Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

COMBINING RETROSPECTIVE APPROXIMATION WITH IMPORTANCE SAMPLING FOR OPTIMISING CONDITIONAL VALUE AT RISK

Anand Deo
Karthyek Murthy
Tirtho Sarker

Singapore University of Technology and Design
8 Somapah Rd
SINGAPORE 487372

ABSTRACT

This paper investigates the use of retrospective approximation solution paradigm in solving risk-averse optimization problems effectively via importance sampling (IS). While IS serves as a prominent means for tackling the large sample requirements in estimating tail risk measures such as Conditional Value at Risk (CVaR), its use in optimization problems driven by CVaR is complicated by the need to tailor the IS change of measure differently to different optimization iterates and the circularity which arises as a consequence. The proposed algorithm overcomes these challenges by employing a univariate IS transformation offering uniform variance reduction in a retrospective approximation procedure well-suited for tuning the IS parameter choice. The resulting simulation based approximation scheme enjoys both the computational efficiency bestowed by retrospective approximation and logarithmically efficient variance reduction offered by importance sampling.

1 INTRODUCTION

Conditional value at risk (CVaR) serves as a widely used risk measure towards assessing tail risks in quantitative risk management and operations research (see McNeil, Frey, and Embrechts 2015; Rockafellar and Uryasev 2000). For a loss $\ell(\boldsymbol{X},\boldsymbol{\theta})$ associated with a decision choice $\boldsymbol{\theta}$ under a random realization $\boldsymbol{X}$, let $v_\beta(\boldsymbol{\theta})$ denote the $(1-\beta)$-th quantile of $\ell(\boldsymbol{X},\boldsymbol{\theta})$. Then its CVaR at the tail-level $\beta \in (0,1)$ is given by,

$$C_\beta(\boldsymbol{\theta}) = E\left[\ell(\boldsymbol{X},\boldsymbol{\theta}) \mid \ell(\boldsymbol{X},\boldsymbol{\theta}) \geq v_\beta(\boldsymbol{\theta})\right],$$

which measures the average loss over the worst β-fraction of the realizations. Under the assumption that $\ell(\boldsymbol{X},\cdot)$ is convex, $C_\beta(\boldsymbol{\theta})$ has been shown to possess favourable properties such as convexity, subadditivity and coherence (see Acerbi and Tasche 2002). Minimizing CVaR $C_\beta(\boldsymbol{\theta})$ over a compact set Θ enjoys the following variational representation (see Rockafellar and Uryasev 2000),

$$c_\beta = \inf_{u\in\mathbb{R},\boldsymbol{\theta}\in\Theta}\left[u+\beta^{-1}E\left(\ell(\boldsymbol{X},\boldsymbol{\theta})-u\right)^+\right] = \inf_{u\in\mathbb{R},\boldsymbol{\theta}\in\Theta} f(u,\boldsymbol{\theta}), \tag{1}$$

which is a convenient starting point for solving optimization problems. Consequently, CVaR has served as one of the primary vehicle for introducing risk aversion in a variety of planning and resource allocation problems in operations research and quantitative finance.

Since (1) is rarely solvable in closed form, Monte Carlo methods are often deployed to obtain an approximate solution. Given N i.i.d. samples of data $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N$ from the distribution of $\boldsymbol{X}$, define the sample averaged objective,

$$\hat{f}_n(u,\boldsymbol{\theta}) = \left[u+\frac{1}{N\beta}\sum_{i=1}^{N}(\ell(\boldsymbol{X}_i,\boldsymbol{\theta})-u)^+\right]$$

Then the sample average approximation (SAA) to the optimisation problem (1) may be constructed as

$$\hat{c}_n = \inf_{u,\boldsymbol{\theta}} \hat{f}_n(u,\boldsymbol{\theta}). \tag{2}$$

Desirable large sample properties such as consistency and asymptotic normality are well-known for SAA estimators (see Shapiro 1991, Theorem 3.2) and their limiting variances scale inversely with the tail level β of interest. This scaling is consistent with the understanding that one would need approximately $\tilde{O}(\beta^{-1})$ to witness loss scenarios exceeding the $(1-\beta)$-quantile $v_\beta(\boldsymbol{\theta})$, as $\beta \to 0$. The computational effort in solving the resulting optimization problems becomes large as a consequence and it becomes necessary to resort to specialized sampling schemes such as importance sampling (see, for example, He et al. 2021 and references therein) or aggregation sampling (Fairbrother et al. 2019).

In the evaluation of objective $C_\beta(\boldsymbol{\theta})$ at a fixed decision choice $\boldsymbol{\theta}$, Importance Sampling (IS) is helpful if one can identify an alternate distribution under which the rare excess loss event $\{\ell(\boldsymbol{X},\boldsymbol{\theta}) \geq v_\beta(\boldsymbol{\theta})\}$ occurs more commonly. Indeed, IS has a rich literature on such change of measure prescriptions offering efficient variance reduction in the estimation of rare event probabilities (see, for example, Heidelberger 1995; Juneja and Shahabuddin 2006, and more recently, Arief et al. 2021; Deo and Murthy 2021a; Ahn and Zheng 2021). For instance with a good change of measure suited for tail probabilities of $\ell(\boldsymbol{X},\boldsymbol{\theta})$, Glynn 1996; Glasserman et al. 2000 demonstrate how the variance reduction offered by the change of measure can be translated to efficient estimation of tail quantiles. Bardou et al. 2009; Egloff and Leippold 2010; He et al. 2021 develop adaptive algorithms which sequentially search for a good sampler choice within a IS distribution family in the estimation of CVaR.

In contrast to the prolific use of IS in tail estimation, its use in solving CVaR-driven optimization problems is limited by the availability of effective IS distribution families suited for complex objectives which arise in planning problems, and more severely, by the need to tailor the alternate sampling distribution choice differently for different decision choices: A change of measure which offers variance reduction for a decision choice $\boldsymbol{\theta} \in \Theta$ may end up being inappropriate for a different decision choice $\boldsymbol{\theta}' \in \Theta$ (see Example 1 in Section 2 and Figure 1 for an illustration). In a setting where identifying a good change of measure itself may require solving a non-trivial optimization problem (see, for eg., Bai et al. 2022), this dependence introduces a circular conceptual difficulty: A sampler choice which attains a small limiting variance depends, in turn, on the optimal decision choice which is the goal of our estimation to begin with. See He et al. 2021 for a detailed description of this circularity issue and how adaptive approaches in Lemaire and Pagès 2010; He et al. 2021 can be useful in overcoming this challenge. The cross-entropy method by Rubinstein 1997, which iteratively updates the IS distribution choice, by minimizing Kullback-Liebler divergence between a proposed IS distribution and the theoretically optimal IS distributions for increasingly rare instances, remains the most widely adopted approach.

The effectiveness of these adaptive schemes rely on working with a IS distribution family expressive enough to mirror the properties of theoretically optimal samplers at different decision choices $\boldsymbol{\theta} \in \Theta$. Their use, on the contrary, becomes practical if the chosen IS distribution family has a simple parametric representation (such as an exponential family) and is easy to sample from. Verifiably efficient prescriptions of IS distribution families have however remained elusive except in elementary instances involving piece-wise linear objectives and elliptical distributions modeling uncertainty in the problem.

In this paper, we propose a simulation based approximation scheme which embeds importance sampling naturally in the well-known retrospective approximation solution paradigm for solving optimization problems (Chen and Schmeiser 2001; Pasupathy 2010). The retrospective approximation framework optimally balances the errors due to sampling and optimization approximations (see Pasupathy 2010) and lends itself naturally to the adaptive selection of IS parameter choices.

For tackling the earlier highlighted challenges pertaining to the proposal and sampling of IS distribution family, we employ IS transformations which are structured sufficiently to induce distributions approximating the theoretically optimal zero variance measures. This deviates from the conventional approach of directly selecting a suitable IS distribution family, which has proven to be model and distribution specific and face

scalability challenges. Specifically, we employ the single parameter family of self-structuring transformations $\{\boldsymbol{T}_h(\boldsymbol{X}) : h > 0\}$ proposed in Deo and Murthy 2021a, where for every $h > 0$, the mapping $\boldsymbol{T}_h : \mathbb{R}^d \to \mathbb{R}^d$ is a deterministic bijective function. Inheriting variance reduction properties exhibited in the context of tail probability estimation in Deo and Murthy 2021a, we exhibit uniform variance reductions offered by this IS transformation family in the evaluation of the objective $f(u, \boldsymbol{\theta})$ in (1) over compact subsets of the decision set Θ. This ability to achieve versatile variance reduction for a wide range of optimization iterates renders the resulting IS estimators as a natural choice towards untying the circularity issue highlighted earlier.

Our proposal to use retrospective approximation together with IS deviates from the existing adaptive IS literature in which Robbins-Monro stochastic approximation remains the preferred solution paradigm. The predominance of stochastic approximation in adaptive IS can be understood from the observation that samples from previous iterates are less suited to be used for the gradient evaluation at the current iterate due the differing changes of measure adopted in each iteration. Since our sampler is based on transformations of the underlying vector $\boldsymbol{X}$ and enjoys robust variance reduction properties, it frees us to explore computationally attractive alternative solution paradigms. While our exposition treats the specific sampler family to be introduced shortly, we note that the enhanced retrospective approximation scheme developed in Section 4 can be used in conjunction with other distribution families as well.

Notation: We use $\xrightarrow{P}$ to denote convergence in probability and $\Rightarrow$ 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 μ and variance σ^2. Let $\|\boldsymbol{x}\|_p$ denote the ℓ_p norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$ and $B_r(\boldsymbol{x})$ denote the l_∞-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 VARIANCE REDUCTION WITH IS TRANSFORMATIONS

In this section we present an IS estimator for approximating CVaR objective. Recall our proposal to induce a suitable IS distribution via a transformation of the random vector $\boldsymbol{X}$. To accomplish this in our context, define the $\mathbb{R}^d$-valued function $\boldsymbol{T}_h(\boldsymbol{x}) := \boldsymbol{x}[s_h]^{\boldsymbol{\kappa}(\boldsymbol{x})}$, where $s_h = h\log\log(1/\beta)$, with $h > 0$, is a stretch factor which stretches the coordinates suitably via the exponent,

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

in order to generate more samples in the extreme risk regions. The stretch factor s_h, when viewed as a function of tail level β, is increasing when the problem is made rarer by letting β smaller (we henceforth drop the dependence on β in our notation). Exponentiation is done component-wise in the above expression for $\boldsymbol{T}_h(\boldsymbol{x})$ as in, $\boldsymbol{T}_h(\boldsymbol{x}) = (x_1 s_h^{\kappa_1(\boldsymbol{x})}, \ldots, x_d s_h^{\kappa_d(\boldsymbol{x})})$. The map $\boldsymbol{T}_h : \mathbb{R}^d \to \mathbb{R}^d$ can be shown to be invertible almost everywhere on $\mathbb{R}^d$ (see Deo and Murthy 2021a, 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,

$$\mathscr{L}_h = \frac{f_{\boldsymbol{X}}(\boldsymbol{Z})}{f_{\boldsymbol{Z}}(\boldsymbol{Z})} = \frac{f_{\boldsymbol{X}}(\boldsymbol{Z})}{f_{\boldsymbol{X}}(\boldsymbol{X})} J_h(\boldsymbol{X}) \tag{3}$$

An explicit expression of the Jacobian, $J_h(\boldsymbol{x}) = \partial \boldsymbol{T}_h(\boldsymbol{x})/\partial \boldsymbol{x}$ in (3), given in (15) in the Appendix, can be obtained by replacing (u/l) in Deo and Murthy 2021a, Algorithm 1 by s_h. With this change-of-measure, we have the following unbiased estimator for the objective function in (1):

$$\hat{f}_{is,n}(u, \boldsymbol{\theta}) = \left[u + \frac{1}{n\beta} \sum_{i=1}^{n} (\ell(\boldsymbol{Z}_i, \boldsymbol{\theta}) - u)^+ \mathscr{L}_{h,i} \right], \tag{4}$$

where $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are drawn i.i.d. from $\boldsymbol{X}$, $\boldsymbol{Z}_i = \boldsymbol{T}_h(\boldsymbol{X}_i)$ and $\mathscr{L}_{h,i}$ denotes the likelihood (3) evaluated at $\boldsymbol{Z}_i$. Subsequently, one may optimise over the IS weighted objective (4). The IS based estimator of minimum CVaR therefore becomes

$$\hat{c}_{is,n} = \inf_{u,\boldsymbol{\theta}} \hat{f}_{is,n}(u,\boldsymbol{\theta}) \tag{5}$$

2.1 A General Distribution Tail Model for Obtaining Variance Reduction Guarantees

We now outline the tail modelling framework under which the IS scheme described above provides substantial variance reduction. We say that $f:\mathbb{R}\to\mathbb{R}$ is regularly varying if for all $x\in\mathbb{R}_+$, $\lim_{t\to\infty} f(tx)/f(t) = x^p$, for some $p\in\mathbb{R}$ (see de Haan and Ferreira 2007, Definition B.1.1). In this case, we write $f\in\mathcal{RV}(p)$. We say that a function $f:\mathbb{R}^d_+\to\mathbb{R}_+$ is *multivariate regularly varying* if for any sequence $\boldsymbol{x}_n$ of $\mathbb{R}^d_+$ satisfying $\boldsymbol{x}_n\to\boldsymbol{x}\neq\boldsymbol{0}$,

$$\lim_{n\to\infty} n^{-1} f(\boldsymbol{h}(n)\boldsymbol{x}_n) = f^*(\boldsymbol{x}), \tag{6}$$

for some limiting $f^*:\mathbb{R}^d_+\to(0,\infty)$ and a component-wise increasing $\boldsymbol{h}(t) = (h_1(t),\ldots,h_d(t))$ satisfying $h_i\in\mathcal{RV}(1/\rho_i),\ \rho_i>0, i=1,\ldots,d$..

Assumption 1. The marginal distribution of $\boldsymbol{X} = (X_1,\ldots,X_d)$ is such that the hazard functions of $\{X_i : i = 1\ldots,d,\ \{\Lambda_i : i = 1\ldots.,d\}$ are eventually strictly increasing and $\Lambda_i\in\mathcal{RV}(\alpha_i)$ for some $\alpha_i>0$. Further the density of $\boldsymbol{X}$ when written in the form,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \exp(-\psi(\boldsymbol{x}))\ \boldsymbol{x}\in\mathbb{R}^+_d,\ \text{satisfies}\ \psi\in\mathcal{MRV}(\psi^*,\boldsymbol{h}). \tag{7}$$

A wide variety of parametric and nonparametric multivariate distributions, including normal, exponential family, elliptical, log-concave distributions and Archimedian copula models satisfy Assumption 1. 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.).

Assumption 2. *There exists a limiting function $\ell^*(\cdot;\cdot)$ and $\rho>0$, such that whenever $\boldsymbol{x}_n\to\boldsymbol{x}\neq\boldsymbol{0}$ and $\boldsymbol{\theta}_n\to\boldsymbol{\theta}$,*

$$n^{-\rho}\ell(n\boldsymbol{x}_n;\boldsymbol{\theta}_n)\to\ell^*(\boldsymbol{x};\boldsymbol{\theta}). \tag{8}$$

such that for every $\boldsymbol{\theta}$, $\ell^*(\cdot,\boldsymbol{\theta})$ is a homogeneous function of $\boldsymbol{x}$.

Assumption 2 is satisfied for example, when $\ell(\boldsymbol{x},\boldsymbol{\theta}) = c(\boldsymbol{\theta}^\intercal\boldsymbol{x})$, such that for $x\neq0$, $c(tx)/t^\rho\to c^*(x)$ as $t\to\infty$. It is also satisfied in more complicated examples, such as two-stage linear optimisation problems where

$$\ell(\boldsymbol{x},\boldsymbol{\theta}) = \boldsymbol{c}^\intercal\boldsymbol{x} + \inf_{\boldsymbol{A}\boldsymbol{y}+\boldsymbol{b}=\boldsymbol{x}} \boldsymbol{\theta}^\intercal\boldsymbol{y}.$$

We refer the reader to Deo and Murthy 2021a, Section 2 and Appendix B for a more elaborate discussion on the conditions under which Assumptions 1 and 2 are satisfied.

2.2 Uniform Variance Reduction Guarantees

One way to quantify the benefits gained by the use of IS transforms is to analyse the variance of the functions $\hat{f}_1$ and $\hat{f}_{is,1}$ as defined in (2) and (4). Note that these are simply the sample variances of the objective functions in SAA and IS-SAA respectively. Proposition 1 below quantifies these.

Proposition 1. Suppose Assumption 1 holds. Define the set $S_r = \{(u,\boldsymbol{\theta}) : u^*(1-r)\le u\le u^*(1+r),\ \theta_i^*(1-r)\le\theta_i\le\theta_i^*(1+r)\}$. Then for any $r,h_{\min},h_{\max}\in(0,\infty)$,

$$\lim_{\beta\to0}\sup_{(u,\boldsymbol{\theta})\in S_r, h\in(h_{\min},h_{\max})}\left|\frac{\log\text{var}\,[[(\ell(\boldsymbol{Z};\boldsymbol{\theta})-u)^+]\mathscr{L}_h]}{\log\text{var}([(\ell(\boldsymbol{X},\boldsymbol{\theta})-u)^+])}-2\right| = 0 \text{ and} \tag{9}$$

$$\sup_{(u,\boldsymbol{\theta})\in S_r, h\in(h_{\min},h_{\max})} \frac{\mathrm{var}(\hat{f}_{is,1}(u,\boldsymbol{\theta}))}{\mathrm{var}(\hat{f}_1(u,\boldsymbol{\theta}))} = o\left(\beta^{\frac{1}{1+\rho(r)}}\right), \tag{10}$$

where $\rho(r)$ decreases to 0 as $r \to 0$.

Proposition 1 quantifies the scale of variance reduction in the objective due to the IS transformation by comparing it with the sample average objective. The first part shows that the asymptotic variance reduction is optimal when viewed in the logarithmic scale (see Bassamboo et al. 2005). Further, since r appears multiplicatively in S_r, equations (9) implies that a uniform reduction in the variance of the objective function is obtained over a substantial neighbourhood of the optimal point. On the other hand, (10) re-expresses this improvement in performance in terms of the rarity parameter β.

Remark 1. Solving (1) accurately requires accurate approximation of the objective function f by its IS-weighted sample average $\hat{f}_{is,n}$. To this end, one seeks to find an alternate measure that resembles the conditional distribution of the random vector $\boldsymbol{X}$ in region $\{\boldsymbol{x} : \ell(\boldsymbol{x},\boldsymbol{\theta}) > u\}$. Whenever $\boldsymbol{X}$ and the function $\ell(\boldsymbol{x},\boldsymbol{\theta})$ satisfy Assumptions 1 and 2 respectively, the distribution of $\boldsymbol{T}_h(\boldsymbol{X})$ resembles that of $\boldsymbol{X}$ in the region $\{\boldsymbol{x} : \ell(\boldsymbol{x};\boldsymbol{\theta}) > u\}$, irrespective of the choice of decision parameter $\boldsymbol{\theta}$ (this can be seen through an application of Deo and Murthy 2021a, Proposition 5.1). Proposition 1 showcases that this translates to an exponential reduction in the variance error incurred while estimating the objective function f using the importance weighted sample average $\hat{f}_{is,n}$.

Example 1. To illustrate the difficulty in applying IS to the optimisation context, and the implications of Proposition 1, consider a simple two dimensional setting, where $\ell(\theta,x_1,x_2) = \theta x_1 + (1-\theta)x_2$. Notice that the region of concentration of the conditional distribution changes greatly when θ is changed from 0.2 to 0.8. Therefore, an IS distribution which provides a large variance reduction in the objective for $\theta = 0.2$ may not for $\theta = 0.8$. However, the uniformity of variance reduction in (9) demonstrates that by use of IS transformations this difficulty may be mitigated. For the same linear loss but with $\boldsymbol{X} \in \mathbb{R}^5$, Figure 1(b) shows that the worst case standard error over S_r is much smaller for IS than it is for SAA for $\beta = 0.001$. This demonstrates that significant benefits are obtained through use of IS transformations for realistic values of β (rather than the theorem truly holding only asymptotically).

Proposition 1 can be used to derive variance guarantees on the error in optimising CVaR with (4).

Theorem 1 Under Assumptions 1 and 2, we have $\sqrt{n}(\hat{c}_{is,n} - c_\beta) \Rightarrow \sigma_{is}(\beta)N(0,1)$ where

$$\sigma^2_{is}(\beta) = \mathrm{var}\left((\ell(\boldsymbol{Z};\boldsymbol{\theta}^*) - v_\beta(\boldsymbol{\theta}^*))^+ \mathscr{L}_h\right)$$

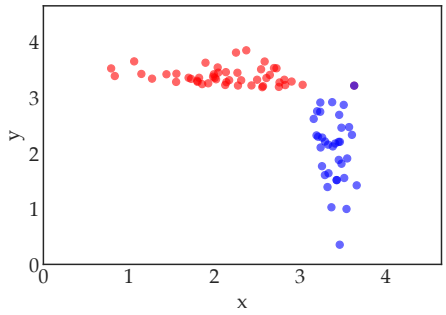
Further, the limiting variance $\sigma^2_{is}(\beta)$ of the IS estimator is smaller than the limiting variance $\sigma^2_c(\beta)$ of the SAA estimator (1) as given by the relationship,

$$\frac{\sigma^2_{is}(\beta)}{\sigma^2_c(\beta)} = o(\beta^{1-\varepsilon}), \text{ for any } \varepsilon > 0,$$

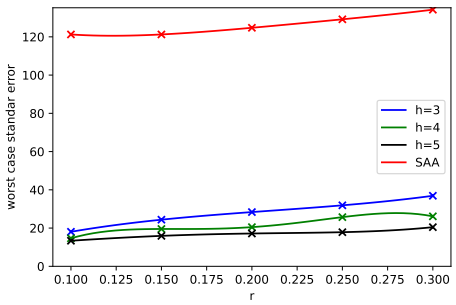
Considering the proposed change of measure for CVaR optimisation, Theorem 1 guarantees a sample complexity of $o(\beta^{-\varepsilon})$ as $\beta \searrow 0$, where $\varepsilon > 0$ can be made arbitrarily small. With the variance reduction guarantee holding for any choice of hyper-parameter $h > 0$, an effective h can be chosen via cross-validation without incurring a change of scaling in sample complexity as demonstrated in Section 4.

3 A VANILLA RETROSPECTIVE APPROXIMATION SCHEME FOR CVAR OPTIMIZATION

The procedure described in Section 2 arrives at the optimal CVaR by generating a *single* sample path $\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\}$ and then solving the resulting IS-weighted SAA problem. However, in practical implementation, solutions are not available in closed form, and sample path problems such as (4) need to be solved using deterministic optimisation algorithms. Thus, obtaining a solution to within a small tolerance level may be a computationally challenging task. With this in mind, sequential procedures are often deployed to solve



(a) Concentration of excess distribution with $\theta = 0.2, 0.8$



(b) Worst case standard errors over S_r

Figure 1: Figures 1(a) plots the conditional distribution of $\theta X_1 + (1-\theta)X_2$ conditional on it exceeding 3.5. Figure 1(b) shows the worst case relative errors (ratio of standard deviation to mean) for the functions $f_{is}(\cdot)$ over S_r and compares them with SAA.

SAA (see Bayraksan and Morton 2011; Pasupathy 2010). One such method is Retrospective Approximation (RA), which generates a *sequence* of SAA sub-problems, and progressively increases sample sizes, while solving the resulting SAA sub-problems with progressively reducing error tolerances. Further, the solution of each stage is fed into as the starting point for the optimiser in the subsequent stage. This eases the overall computational burden as follows: the initial stages obtain a rough estimate of the optimal parameters while not expending too much computation. The later stages are accurate, since the initial solution to optimisation algorithm used to solve the SAA sub-problem is close to the true solution.

In our context, given a sequence of sample sizes $\{m_k : k \geq 1\}$, this amounts to minimising a sequence of random functions $\{f_{is,m_k}(u,\boldsymbol{\theta}) : k \geq 1\}$, where $f_{is,n}(u,\boldsymbol{\theta}) \to f(u,\boldsymbol{\theta})$ in probability as $n \to \infty$. Further, at each stage k, we set the initial solution to the $(k+1)th$ sub-problem to be $(u_k, \boldsymbol{\theta}_k)$ the minimiser of f_{m_k}. Algorithm 1 gives a RA based implementation of the IS-weighted SAA algorithm from Section 2. Denote the solution of the sample path problem in (14) by $(u_k^*, \boldsymbol{\theta}_k^*)$. Notice that this is the *true solution* of the optimisation problem and not the one computed to ε_k precision. Our first result shows that use of IS significantly reduces sample errors in the solution to the sequence of sub-problems. Define

$$\boldsymbol{g}(u;\boldsymbol{\theta}) = \left[1 - \beta^{-1} E\left(\boldsymbol{I}(\ell(\boldsymbol{X},\boldsymbol{\theta}) \geq u)\right), \beta^{-1} E\left(\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{X},\boldsymbol{\theta}) \boldsymbol{I}(\ell(\boldsymbol{X},\boldsymbol{\theta}) \geq u)\right)\right] \text{ and} \tag{11}$$

$$\mathbf{G}(\boldsymbol{x};u,\boldsymbol{\theta},r) = \left[1 - \beta^{-1} \boldsymbol{I}(\ell(\boldsymbol{x},\boldsymbol{\theta}) \geq u) \mathscr{L}_r(\boldsymbol{x}) \ , \ \beta^{-1} \left(\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{x},\boldsymbol{\theta}) \boldsymbol{I}(\ell(\boldsymbol{x},\boldsymbol{\theta}) \geq u)\right) \mathscr{L}_r(\boldsymbol{x})\right], \tag{12}$$

where, $\nabla_{\boldsymbol{\theta}}$ denotes the derivative with respect to $\boldsymbol{\theta}$.

Proposition 2. Suppose Assumptions 1 and 2 are satisfied. Further, let the problem in (1) have a unique minimiser. Then, as $k \to \infty$, we have

$$\sqrt{m_k}[(u_k^*, \boldsymbol{\theta}_k^*) - (v_\beta^*(\boldsymbol{\theta}), \boldsymbol{\theta}^*)] \xrightarrow{L} N(\mathbf{0}, \Sigma_h), \text{ where } \Sigma_h = [\nabla \boldsymbol{g}(u^*;\boldsymbol{\theta}^*)]^{-\intercal} \text{var}\left(\boldsymbol{G}(\boldsymbol{Z};u^*,\boldsymbol{\theta}^*,h)\right) [\nabla \boldsymbol{g}(u^*;\boldsymbol{\theta}^*)]^{-1} \tag{13}$$

Further, the Frobenius norm of the limiting covariance satisfies $\|\Sigma_h\|_F = o(\beta^{-\varepsilon})$ for any $\varepsilon > 0$.

Proposition 2 showcases the utility of using IS in conjunction with RA to get a better performance in terms of estimation errors. In particular, the second part above suggests that the sample complexity of estimation grows as $o(\beta^{-\varepsilon})$ rather than $\tilde{O}(\beta^{-1})$. Notice that Proposition 2 gives no mention of the computation required to solve the optimisation problem in Algorithm 1, and instead focuses on the quality of the optimal solutions to the sample path sub-problems. In the next discussion, we argue the computational benefits obtained by the use of retrospective approximation continue to hold even when IS is used. Assumption 3 below imposes a mild condition on m_k, ε_k (see Pasupathy 2010, Assumptions C.1-C.3).

Algorithm 1: Retrospective Approximation based CVaR Optimisation – OPT(h)

Input: Target tail probability level β, initial solution $u_0, \boldsymbol{\theta}_0$, sample sizes $m_1, \ldots, m_k$, error tolerances $\varepsilon_1, \ldots, \varepsilon_k$.

For $k \geq 1$**, do**

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

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

2. Solve the IS based optimisation:

$$\hat{c}_{is,m_k} := \inf_{u,\boldsymbol{\theta}} \left[u + \frac{1}{m_k \beta} \sum_{i=1}^{m_k} \left(\ell(\boldsymbol{Z}_i, \boldsymbol{\theta}) - u\right)^+ \mathscr{L}_{h,i}, \right] \text{ to a tolerance of } \varepsilon_k \tag{14}$$

with an initial solution $(u_{k-1}, \boldsymbol{\theta}_{k-1})$. Return also the optimiser of (14), $(u_k, \boldsymbol{\theta}_k)$.

Assumption 3. Suppose that the sequence $\{(\varepsilon_k, m_k) : k \geq 1\}$ satisfy the following requirements:

1. If the optimisation procedure used to solve the individual sample path problems exhibits linear convergence, $\liminf_{k\to\infty} \varepsilon_{k-1}\sqrt{m} > 0$. If this procedure exhibits polynomial convergence, $\liminf_{k\to\infty} \log 1/\sqrt{m_{k-1}}(\log \varepsilon_k)^{-1} > 0$.
2. $\limsup_{k\to\infty} (\sum_{j=1}^{k} m_j)^2/\varepsilon_k^2 < \infty$
3. $\limsup_{k\to\infty} m_k^{-1} \sum_{j=1}^{k} m_j < \infty$.

Observe that there are two competing errors in the solution output by Algorithm 1. First, there is the sample error, which can be quantified by the limit theorem in Proposition 2. Second, there is the error due to solving the optimisation problem imperfectly (up to ε_k accuracy). Assumption 3 imposes conditions so that these errors are balanced out and an optimal rate of convergence of the solution is obtained. Notice that the total work done in running Algorithm 1 for k epochs equals $W_k = \sum_{j\leq k} N_j m_j$, where N_j is the (random) number of calls made to the deterministic algorithm used to solve (14) to ε_j accuracy. Proposition 3 establishes that the work normalised error for the RA based IS-weighted CVaR optimisation remains bounded.

Proposition 3. Under Assumptions 1 - 3, we have $W_k[\sigma_{is}(\beta)]^{-2} \|(u_k, \boldsymbol{\theta}_k) - (v_\beta(\boldsymbol{\theta}^*), \boldsymbol{\theta}^*)\|_2^2 = O_p(1)$.

Notice the appearance of the limiting variance $\sigma_{is}(\beta)$ from (3) above. This suggests that the IS scheme continues to enjoy a significant reduction over SAA in work normalised errors. Indeed this is showcased in experiments in Section 5, where we show that the amount of computing effort required to solve the CVaR optimisation problem to a fixed accuracy is substantially less when IS is used.

4 ENHANCED RETROSPECTIVE APPROXIMATION WITH ADAPTIVITY IN IS CHOICE

One major shortcoming of Algorithm 1 is the lack of optimisation over the cross-validation parameter h. In this discussion, we develop an enhanced version of the retrospective approximation based CVaR optimisation which includes a subroutine to iteratively arrive at a value of h which gives a lower error. To simplify matters, we assume the existence of a noisy oracle, which given n samples of data and $(u, \boldsymbol{\theta})$ returns a value $\hat{h}_n(u, \boldsymbol{\theta})$, such that $\hat{h}_n(u_n, \boldsymbol{\theta}_n) \xrightarrow{P} h(u, \boldsymbol{\theta})$ whenever $(u_n, \boldsymbol{\theta}_n) \to (u, \boldsymbol{\theta})$ as $n \to \infty$. We further assume that $h(u^*, \boldsymbol{\theta}^*)$ is a good hyper-parameter choice for problem (the specific implementation for h_n and h is presented in Section 5). Under this assumption, Algorithm 2 presents an enhancement to the RA based IS-weighted CVaR optimisation scheme. It possesses the following two benefits over prominent adaptive IS schemes:

Algorithm 2: Enhanced Retrospective Approximation based CVaR Optimisation

Input: Target tail probability level β, samples $\boldsymbol{X}_1, \ldots,$ from $f_{\boldsymbol{X}}(\cdot)$, initial seeds $u_0, \boldsymbol{\theta}_0, h_0$.
For $k \geq 1$**, do**
1. IS-Weighted CVaR optimisation: With a sample size of $m_{k,1}$ and error tolerance $\varepsilon_{k,1}$, implement steps 1 and 2 of Algorithm 1 starting from $(u_{k-1}, \boldsymbol{\theta}_{k-1})$ and with $h = h_{k-1}$.
2 Update the cross validation parameter: Using a sample size of $m_{k,2}$, return the updated value $h_k = \hat{h}_{m_{k,2}}(u_k, \boldsymbol{\theta}_k)$

For every value of $\{h(u, \boldsymbol{\theta}) : u > 0, \boldsymbol{\theta} \in \Theta\}$, the distribution for $\boldsymbol{Z}$ may be generated easily by suitably transforming the samples generated from the distribution of $\boldsymbol{X}$. The ability to cheaply obtain samples from an alternate distribution, as assumed with existing IS techniques, could be an issue when the IS distribution family need to be expressive enough to approximate the zero variance measures for complex optimization objectives.

Similarly, given a value of $(u, \boldsymbol{\theta})$, Algorithm 2 deals with the selection of $h(u, \boldsymbol{\theta})$ directly from samples of $\boldsymbol{X}$, rather than assuming a black box access to a good choice of $h(u, \boldsymbol{\theta})$. For a fixed choice of $u, \boldsymbol{\theta}$, the selection of $h(u, \boldsymbol{\theta})$ is fortunately a one-dimensional problem which can be tackled effectively with bisection or grid search and using common random numbers. Using the same samples of $\boldsymbol{X}$ to reduce the variance in IS parameter selection is not possible with most existing IS approaches. As will be explained shortly, the computed values of the hyper-parameter will converge in probability to $h(u^*, \boldsymbol{\theta}^*)$. Since the entire cross validation procedure only requires access to samples of $\boldsymbol{X}$, it is more easily implementable than adaptive techniques which require black box access to a good importance sampler at each stage.

Observe that given the $(k-1)$th stage solution, $(h_{k-1}, u_{k-1}, \boldsymbol{\theta}_{k-1})$ and $m_{k,1}$ number of samples, as a consequence of the discussion in Pasupathy 2010, Theorem 2, $(u_k, \boldsymbol{\theta}_k) = (u^*, \boldsymbol{\theta}^*) + O_p(m_{k,1}^{-1/2} + \varepsilon_k)$. Further, notice that with $m_{k,2}$ samples used for cross-validation in the first stage, one has $\hat{h}_{m_{k,2}}(u_{k-1}, \boldsymbol{\theta}_{k-1}) = h(u_k, \boldsymbol{\theta}_k) + o_p(1)$. From the continuity of $h(\cdot)$ in the parameter, and the consistency of $(u_{k-1}, \boldsymbol{\theta}_{k-1})$ as $k \to \infty$, we further have that $\hat{h}_{m_{k,2}}(u_{k-1}, \boldsymbol{\theta}_{k-1}) = h(u^*, \boldsymbol{\theta}^*) + o_p(1)$. The above discussion suggests that Algorithm 2 arrives at a statistically consistent estimate of the desired optimal hyper-parameter. It is of note to observe that the convergence to the optimal $h(u^*, \boldsymbol{\theta}^*)$ appears to be insensitive to the choice of h_0. Indeed, this is demonstrated experimentally in Section 5.

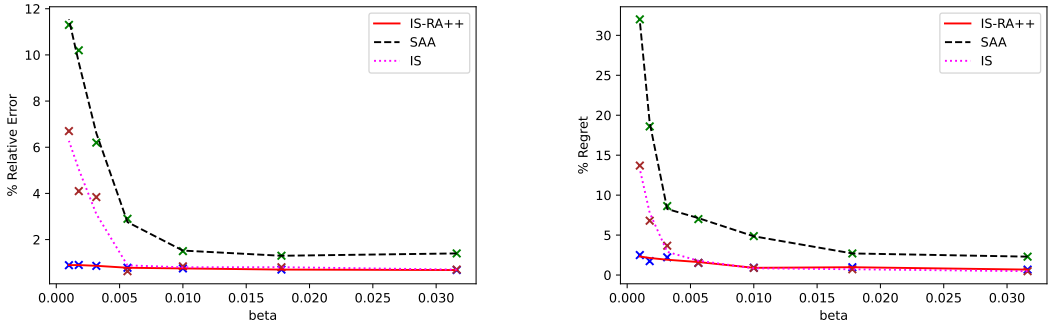
A natural question to ask given the above discussion, is whether one may derive a limit theorem which states that as $k \to \infty$, $\sqrt{m}_k(\hat{c}_{is,m_k} - c_\beta) \xrightarrow{L} N(0, [\sigma^*_{is}(\beta)]^2)$, where $[\sigma^*_{is}(\beta)]^2$ is the variance of the objective evaluated at $(u^*, \boldsymbol{\theta}^*)$ under the distribution $\boldsymbol{Z}_{h(u^*, \boldsymbol{\theta}^*)}$. Addressing this is an interesting follow-up direction. We next attempt to justify the benefits obtained by the use of Algorithm 2 through numerical experiments.

5 NUMERICAL EXPERIMENTS

5.1 Minimization of CVaR for Linear Portfolios

In order to demonstrate the efficacy of our algorithms in practice, we consider solving the constrained minimum CVaR minimisation portfolio optimisation problem. In this setting, $\ell(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^\intercal \boldsymbol{x}$, and the set $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\theta}^\intercal \mathbf{1} = 1\}$. The marginals of $\boldsymbol{X}$ are taken to have the c.d.f.s $F_i(x) = P(X_i \leq x) = 1 - \mathrm{e}^{-x^{\alpha_i}}$ where $\alpha_i = 0.5 \ \forall i$. Dependence is modelled through a Gaussian copula whose covariance matrix $\boldsymbol{R}$ is designed to capture a realistic degree of correlation among various asset returns. We compare the effectiveness of the estimators returned by Algorithms 1 and 2 with plain SAA using three performance measures:

First, we use relative mean square error as a metric to compare the quality of the optimal value output by each of the algorithms. That is, given the output of an algorithm, call it $\hat{c}_\beta$, we compute the relative root mean square error as $\sqrt{\frac{1}{K}\sum_{i=1}^{K}(\hat{c}_{\beta,i}/c_\beta - 1)^2}$ where $\hat{c}_{\beta,i}$ are outputs of independent runs of the algorithm and c_β is defined in (1). We use relative errors as metrics of performance both here as well in the next



(a) Relative RMSE in CVaR optimisation

(b) Relative regret in CVaR optimisation

Figure 2: Figure 2(a) compares the relative RMSE in CVaR optimisation using Algorithms 1 and 2 with plain SAA. Figure 2(b) compares the corresponding out-of-sample regret. In each of the figures, cross-marks respectively denote estimated quantities.

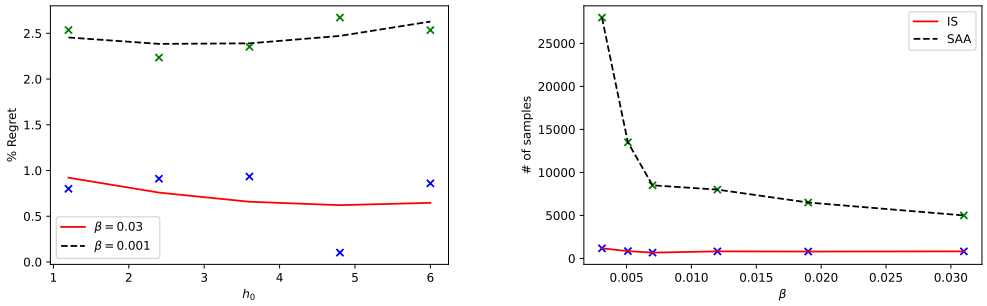
experiment to facilitate a scale free (i.e: β independent) comparison between different algorithms. In order to compute the optimal value c_β, we solve the SAA problem (2) with $N = 10^6$ samples. In each case, the value of β is varied from $\beta = 0.037$ to $\beta = 5 \times 10^{-4}$. In order to ensure a fair comparison between each of the three methods, we keep the total sample budget fixed and equal to 2500. For the implementation in Algorithm 2, this budget is divided among two epochs as $m_1 = 500$ and $m_2 = 2000$. Further, for every $(u, \boldsymbol{\theta})$, we let $\hat{h}_n(u, \boldsymbol{\theta}) = \inf_h \frac{1}{n} \sum_{i=1}^n \mathbf{I}(\ell(\boldsymbol{Z}_{h,i}; \boldsymbol{\theta}) \geq u) \mathscr{L}_{h,i}^2$. Observe that from the law of large numbers this converges to the second moment of the gradient of the objective function with respect to u. Since the final errors in decision (see Proposition 2 for example) depend on this quantity having a smaller variance, $\hat{h}_n(u, \boldsymbol{\theta})$ as chosen above is reasonable. For Algorithm 1 for all β, we choose $h = 2.5$. Figure 2(a) details the results.

We next use relative regret as a metric to compare out-of-sample performance each of the algorithms. Given the output of an algorithm $\hat{\boldsymbol{\theta}}$, we define the relative regret as $\hat{r}_\beta(\hat{\boldsymbol{\theta}}) = \frac{1}{K} \sum_{i=1}^K (c_\beta(\hat{\boldsymbol{\theta}}_i)/c_\beta - 1) \times 100\%$, where $\hat{\boldsymbol{\theta}}_i$ are outputs of independent runs of the algorithm. The first observation is that IS-RA significantly outperforms SAA in terms of relative errors obtained. Secondly, adaptively optimising over h as in Algorithm 2 leads to a further improvement over Algorithm 1. Finally, Figure 3(a) displays the relative regret incurred as a function of the starting point of the h_0. Notice that there appears to be a degree of robustness in the performance of the algorithm to the specific value of h_0 selected.

Finally, in order to compare in terms of the effort required to obtain a desired out of sample accuracy, we compute the number of samples required by each method to obtain 1% relative regret; refer Figure3(b). Observe that for $\beta = 0.037$, for IS, this is ≈ 600, while SAA requires ≈ 5500 samples. This difference is even more pronounced when $\beta = 0.003$, where SAA requires roughly 28000 samples, while IS only requires 1175.

5.2 A Portfolio Credit Risk Example

Consider the problem of selecting a portfolio of loans from among K classes (industries). Given a that a set of common market factors $\boldsymbol{X}$ realise a value $\boldsymbol{x}$, a loan belonging to class i defaults with probability $p_i(\boldsymbol{x})$ independently of everything else. We further suppose that the loss given default for loan k is given by e_k (random), independent of the market variable. Denoting n_i as the number of loans belonging to class i, the total loss given a portfolio $\boldsymbol{\theta}$ is $\ell(\boldsymbol{\theta}, \boldsymbol{X}) = \sum_{i=1}^K \theta_i (\sum_{k=1}^{n_k} e_k \mathbf{I}(\text{loan k defaults}))$. The objective is to find the portfolio with the minimum CVaR, subject to the total returns from the investment exceeding a nominal threshold, that is $\boldsymbol{\theta}^\intercal \boldsymbol{r} \geq q$ where the portfolio returns are given by $\boldsymbol{r}$. In this example, $K = 2$ and $\boldsymbol{X} \in \mathbb{R}^4$



(a) Relative regret as a function of h_0 (b) Samples required to obtain 1% relative regret

Figure 3: Figure 3(a) plots the relative regret incurred in optimisation as a function of the initial value h_0. Figure 3(b) plots the number of samples required by each method to obtain 1% relative regret. In each of the figures, cross-marks respectively denote estimated quantities.

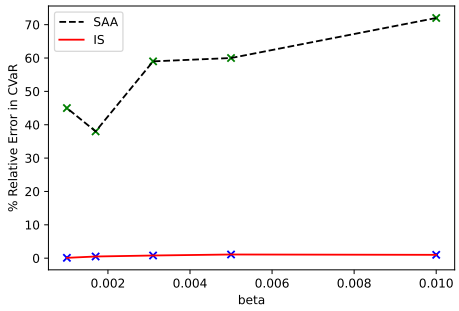


Figure 4: Error in CVaR Computation for the Portfolio Credit Risk problem

has Weibull marginals with a Gaussian copula used to model joint dependence. We further assume that $p_i(\boldsymbol{x})$ has a logistic form. Importance sampling is deployed only on the common market factors $\boldsymbol{X}$, and the IS weighted algorithm is solved as before. In this implementation, $N = 2000$ while $n_1 = n_2 = 5000$ and e_i are uniform random variables. Figure 4 displays the results of the experiment. Notice that the use of IS on the common market variable significantly improves lowers errors in optimal VaR estimation. The relative error in CVaR estimation is roughly 1% when IS is used as opposed to over 50% without the use of IS.

6 OUTLINES OF KEY PROOF STEPS

Proof of Proposition 1: For notational convenience, let $M_{2,u}$ denote the second moment of $[\mathscr{L}_h(\ell(\boldsymbol{Z},\boldsymbol{\theta}) - u)^+]$. Note that this is an upper bound for the variance which we want to compute. Further, define $t(u) = \Lambda_{\min}(u^{1/\rho})$, $q_\infty(u)$ as the inverse of $\Lambda_{\min}(u)$ and $\boldsymbol{Y}_u = [t(u)]^{-1}\boldsymbol{\Lambda}(\boldsymbol{X})$. Changing variables from $\boldsymbol{X}$ to $\boldsymbol{Y}_u$ in the expectation below (see (EC.16) onward in the proof of Lemma EC.6 of Deo and Murthy 2021a for detailed steps in a similar change of variables exercise), we obtain $M_{2,u} = \mathbb{E}\left[(\ell(\boldsymbol{X},\boldsymbol{\theta}) - u)^2 \frac{f_{\boldsymbol{X}}(\boldsymbol{X})}{f_{\boldsymbol{X}}(\boldsymbol{T}_h^{-1}(\boldsymbol{X}))} J(\boldsymbol{T}_h^{-1}(\boldsymbol{X})); \ell(\boldsymbol{X},\boldsymbol{\theta}) \geq u\right]$. This simplifies to $\mathbb{E}\left[\exp(-t(u)F_u(\boldsymbol{Y}_u))\right]$ where $F_u(\boldsymbol{p}) = a_u(\boldsymbol{p}) + b_u(\boldsymbol{p}) + c_u(\boldsymbol{p}) - 2d[t(u)]^{-1}\log t(u) + \chi_{\text{lev}_1^+(\ell_{u,\boldsymbol{\theta}})}(\boldsymbol{p})$; $a_u(\boldsymbol{p}) = [t(u)]^{-1}\left[\log[f_{\boldsymbol{Y}}(\boldsymbol{\psi}_u(t(u)\boldsymbol{p})/f_{\boldsymbol{Y}}(t(u)\boldsymbol{p})]\right]$, $b_u(\boldsymbol{p}) = [t(u)]^{-1}\left[\sum_{i=1}^d \left[\log\lambda_i(\boldsymbol{T}_{h,i}^{-1}(\boldsymbol{q}(t(u)\boldsymbol{p}))) - \log\lambda_i(q_i(t(u)p_i))\right] - \log J_h(\boldsymbol{T}^{-1}(\boldsymbol{q}(t(u)\boldsymbol{p})))\right]$ and $c_u(\boldsymbol{p}) = -2[t(u)]^{-1}\log(\ell(\boldsymbol{q}(t(u)\boldsymbol{p},\boldsymbol{\theta}) - u))$. Define $p_u = P(\ell(\boldsymbol{X},\boldsymbol{\theta}) \geq u)$. Then an application of the general Varad-

han's integral lemma (Varadhan 1988, Theorem 2.1), for all $(u,\boldsymbol{\theta})$, we have that $M_{2,u} = o(p_u^{2-o(1)})$ (see the proof of Deo and Murthy 2021b, Theorem 1). Similarly, we have that $\text{var}[(\ell(\boldsymbol{X},\boldsymbol{\theta})-u)\mathbf{I}(\ell(\boldsymbol{X},\boldsymbol{\theta})\geq u)] = \Theta(p_u)$. Combining these two proves the claim for a fixed $(u,\boldsymbol{\theta})$. Notice that $\inf\{u : u \in S_r\} \geq \log^{\varepsilon} 1/\beta$ for all β small enough. Further, $\inf_{h>h_{\min}} s_h = o(-\log^{\varepsilon}\beta)$. The uniformity over $h,u,\boldsymbol{\theta}$ is now obtained using the uniform convergence of ℓ as in Assumption 2 and then following the proof of Deo and Murthy 2021a, Lemma D.2. For part 2, notice that from the above discussion, for any $(u,\boldsymbol{\theta})$, the ratio of the variances of $\hat{f}_{is,1}$ and $\hat{f}_1$ is given by $o(p_u^{1-\varepsilon})$. Under Assumptions 1 and 2, it can be established that for some continuous function $a(\cdot)$, $C_\beta(\boldsymbol{\theta}) = a(\boldsymbol{\theta})(1+o(1))q_\infty^\rho(-\log\beta)$. As a consequence of Rockafellar and Wets 2009, Theorem 7.33, $\boldsymbol{\theta}^*$ which minimises $C_\beta(\boldsymbol{\theta})$ satisfies $\boldsymbol{\theta}^* = \boldsymbol{\theta}_1 + o(1)$, where $\boldsymbol{\theta}_1 \in \arg\min a(\boldsymbol{\theta})$. Then, the optimal $u^* = v_\beta(\boldsymbol{\theta}^*) = I^* q_\infty^\rho(-\log\beta)[1+o(1)]$, where $I^* = \inf_{\boldsymbol{\theta}} a(\boldsymbol{\theta})$. This implies that $u \in S_r$, $u \geq (1-r)I^* q_\infty^\rho(-\log\beta)$. Finally, it can be seen that $\log p_u = \Lambda_{\min}(u/a(\boldsymbol{\theta}))(1+o(1))$ (apply Deo and Murthy 2021a, Theorem 4.2). Plugging everything together, we have that $\log p_u \leq (1-r)^{\alpha_{\min}}(I^*/a(\boldsymbol{\theta}))\log\beta$. Now, the continuity of $a(\cdot)$ implies that whenever $(u,\boldsymbol{\theta}) \in S_r$, and $h > h_{\min}$, $p_u \leq \beta^{1/1-\rho(r)}$, for $\rho(r) \to 0$. □

Proof of Theorem 1: Notice first that the limiting function in the optimisation, f may be written as the expected value of $\hat{f}_{is,1}$. Further, observe that due to Assumptions 1 and 2 , $\hat{f}_{is,1}$ has a finite variance. Therefore, by an application (Shapiro 1991, Theorem 3.2), one obtains the first part of Theorem 1. The quantification of the variance reduction is obtained following the proof of Proposition 1, with $p_{v_{\beta(\boldsymbol{\theta}^*)}} = \beta$. □

Proof of Proposition 2 It is sufficient to verify the conditions from Pasupathy 2010, Theorem 4. First, note that the optimal solutions $(u_k^*, \boldsymbol{\theta}_k^*)$ are zeros $\nabla\hat{f}_{is,m_k}(u,\boldsymbol{\theta})$. Further, we have that $E\boldsymbol{G}(u,\boldsymbol{\theta}) = g(u,\boldsymbol{\theta})$ for all $(u,\boldsymbol{\theta})$. Finally, owing to Assumptions 1 and 2, the usual conditions required for a CLT hold, and we therefore have $\sqrt{m_k}(\nabla\hat{f}_{is,m_k}(u,\boldsymbol{\theta}) - \boldsymbol{g}(u,\boldsymbol{\theta})) \to N(0, \text{var}(\boldsymbol{G}(\boldsymbol{Z};u,\boldsymbol{\theta};h)))$. By the assumptions of the theorem, there exists a unique minimiser, and the gradients of $\boldsymbol{g}$ is non-singular. Finally, the uniform convergence assumption holds as a result of the discussion following the statement of Pasupathy 2010, Theorem 4. □

Proof of Proposition 3 This follows from Assumption 3, Proposition 2 and the proof of Pasupathy 2010, Theorem 5. □

A Expression for the Jacobian J_h

Recall that $J_h(\cdot)$ is the Jacobian of the transformation $\boldsymbol{T}_h(\cdot)$. This is given by

$$J_h(\boldsymbol{x}) := \left[\prod_{i=1}^{d} \tilde{J}_i(\boldsymbol{x})\right] \times \frac{s_h^{\mathbf{1}^\intercal \boldsymbol{\kappa}(\boldsymbol{x})}}{\max_{i=1,\ldots,d} \tilde{J}_i(\boldsymbol{x})} \text{ where } \tilde{J}_i(\boldsymbol{x}) := 1 + \frac{\rho^{-1}\log(s_h)}{\|\log(1+|\boldsymbol{x}|)\|_\infty} \frac{|x_i|}{1+|x_i|}, \quad i = 1,\ldots,d.$$

ACKNOWLEDGEMENTS

The authors acknowledge support from Singapore Ministry of Education grant MOE2019-T2-2-163.

REFERENCES

Acerbi, C., and D. Tasche. 2002. "On the coherence of expected shortfall". *Journal of Banking & Finance* 26(7):1487–1503.

Ahn, D., and L. Zheng. 2021. "Efficient simulation for linear programming under uncertainty". In *2021 Winter Simulation Conference (WSC)*, 1–12. IEEE.

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. 2022, feb. "Rare-Event Simulation for Neural Network and Random Forest Predictors". *ACM Trans. Model. Comput. Simul.*. Just Accepted.

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.

Bayraksan, G., and D. P. Morton. 2011. "A sequential sampling procedure for stochastic programming". *Operations Research* 59(4):898–913.

Chen, H., and B. W. Schmeiser. 2001, mar. "Stochastic Root Finding Via Retrospective Approximation". *IIE Transactions* 33(3):259–275.

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

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

Deo, A., and K. Murthy. 2021b. "Efficient Black-Box Importance Sampling for VaR and CVaR Estimation". *arXiv preprint arXiv:2106.10236*.

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

Fairbrother, J., A. Turner, and S. W. Wallace. 2019, nov. "Problem-driven Scenario Generation: An Analytical Approach for Stochastic Programs With Tail Risk Measure". *Mathematical Programming* 191(1):141–182.

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

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*.

Heidelberger, P. 1995, January. "Fast Simulation of Rare Events in Queueing and Reliability Models". *ACM Trans. Model. Comput. Simul.* 5(1):43–85.

Juneja, S., and P. Shahabuddin. 2006. "Chapter 11 Rare-Event Simulation Techniques: An Introduction and Recent Advances". In *Simulation*, edited by S. G. Henderson and B. L. Nelson, Volume 13 of *Handbooks in Operations Research and Management Science*, 291 – 350. Elsevier.

Lemaire, V., and G. Pagès. 2010. "Unconstrained Recursive Importance Sampling". *The Annals of Applied Probability* 20(3):1029–1067.

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

Pasupathy, R. 2010. "On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization". *Operations Research* 58(4-part-1):889–901.

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

Rockafellar, R. T., and R. J.-B. Wets. 2009. *Variational analysis*, Volume 317. Springer Science & Business Media.

Rubinstein, R. Y. 1997. "Optimization of computer simulation models with rare events". *European Journal of Operational Research* 99(1):89–112.

Shapiro, A. 1991. "Asymptotic Analysis of Stochastic Programs". *Annals of Operations Research*.

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 Postdoctoral Researcher at Singapore University of Technology and Design. His research interests are in quantitative risk management, operations research and machine learning. 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.

TIRTHO SARKER is a Research Assistant at Singapore University of Technology and Design. His e-mail address is tirtho_sarker@sutd.edu.sg.