d_+$ and $r > 0$, let $B_r(\boldsymbol{a}) = \{\boldsymbol{y} \in \mathbb{R}^d_+ : \|\boldsymbol{a} - \boldsymbol{y}\|_\infty \leq r\}$ be a ball of radius $r$, centred at $\boldsymbol{a}$, under the $\|\cdot\|_\infty$ norm. Denote $B_r(\boldsymbol{0})$ by $B_r$. Define $\boldsymbol{q} = \boldsymbol{\Lambda}^{-1}$ be the component-wise inverse, $\boldsymbol{\psi}_\beta = \boldsymbol{\Lambda} \circ \boldsymbol{T}^{-1} \circ \boldsymbol{q}$ and $t(\beta) = \Lambda(v_\beta^{1/\rho})$. Let $\lambda_i(x) = f_{X_i}(x)/(1 - F_{X_i}(x))$ denote the hazard rate of $X_i$ and $E[X;A] = E[XI(A)]$. Define $L_\beta(\boldsymbol{p}) = [v_\beta]^{-1} L(\boldsymbol{q}(t(\beta)\boldsymbol{p}))$. Finally, let $\boldsymbol{Y}_\beta = [t(\beta)]^{-1}\boldsymbol{Y}$. For notational convenience, let $M_{2,\beta}$




(a) Cross validation over $h$

(b) Relative error in VaR/ CVaR estimation using IS

Figure 2: Figure 2(a) displays the relative RMSE in CVaR computation against the parameter $h$ (for the target level $\beta = 10^{-6}$) and Figure 2(b) shows the relative errors observed for $\beta \in [10^{-3.5}, 10^{-7}]$. The solid curves are fit to the observed relative errors in VaR (black marks) and CVaR (green crosses) estimation. In Figure 2(b), $h = 2.6$.

denote the second moment of $[\mathscr{L}_R(L(\boldsymbol{Z}) - v_\beta)^+]$. For $A \subseteq \mathbb{R}^d_+$, let $\chi_A(\cdot)$ denote its characteristic function; that is, $\chi_A(\boldsymbol{x}) = \infty$ if $\boldsymbol{x} \notin A$ and $\chi_A(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in A$. Further, for a function $f : \mathbb{R}^d_+ \to \mathbb{R}$ and $a \in \mathbb{R}$, let $\text{lev}^+_a(f) = \{\boldsymbol{x} \in \mathbb{R}^d_+ : f(\boldsymbol{x}) \geq a\}$ denote the super-level set of $f$. Let $f_{\text{LD}}(\boldsymbol{x}) = L^*(\boldsymbol{x}^{1/\alpha})$.

With this notation, see that $\boldsymbol{Y}_\beta = [t(\beta)]^{-1}\boldsymbol{Y} = [t(\beta)]^{-1}\boldsymbol{\Lambda}(\boldsymbol{X})$. Changing variables from $\boldsymbol{X}$ to $\boldsymbol{Y}_\beta$ in the expectation below (see (EC.16) onward in the proof of Lemma EC.6 of Deo and Murthy 2021 for detailed steps in a similar change of variables exercise), we obtain

$$M_{2,\beta} = \mathbb{E}\left[(L(\boldsymbol{X}) - v_\beta)^2 \frac{f_{\boldsymbol{X}}(\boldsymbol{X})}{f_{\boldsymbol{X}}(\boldsymbol{T}^{-1}(\boldsymbol{X}))} J(\boldsymbol{T}^{-1}(\boldsymbol{X})); L(\boldsymbol{X}) \geq v_\beta\right] = \mathbb{E}\left[\exp(-t(\beta)F_\beta(\boldsymbol{Y}_\beta))\right] \tag{17}$$

$$\text{where } F_\beta(\boldsymbol{p}) = a_\beta(\boldsymbol{p}) + b_\beta(\boldsymbol{p}) + c_\beta(\boldsymbol{p}) - 2d[t(\beta)]^{-1}\log t(\beta) + \chi_{\text{lev}^+_1(L_\beta)}(\boldsymbol{p})] \tag{18a}$$

$$\text{where } a_\beta(\boldsymbol{p}) = [t(\beta)]^{-1}\left[\log f_{\boldsymbol{Y}}(\boldsymbol{\psi}_u(t(\beta)\boldsymbol{p}) - \log f_{\boldsymbol{Y}}(t(\beta)\boldsymbol{p})\right], \text{ and} \tag{18b}$$

$$b_\beta(\boldsymbol{p}) = [t(\beta)]^{-1}\left[\sum_{i=1}^d \left[\log\lambda_i(\boldsymbol{T}_i^{-1}(\boldsymbol{q}(t(\beta)\boldsymbol{p}))) - \log\lambda_i(q_i(t(\beta)p_i))\right] - \log J(\boldsymbol{T}^{-1}(\boldsymbol{q}(t(\beta)\boldsymbol{p})))\right] \text{ and} \tag{18c}$$

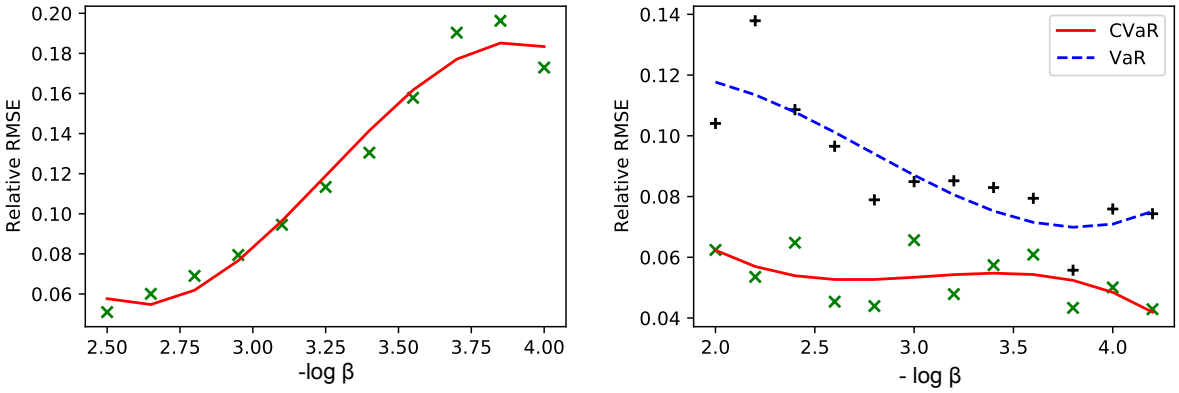
$$c_\beta(\boldsymbol{p}) = -2[t(\beta)]^{-1}\log(L(\boldsymbol{q}(t(\beta)\boldsymbol{p}) - v_\beta). \tag{18d}$$

Notice that $c_\beta(\boldsymbol{p})$ is well defined for all $\boldsymbol{p} \in \text{lev}^+_1(L_\beta)$. Observe that from (13), $v_\beta \sim q^\rho(-I^* \log\beta)$. Next, recall that $\Lambda \in \mathcal{RV}(\alpha)$ and that $\Lambda^{-1} = q$. Hence, from de Haan and Ferreira 2007, Proposition B.1.9 (viii), $q \in \mathcal{RV}(1/\alpha)$. Therefore, $r_\beta/v_\beta \to 0$ and $r_\beta \to \infty$ as $\beta \to 0$. Hence, the following conclusions of Deo and Murthy 2021, Lemmas A.8-A.11 and Corollary A.3 hold: for $\varepsilon, r > 0$, for all sufficiently small enough $\beta$,

$$\sup_{\boldsymbol{p}\in\mathbb{R}^d_+} b_\beta(\boldsymbol{p}) \geq 0 \tag{19a}$$

$$\text{lev}^+_1(L_\beta) \cap B_{\delta_1} = \emptyset \text{ for some } \delta_1 > 0, \tag{19b}$$

$$a_\beta(\boldsymbol{p}) \geq I(\boldsymbol{p}) + o(1) \text{ uniformly over } \boldsymbol{p} \in \text{lev}^+_1(L_\beta) \cap B_r, \text{ and } \liminf_{\beta\to 0} \chi_{\text{lev}^+_1(L_\beta)}(\boldsymbol{p}_\beta) \geq \chi_{\text{lev}^+_1(f_{\text{LD}})}(\boldsymbol{p}) \tag{19c}$$




(a) Relative error in CVaR estimation without IS

(b) Relative error in VaR/CVaR estimation with IS

Figure 3: Figure 3(a) displays the relative RMSE in CVaR computation without IS using $10^4$ samples. Figure 3(b) displays the relative errors in IS based VaR/CVaR computation with the parameter $h = 4.6$. In each of the figures, black and green marks respectively denote estimated VaR and CVaR values respectively.

whenever $\boldsymbol{p}_\beta \to \boldsymbol{p}$. Let $\hat{\boldsymbol{p}} = \boldsymbol{p}/\|\boldsymbol{p}\|_\infty$ be the unit vector in the direction of $\boldsymbol{p}$. Rewrite

$$L(\boldsymbol{q}(t(\beta)\boldsymbol{p})) = \frac{L\left(\frac{\boldsymbol{q}(t(\beta)\|\boldsymbol{p}\|_\infty \hat{\boldsymbol{p}})}{q(t(\beta)\|\boldsymbol{p}\|_\infty)} q(t(\beta)\|\boldsymbol{p}\|_\infty)\right)}{q^\rho(t(\beta)\|\boldsymbol{p}\|_\infty)} q^\rho(t(\beta)\|\boldsymbol{p}\|_\infty) = L^*(\hat{\boldsymbol{p}}^{1/\alpha}) q^\rho(t(\beta)\|\boldsymbol{p}\|_\infty)(1+o(1)),$$

uniformly over $\|\boldsymbol{p}\|_\infty \geq \delta$; the second equality in the above is obtained upon noting that $q \in RV(1/\alpha)$, and using the continuous convergence of $L(\cdot)$ as specified in Assumption 1. Further, as $x \to \infty$, for any $\varepsilon > 0$,

$$\frac{\log q(x)}{\varepsilon x} \to 0 \quad \text{see de Haan and Ferreira 2007, Proposition B.1.9 (1).}$$

Therefore, (19b) suggests that uniformly over $\text{lev}_1^+(L_\beta)$, $\log L(\boldsymbol{q}(t(\beta))\boldsymbol{p}) \leq \varepsilon t(\beta)\|\boldsymbol{p}\|_\infty$, for all $\beta$ sufficiently small. Further, since $t(\beta) \leq \exp(\varepsilon t(\beta))$ for all small enough $\beta$,

$$c_\beta(\boldsymbol{p}) \geq -\varepsilon\|\boldsymbol{p}\|_\infty \text{ uniformly over } \text{lev}_1^+(L_\beta). \tag{20}$$

Now for any $\varepsilon > 0$, from the bounds in (19a), (19c) and (20), one obtains that whenever $\boldsymbol{p}_\beta \to \boldsymbol{p}$ as $\beta \to 0$,

$$\liminf_{\beta\to 0} F_\beta(\boldsymbol{p}_\beta) \geq I(\boldsymbol{p}) - \varepsilon\|\boldsymbol{p}\|_\infty + \chi_{\text{lev}_1^+(f_{\text{LD}})}(\boldsymbol{p}).$$

Noting that $\boldsymbol{Y}_\beta$ satisfies an LDP with rate function $I(\cdot)$, an application of the general Varadhan's integral lemma (see Varadhan 1988, Theorem 2.2) yields,

$$\limsup_{\beta\to 0} [t(\beta)]^{-1} \log M_{2,\beta} \leq -\inf_{\boldsymbol{p}\in \text{lev}_1^+(f_{\text{LD}})} [2I(\boldsymbol{p}) - \varepsilon\|\boldsymbol{p}\|_\infty]. \tag{21}$$

Since $\boldsymbol{Y}$ has standard exponential marginals, $I(\boldsymbol{p}) \geq \|\boldsymbol{p}\|_\infty$ for all $\boldsymbol{p}$ (see Deo and Murthy 2021, Lemma 3.4 (d)). The infimum in (21) therefore occurs in a compact set. As $\varepsilon > 0$ above is arbitrary, we have

$$\limsup_{\beta\to 0} [t(\beta)]^{-1} \log M_{2,\beta} \leq -2 \inf_{\boldsymbol{p}\in \text{lev}_1^+(f_{\text{LD}})} I(\boldsymbol{p}) = -2I^*. \tag{22}$$

Next, as a consequence of (13), $(1+o(1))I^*t(\beta) = -\log P(L(\boldsymbol{X}) > v_\beta) = -\log\beta$ as $\beta \to 0$. With $t(\beta) = -(1/I^* + o(1))\log\beta$, we have $\log M_{2,\beta} \leq (2-\delta)\log\beta$. With the choice of $\delta > 0$ being arbitrary, we therefore have $M_{2,\beta} = o(\beta^{2-\delta})$ for any $\delta > 0$. Finally, for the Monte Carlo estimator without change-of-measure, $\beta^2\sigma_c^2(\beta) = \mathbb{E}\left([(L(\boldsymbol{X}) - v_\beta)^+]^2\right) - \left(\mathbb{E}\left((L(\boldsymbol{X}) - v_\beta)^+\right)\right)^2$. Notice that

$$\mathbb{E}\left([(L(\boldsymbol{X}) - v_\beta)^+]^2\right) = \int_{L(\boldsymbol{x}) \geq v_\beta} (L(\boldsymbol{x}) - v_\beta)^2 f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} \geq P(L(\boldsymbol{X}) \geq v_\beta + 1) = \beta(1 + o(1)). \tag{23}$$

Further, notice that following the analysis from (17), for any $\delta > 0$, $\mathbb{E}\left((L(\boldsymbol{X}) - v_\beta)^+\right) \leq \beta^{1-\delta}$. Hence, $\left(\mathbb{E}\left((L(\boldsymbol{X}) - v_\beta)^+\right)\right)^2 = o(\mathbb{E}\left([(L(\boldsymbol{X}) - v_\beta)^+]^2\right))$. Thus, we have that for all $\delta > 0$,

$$\frac{\sigma_{is,c}^2(\beta)}{\sigma_c^2(\beta)} = o(\beta^{1-\delta}). \quad \square$$

**Verification of the conditions of Sun and Hong 2010, Corollaries 1 and 2 :** Here we perform the pending verification of Sun and Hong 2010, Assumption 2. Notice that existence of $f_L(\cdot)$ automatically implies that Sun and Hong 2010, Assumption 1 holds, which is a sufficient condition for the central limit theorem to hold. Fix any $p > 2$. Notice that using a similar change of variables arguments as in the beginning of the proof (with $\tilde{\mathbb{E}}$ denoting expectation under the IS measure), $\widetilde{\mathbb{E}}\left(\mathscr{L}_R^p \mathbf{I}(L(\boldsymbol{X}) \geq v_\beta + \varepsilon)\right)$ is bounded above by

$$\mathbb{E}\left[\exp(-(p-1)t(\beta)F_\beta(\boldsymbol{Y}_\beta))\right], \tag{24}$$

for any $\varepsilon > 0$. Following (19a) through to (19c), $\tilde{\mathbb{E}}\left(\mathscr{L}_R^p \mathbf{I}(L(\boldsymbol{X}) \geq v_\beta + \varepsilon)\right) \leq \exp(t(\beta)p\varepsilon)$ for any $p > 2$. $\square$

## ACKNOWLEDGEMENTS

Support from Singapore Ministry of Education grant MOE2019-T2-2-163 is gratefully acknowledged.

## REFERENCES

Arief, M., Z. Huang, G. K. S. Kumar, Y. Bai, S. He, W. Ding, H. Lam, and D. Zhao. 2021. "Deep Probabilistic Accelerated Evaluation: A Robust Certifiable Rare-Event Simulation Methodology for Black-Box Safety-Critical Systems". In *International Conference on Artificial Intelligence and Statistics*, 595–603. Proceedings of Machine Learning Research.

Bai, Y., Z. Huang, H. Lam, and D. Zhao. 2020. "Rare-Event Simulation for Neural Network and Random Forest Predictors". *arXiv preprint arXiv:2010.04890*.

Ban, G.-Y., N. El Karoui, and A. E. Lim. 2018. "Machine Learning and Portfolio Optimization". *Management Science* 64(3):1136–1154.

Ban, G.-Y., and C. Rudin. 2019. "The Big Data Newsvendor: Practical Insights from Machine Learning". *Operations Research* 67(1):90–108.

Bardou, O., N. Frikha, and G. Pagès. 2009. "Computing VaR and CVaR Using Stochastic Approximation and Adaptive Unconstrained Importance Sampling". *Monte Carlo Methods and Applications* 15(3):173–210.

Bassamboo, A., S. Juneja, and A. Zeevi. 2005. "Expected Shortfall in Credit Portfolios with Extremal Dependence". In *Proceedings of the 2005 Winter Simulation Conference, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines*, 1849 – 1858. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Bienstock, D., M. Chertkov, and S. Harnett. 2014. "Chance-Constrained Optimal Power Flow: Risk-aware Network Control Under Uncertainty". *SIAM Review* 56(3):461–495.

Bucklew, J. 2013. *Introduction to Rare Event Simulation*. Springer Science & Business Media.

Cortez, P., and A. d. J. R. Morais. 2007. "A Data Mining Approach to Predict Forest Fires Using Meteorological Data".

de Haan, L., and A. Ferreira. 2007. *Extreme Value Theory: An Introduction*. Springer Science & Business Media.

Deo, A., and K. Murthy. 2020. "Optimizing Tail Risks Using an Importance Sampling Based Extrapolation for Heavy-tailed Objectives". In *2020 59th IEEE Conference on Decision and Control (CDC)*, 1070–1077.

Deo, A., and K. Murthy. 2021. "Achieving Efficiency in Black Box Simulation of Distribution Tails with Self-structuring Importance Samplers". *arXiv preprint arXiv:2102.07060*.

Egloff, D., and M. Leippold. 2010. "Quantile Estimation with Adaptive Importance Sampling". *The Annals of Statistics* 38(2):1244–1278.

Elmachtoub, A. N., and P. Grigas. 2021. "Smart "Predict, then Optimize"". *Management Science*.

Glasserman, P. 2013. *Monte Carlo Methods in Financial Engineering*, Volume 53. Springer Science & Business Media.

Glasserman, P., P. Heidelberger, and P. Shahabuddin. 2000. "Variance Reduction Techniques for Estimating Value-at-Risk". *Management Science* 46(10):1349–1364.

Glasserman, P., P. Heidelberger, and P. Shahabuddin. 2002. "Portfolio Value-at-Risk with Heavy-tailed Risk Factors". *Mathematical Finance* 12(3):239–269.

Glynn, P. W. 1996. "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*, 180–185. Publishing House of St. Petersburg University.

He, S., G. Jiang, H. Lam, and M. C. Fu. 2021. "Adaptive Importance Sampling for Efficient Stochastic Root Finding and Quantile Estimation". *arXiv preprint arXiv:2102.10631*.

Lim, A. E., J. G. Shanthikumar, and G.-Y. Vahn. 2011. "Conditional Value-at-Risk in Portfolio Optimization: Coherent but Fragile". *Operations Research Letters* 39(3):163–171.

Lu, J.-C., and G. K. Bhattacharyya. 1990. "Some New Constructions of Bivariate Weibull Models". *Annals of the Institute of Statistical Mathematics* 42(3):543–559.

McNeil, A. J., R. Frey, and P. Embrechts. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton university press.

Rockafellar, R. T., and S. Uryasev. 2000. "Optimization of Conditional Value-at-Risk". *Journal of Risk* 2:21–42.

Rubinstein, R. Y., and D. P. Kroese. 2013. *The Cross-entropy Method: a Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer Science & Business Media.

Serfling, R. J. 2009. *Approximation Theorems of Mathematical Statistics*, Volume 162. John Wiley & Sons.

Sun, L., and L. J. Hong. 2010. "Asymptotic Representations for Importance-Sampling Estimators of Value-at-Risk and Conditional Value-at-Risk". *Operations Research Letters* 38(4):246–251.

Tamar, A., Y. Chow, M. Ghavamzadeh, and S. Mannor. 2016. "Sequential Decision Making with Coherent Risk". *IEEE Transactions on Automatic Control* 62(7):3323–3338.

Trindade, A. A., S. Uryasev, A. Shapiro, and G. Zrazhevsky. 2007. "Financial Prediction with Constrained Tail Risk". *Journal of Banking & Finance* 31(11):3524–3538.

Varadhan, S. R. S. 1988. "Large deviations and Applications". In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, edited by P.-L. Hennequin, 1–49. Berlin, Heidelberg: Springer Berlin Heidelberg.

## AUTHOR BIOGRAPHIES

**ANAND DEO** is a Senior Research Assistant at Singapore University of Technology and Design. His research interests span applied probability, quantitative risk management, operations research, and machine learning. Formerly, he was a PhD student at the Tata Institute of Fundamental Research, Mumbai. His e-mail address is deo_avinash@sutd.edu.sg.

**KARTHYEK MURTHY** is an Assistant Professor in Singapore University of Technology and Design. His research centers around building models and methods for incorporating competing considerations such as risk, robustness, and fairness in data-driven optimization problems affected by uncertainty. His e-mail address is karthyek_murthy@sutd.edu.sg.