Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

IMPORTANCE SAMPLING FOR MINIMIZATION OF TAIL RISKS: A TUTORIAL

Anand Deo[1], and Karthyek Murthy[2]

[1]Indian Institute of Management Bangalore, Bilekahalli, Bangalore, 560076, INDIA
[2]Singapore University of Technology and Design, 8 Somapah Road, 487372, SINGAPORE

ABSTRACT

This tutorial provides an introductory overview of how one may use importance sampling to drastically reduce the sample requirements in solving stochastic optimization and elementary simulation optimization problems incorporating tail risk measures. Sample average approximations, while appealing due to their universality in use, require a large number of samples due to the rarity with which relevant tail events get observed. Importance Sampling is among the most potent methods for reducing the sample requirements in estimating rare event probabilities. Can importance sampling be used with similar effectiveness for solving optimization formulations (involving rare events) as well, and if so, what are the key ingredients required to operationalize this idea? Focusing on these questions, this tutorial aims to demonstrate (i) how to arrive at an effective change of measure prescription at every decision, and (ii) the prominent techniques available for integrating such a prescription within a solution paradigm for optimization.

1 INTRODUCTION

When building optimization models for planning and decision-making under uncertainty, a risk-neutral approach would seek to identify the best decision on average. Relying on independent yet similar repeated chances, a risk-neutral approach is justified by the law of large numbers. An unfortunate run entailing unacceptable losses in the first few runs, for example, may however make it harder to sustain and continue the operations even if an eventual upturn is inevitable. Therefore, the practice of risk management strives to look beyond expected outcomes and proactively shape the loss distribution, particularly aspects such as variability, risks posed by extreme losses, etc.

Mean-risk optimization models, which seeks to minimize a measure of risk while meeting a target mean return, is a prominent approach by which a modeler may introduce risk-aversion while optimizing under uncertainty. When performing optimization, a convex risk measure like conditional-value at risk (CVaR) becomes particularly appealing due to its ability to quantitatively capture distribution tail risks while retaining the convexity of the objective. Roughly speaking, CVaR at a quantile level $1-\beta$ captures the loss due to top β-fraction of the samples. Since the introduction of CVaR for optimization under uncertainty in Rockafellar and Uryasev (2000) and Uryasev (2000), mean-CVaR optimization modeling has become one of the most common vehicle for managing risk in numerous operations research applications, and as well in a number of related engineering disciplines.

As with most stochastic optimization formulations, solving a mean-CVaR optimization model is typically tackled by approximating the mean and CVaR by their respective Monte Carlo sample average approximations (SAA). Despite the bottleneck that only a small fraction of the samples contribute to the evaluation of the CVaR criterion, SAA remains the most preferred solution approach due to its simplicity and near-universality in use: The SAA procedure and methods for inferring its solution quality remain unchanged as long as the objective and constraints possess finite second moments (see e.g., Shapiro 1991; Homem-de Mello and Bayraksan 2014). For a risk-averse optimization formulation involving CVaR at quantile level $1-\beta$, the sample requirement for SAA to work gets blown up by a large multiplicative

factor of $O(1/\beta)$, as $\beta \to 0$, when compared to the risk-neutral counterparts. This unfortunately leads to extraordinarily large formulations if one is employing a deterministic solver, and slower convergence if one is employing stochastic gradients.

If we view the literature on tackling rare tail events in simulation, we witness Importance Sampling (IS) as one of the most prominent variance reduction approaches used for substantially reducing the sample requirement involved in estimating rare event probabilities and related expectations. Can importance sampling be used, to a similar degree of effectiveness, in optimization as well? This tutorial is dedicated to concisely introducing the ingredients required for using importance sampling for optimization under rare events. As the literature on simulation of rare events is rich and several beautiful expository reviews have been written on importance sampling for rare events, this tutorial will restrict its discussion to the scope of using IS for optimization, specifically for mean-CVaR formulations. The focus on CVaR is for the purposes of clarity, and the discussed methods extend even if CVaR is replaced by a different tail risk measure (such as expected excess loss) which preserves convexity.

The basic idea behind IS is to accelerate the occurrences of the tail risk events by sampling from alternate distributions which place greater emphasis on the risk scenarios of interest. Observed samples are then suitably reweighed to eliminate the bias introduced. This tutorial will specifically focus on the following two ingredients: (1) how one may arrive at an effective change of measure for evaluating at the objective at any given decision; and (2) how one may incorporate these decision-dependent changes of measure in a solution paradigm for minimizing CVaR.

The rest of the tutorial is organized as follows: Section 2 provides a definition of CVaR and introduces the risk-averse optimization formulation we shall be primarily considering in this paper. Section 3 provides an introduction to importance sampling and discusses how one may arrive at an effective change of measure for a given decision with illustrative examples. Section 4 introduces a retrospective approximation approach, which can be understood as performing SAA with importance samples, while incorporating lazy-updates for changing the importance sampling distribution. Section 5 presents an adaptive stochastic approximation procedure which could be suitable if iterative gradient-descent methods are preferred. The methods are accompanied by results on the magnitude of sample reductions offered by efficient IS, relative to SAA, in obtaining an optimal solution of desired quality.

2 RISK-AVERSE OPTIMIZATION DRIVEN BY CONDITIONAL VALUE AT RISK

In this section, we define the notion of conditional value at risk (CVaR), describe two prominent CVaR-driven optimization formulations, and briefly note the merits and challenges in using SAA to solve them.

2.1 Conditional Value at Risk and its Variational Representation

Let $\boldsymbol{X}$ be a random vector modeling the collection of uncertain variables affecting an optimization problem. Suppose that $\ell(\boldsymbol{x}, \boldsymbol{\theta})$ denotes the loss incurred for a choice of decision $\boldsymbol{\theta}$ when the random vector $\boldsymbol{X}$ realises the value $\boldsymbol{x}$. The Value at Risk (VaR) of the loss $\ell(\boldsymbol{X}, \boldsymbol{\theta})$ at a tail level $\beta \in (0,1)$ is simply the loss quantile

$$v_\beta(\boldsymbol{\theta}) = \inf\{u : P(\ell(\boldsymbol{X}, \boldsymbol{\theta}) \geq u) \leq \beta\}.$$

The Conditional Value at Risk (CVaR) of the loss $\ell(\boldsymbol{X}, \boldsymbol{\theta})$ at a tail level β is the average loss given that $\ell(\boldsymbol{X}, \theta)$ exceeds the respective value at risk: specifically,

$$C_\beta(\boldsymbol{\theta}) = E\left[\ell(\boldsymbol{X}, \boldsymbol{\theta}) \mid \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq v_\beta(\boldsymbol{\theta})\right]. \tag{1}$$

In this paper, we shall be considering the challenges in estimating and optimizing CVaR when the tail-level β is close to zero. The following variational representation, due to Rockafellar and Uryasev (2000), makes CVaR conducive for optimization:

$$C_\beta(\boldsymbol{\theta}) = \inf_{u \in \mathbb{R}} \left\{u + \beta^{-1} E\left(\ell(\boldsymbol{X}, \boldsymbol{\theta}) - u\right)^+\right\}. \tag{2}$$

Here $(\ell(\boldsymbol{X},\boldsymbol{\theta})-u)^+$ denotes the positive part $\max\{\ell(\boldsymbol{X},\boldsymbol{\theta})-u,0\}$, which captures the extent of excess loss above a level u. The value at risk $v_\beta(\boldsymbol{\theta})$ is an optimal solution in the variational representation (2) (see Rockafellar and Uryasev 2000), and it is readily verifiable that substituting the choice $u = v_\beta(\boldsymbol{\theta})$ in the objective in (2) yields the right-hand side in (1).

2.2 Stochastic Optimization Formulations Incorporating CVaR

Formulation 1 (Minimizing CVaR, potentially with constraints on the mean) Equipped with the above variational representation, if one wishes to minimize CVaR of a loss $\ell(\boldsymbol{X},\boldsymbol{\theta})$ over decision alternatives $\boldsymbol{\theta}$ in the set $\Theta \subseteq \mathbb{R}^p$, they may do so by solving the right-hand side of (3) below.

$$c_\beta := \inf_{\boldsymbol{\theta}\in\Theta} C_\beta(\boldsymbol{\theta}) = \inf_{u\in\mathbb{R},\,\boldsymbol{\theta}\in\Theta} f(u,\boldsymbol{\theta}), \tag{3}$$

where $f(u,\boldsymbol{\theta}) = E[F(\boldsymbol{X};u,\boldsymbol{\theta})]$ and

$$F(\boldsymbol{x};u,\boldsymbol{\theta}) := u + \beta^{-1}\left(\ell(\boldsymbol{X},\boldsymbol{\theta})-u\right)^+. \tag{4}$$

Observe that if the loss $\ell(\boldsymbol{x},\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$, for any fixed $\boldsymbol{x}$, then the convexity is retained in (3). The set Θ can be modeled to include constraints on the decisions one may wish to impose, as illustrated in Example 1 below.

Example 1 (Portfolio optimization)**.** The task of constructing a linear portfolio with minimum risk while meeting a target return is among the simplest yet instructive examples one may consider. Suppose $\boldsymbol{X}$ is an $\mathbb{R}^d$-valued random vector modeling the returns of d-assets. For a linear portfolio model which places a weight θ_i over the asset i, for $i = 1,\ldots,n$, the return realization gets specified by $\boldsymbol{\theta}^\intercal\boldsymbol{X}$. In this case, we can take the portfolio loss to be $\ell(\boldsymbol{x},\boldsymbol{\theta}) = -\boldsymbol{\theta}^\intercal\boldsymbol{x}$. It is convenient to require the weights placed over the d assets to add up to 1. Therefore, when an investor seeks to meet a target return $t \in (0,+\infty)$, the constraint set Θ can be specified as in

$$\Theta = \{\boldsymbol{\theta}\in\mathbb{R}^d_+ : \mathbf{1}^\intercal\boldsymbol{\theta} = 1,\ \boldsymbol{\mu}^\intercal\boldsymbol{\theta} \geq t\}, \tag{5}$$

where the vector $\boldsymbol{\mu}$ is the mean vector of the d assets.

Formulation 2 (Mean-CVaR optimization) Another convenient model for introducing risk aversion is to consider a convex combination of the CVaR criterion and the risk-neutral expected value objective as below:

$$\inf_{\boldsymbol{\theta}\in\Theta}\left\{\lambda E[\ell(\boldsymbol{X},\boldsymbol{\theta})] + (1-\lambda)C_\beta(\boldsymbol{\theta})\right\}, \tag{6}$$

where Θ is a convex subset of the euclidean space and λ is a parameter governing the risk-appetite of a decision-maker. One may use a smaller value of λ to specify a smaller appetite for risk. From the variational representation in (2), the above mean-CVaR model simplifies to

$$\inf_{u\in R,\boldsymbol{\theta}\in\Theta}\left\{\lambda E\left[\ell(\boldsymbol{X},\boldsymbol{\theta})\right] + (1-\lambda)E\left[F(\boldsymbol{X};u,\boldsymbol{\theta})\right]\right\},$$

where $F(\cdot)$ is defined, as before, in (4).

Example 2 (Risk-averse two stage linear programs)**.** In two-stage formulations, a decision-maker takes an action in the first-stage; and in the wake of the realization of the random vector $\boldsymbol{X}$, he/she additionally gets to make a recourse decision augmenting the first-stage decision. Usually, a recourse decision is interpreted as utilizing the extra information to compensate for any bad effects that might have been experienced as a result of first-stage action. A risk-averse two-stage linear program can be formulated as in (6), with

$$\ell(\boldsymbol{x},\boldsymbol{\theta}) = c^\intercal\boldsymbol{\theta} + Q(\boldsymbol{x},\boldsymbol{\theta}),$$

where $Q(\boldsymbol{x}, \boldsymbol{\theta})$ is the optimal value of a second-stage linear program. The following is an example of a second-stage formulation:

$$Q(\boldsymbol{x}, \boldsymbol{\theta}) = \inf\{\boldsymbol{y}^\intercal \boldsymbol{x} : T\boldsymbol{\theta} + W\boldsymbol{y} = \boldsymbol{h}, \boldsymbol{y} \geq \boldsymbol{0}\}$$

where T and W are suitably dimensioned matrices referred to as "tender" and "recourse" matrices. In this example, we have taken these matrices to deterministic. Allowing them to be random provides additional modeling power. One may refer to Shapiro et al. (2021), Chapters 1-2 for a comprehensive introduction to two-stage stochastic programming formulations and applications.

2.3 Sample-Average Approximation (SAA)

For solving (3), one may consider its sample average approximation in which the expectations $E(\ell(\boldsymbol{X}, \boldsymbol{\theta}) - u)^+$ in the objective, for all $\boldsymbol{\theta} \in \Theta$, are replaced by the respective average over independent observations of $\boldsymbol{X}$. In particular, given n i.i.d. samples of data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the distribution of $\boldsymbol{X}$, the sample averaged objective is denoted by,

$$\hat{f}_n(u, \boldsymbol{\theta}) \;=\; u + (n\beta)^{-1} \sum_{i=1}^{n} (\ell(\boldsymbol{X}_i, \boldsymbol{\theta}) - u)^+. \tag{7}$$

Then a sample average approximation (SAA) to the optimization problem (3) may be specified as,

$$\hat{c}_n = \inf_{u \in \mathbb{R}, \boldsymbol{\theta} \in \Theta} \hat{f}_n(u, \boldsymbol{\theta}). \tag{8}$$

Likewise, a sample-average approximation to the mean-CVaR formulation (6) is given by,

$$\inf_{u \in \mathbb{R}, \boldsymbol{\theta} \in \Theta} \left\{ \lambda / n \sum_{i=1}^{n} \ell(\boldsymbol{X}_i, \boldsymbol{\theta}) + (1 - \lambda) \hat{f}_n(u, \boldsymbol{\theta}) \right\}. \tag{9}$$

Approximating the true expectations in the objective by their sample averages is the simplest approximation one can perform, and it has a broad appeal due to its simplicity and near-universal applicability. The (sample requirements for the) resulting formulations could be extraordinarily large however when handling tail expectations: Observe that the term $(\ell(\boldsymbol{X}_i, \boldsymbol{\theta}) - u)^+$ appearing in (7) is often zero for most observations $\boldsymbol{X}_i$, when a search is conducted over the variable u to find its optimal value $v_\beta(\boldsymbol{\theta})$. This is due to the value-at-risk $v_\beta(\boldsymbol{\theta})$ being the $(1-\beta)$-th quantile of the loss distribution. Thus, a large fraction of terms in the summation in (7) will be zero, which is in line with CVaR being a tail risk measure.

Large sample properties such as consistency and asymptotic normality are well-known for SAA estimators (see Shapiro (1991), Theorem 3.2)). An application of these properties to identify the number of samples n required in (8) to approximate (3) reveals that the number of samples required scales inversely proportional to the tail level β of interest. This observation is consistent with the understanding that one would need approximately $\tilde{O}(\beta^{-1})$ samples, as $\beta \searrow 0$, in order to witness loss scenarios exceeding the $(1-\beta)$-th quantile captured by the value at risk.

To gather a sense of the magnitude of the number of samples required, consider the following portfolio optimization example from Caccioli et al. (2018): Even at a tail level of $\beta = 1/40$, it has been observed that one would need about 14 years of observations to achieve a 10% relative error in the optimum portfolio's CVaR for 100 stocks. A more detailed discussion on this example is available in Caccioli et al. (2018). The perils of minimizing CVaR with insufficient samples are also discussed with the help of a detailed empirical study in Lim et al. (2011). The computational effort required for solving (8) becomes exorbitantly large as a consequence of the large sample requirement, specifically if the tail level β is small. Small values of β may be particularly pertinent if a situation demands high reliability, as may be required in settings such as electric power dispatch and the design of cyber-physical systems.

To keep the discussion centred on overcoming the challenges due to this rarity in CVaR minimization, we focus on the formulation (3) in the rest of this tutorial and assume that the constraints in the decision

set Θ are specified in terms of explicitly known quantities: For example, if the constraint set is as in (5), we assume that the mean vector $\boldsymbol{\mu}$ is known, even though in practice, the mean has to be estimated by the empirical mean or its variants. The importance sampling techniques we develop in the subsequent sections are for approximating $E[F(\boldsymbol{X}; u, \boldsymbol{\theta})]$, and they can be used for both the CVaR minimization (3), mean-CVaR (6) formulations, and as well their variants. In particular, we shall develop importance sampling based estimator $\hat{f}_{\mathtt{is},n}(u, \boldsymbol{\theta})$ for $E[F(\boldsymbol{X}; u, \boldsymbol{\theta})]$, which can be used as a replacement for the sample average approximation $\hat{f}_n(u, \boldsymbol{\theta})$ in the same way in both the SAA objectives (7) and (9).

3 VARIANCE REDUCTION WITH IMPORTANCE SAMPLING

This section introduces the basic idea behind importance sampling (IS) and outlines techniques one may employ to derive IS distributions for any fixed decision choice $\boldsymbol{\theta} \in \Theta$.

3.1 Importance Sampling for Estimation of Tail Risks

Consider any fixed choice of $\boldsymbol{\theta} \in \Theta$. Recall that the main difficulty in the estimation of $E[(\ell(\boldsymbol{X}, \boldsymbol{\theta}) - u)^+]$ in (3) is the lack of samples in the excess loss region $\{\ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u\}$, for values of u around the $(1-\beta)$-th quantile of the loss. IS attempts to overcome this drawback by instead drawing samples from an alternative distribution under which this tail event occurs more frequently. The bias incurred by sampling from a different distribution is compensated by suitably weighing the resulting observations. Specifically, let $\boldsymbol{Z}$ be another random vector of our choice whose probability density $f_{\boldsymbol{Z}}(\boldsymbol{z}) > 0$ is absolutely continuous with respect to that of $\boldsymbol{X}$: that is $f_{\boldsymbol{Z}}(\boldsymbol{z}) > 0$ whenever $f_{\boldsymbol{X}}(\boldsymbol{z}) > 0$. Now, consider the following weighted estimator for the objective,

$$\hat{f}_{\mathtt{is},n}(u, \boldsymbol{\theta}) = u + \frac{1}{n\beta} \sum_{i=1}^{n} (\ell(\boldsymbol{Z}_i, \boldsymbol{\theta}) - u)^+ \frac{f_{\boldsymbol{X}}(\boldsymbol{Z}_i)}{f_{\boldsymbol{Z}}(\boldsymbol{Z}_i)} \tag{10}$$

where $\boldsymbol{Z}_1, \ldots \boldsymbol{Z}_n$ are sampled i.i.d. from the distribution of $\boldsymbol{Z}$. The "weights" in the above estimators are likelihood ratios $f_{\boldsymbol{X}}(\boldsymbol{Z}_i)/f_{\boldsymbol{Z}}(\boldsymbol{Z}_i)$ accompanying each observation of the excess loss $(\ell(\boldsymbol{Z}_i, \boldsymbol{\theta}) - u)^+$. Had there been no change in the density from which the samples are obtained (that is, if the samples are obtained from the original density $f_{\boldsymbol{X}}(\cdot)$ itself), then observe that we get back the SAA objective in (7).

Suppose that $\hat{f}_{\mathtt{is},n}$ has finite variance, and let any alternative distribution choice $f_{\boldsymbol{Z}}$ which possess this property be labeled as "admissible". Then observe that as the number of samples increase, the IS objective $\hat{f}_{\mathtt{is},n}(u, \boldsymbol{\theta})$ approximates the desired objective $f(u, \boldsymbol{\theta}) = u + \beta^{-1} E[(\ell(\boldsymbol{X}, \boldsymbol{\theta}) - u)^+]$ due to the following:

$$\begin{aligned}
\hat{f}_{\mathtt{is},n}(u, \boldsymbol{\theta}) \to u + \beta^{-1} E\left[(\ell(\boldsymbol{Z}, \boldsymbol{\theta}) - u)^+ \frac{f_{\boldsymbol{X}}(\boldsymbol{Z})}{f_{\boldsymbol{Z}}(\boldsymbol{Z})}\right] &= u + \beta^{-1} \int_{\boldsymbol{z}} (\ell(\boldsymbol{z}, \boldsymbol{\theta}) - u)^+ \frac{f_{\boldsymbol{X}}(\boldsymbol{z})}{f_{\boldsymbol{Z}}(\boldsymbol{z})} f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z} \\
&= u + \beta^{-1} \int_{\boldsymbol{z}} (\ell(\boldsymbol{z}, \boldsymbol{\theta}) - u)^+ f_{\boldsymbol{X}}(\boldsymbol{z}) d\boldsymbol{z} = u + \beta^{-1} E[(\ell(\boldsymbol{X}, \boldsymbol{\theta}) - u)^+] = f(u, \boldsymbol{\theta}),
\end{aligned}$$

as $n \to \infty$, and therefore the estimator in (10) is consistent. In the above chain, the first equality holds due to law of large numbers. The above equations also show that there is no bias introduced in this approximation: That is, $E[\hat{f}_{\mathtt{is},n}(u, \boldsymbol{\theta})] = f(u, \boldsymbol{\theta})$, for every choice of u and $\boldsymbol{\theta}$.

While every admissible change of distribution approximates the target objective as above, the key to approximating well with a substantially smaller number of samples relies on making a good choice for the IS distribution $\boldsymbol{Z}$. Indeed in the estimation of rare event probabilities, there is a considerable litreature on how to arrive at good choices for IS distributions: See, for example, Heidelberger (1995), Asmussen and Glynn (2007), Juneja et al. (2007), or, more recently Blanchet et al. (2019), Bai et al. (2022), Deo and Murthy (2021) for treatments on objectives which may have a greater relevance from an optimization point of view.

Keeping typical objectives arising in optimization in view, we next describe two approaches which could be considered for arriving at an effective IS distribution for any fixed decision $\boldsymbol{\theta} \in \Theta$.

3.2 Approach 1: IS via Exponential Twisting Based on Dominating Points

The first approach for deriving a good importance sampling change of measure is to explicitly use the distribution of $\boldsymbol{X}$ and the loss $\ell(\boldsymbol{X},\boldsymbol{\theta})$ to carefully arrive at an IS distribution choice. Executing this approach typically involves two steps.

Step 1: Use the log-moment generating function of $\boldsymbol{X}$ to identify the so-called "dominating points" of the excess loss set $\{\boldsymbol{x}:\ell(\boldsymbol{x},\boldsymbol{\theta})\geq u\}$, with reference to the given distribution for $\boldsymbol{X}$. Roughly speaking, the dominating points are a collection of points in the excess loss set $\{\boldsymbol{x}:\ell(\boldsymbol{x},\boldsymbol{\theta})\geq u\}$ such that each dominating point captures, in a local sub-region of $\{\boldsymbol{x}:\ell(\boldsymbol{x},\boldsymbol{\theta})\geq u\}$, the most likely way the excess loss event happens.
Step 2: Once the dominating points are identified, one typically chooses the IS distribution to be a mixture distribution, with each component distribution of the mixture being chosen to be an "exponentially tilted" distribution whose mean coincides with one of the dominating points. Thus one may need as many component distributions as the number of dominating points.

By placing the emphasis on the dominating points, the guiding principle here is to ensure that locally within sub-regions in $\{\boldsymbol{x}:\ell(\boldsymbol{x},\boldsymbol{\theta})\geq u\}$, the points which are more likely to be observed are indeed given more probability mass. See, for e.g.., Arief et al. (2021), Definition 2 for a definition of dominating points.

Definition 1 (Exponentially tilted densities) Given a probability density $f_{\boldsymbol{X}}$ for the random vector $\boldsymbol{X}$, we call a new density g to be exponentially titled version of $f_{\boldsymbol{X}}$ with a tilt factor $\boldsymbol{b}$ if

$$g(\boldsymbol{x})\propto\exp\left(\boldsymbol{b}^{\intercal}\boldsymbol{x}\right)f_{\boldsymbol{X}}(\boldsymbol{x}),$$

and $E[\exp(\boldsymbol{b}^{\intercal}\boldsymbol{X})]$ is finite. In particular, we have

$$g(\boldsymbol{x})=\frac{\exp\left(\boldsymbol{b}^{\intercal}\boldsymbol{x}\right)}{E[\exp(\boldsymbol{b}^{\intercal}\boldsymbol{X})]}f_{\boldsymbol{X}}(\boldsymbol{x})=\exp\left(\boldsymbol{b}^{\intercal}\boldsymbol{x}-\Lambda(\boldsymbol{b})\right)f_{\boldsymbol{X}}(\boldsymbol{x}),\tag{11}$$

where $\Lambda(\boldsymbol{r})$ is the log-moment generating function defined by $\Lambda(\boldsymbol{r})=\log E[\exp(\boldsymbol{r}^{\intercal}\boldsymbol{X})]$, for $\boldsymbol{r}\in\mathbb{R}^d$.

Lemma 1. Suppose that the log-moment generating function $\Lambda(\boldsymbol{r})=E[\exp(\boldsymbol{r}^{\intercal}\boldsymbol{X})]$ is finite and differentiable at $\boldsymbol{r}=\boldsymbol{b}$. If $\boldsymbol{Z}$ is distributed according to the exponentially tilted density $g(\cdot)$ in (11), then $E[\boldsymbol{Z}]=\nabla\Lambda(\boldsymbol{b})$.

Proof. The conclusion follows directly from the following two deductions (the derivative and integral below can be interchanged by dominated convergence):

$$E[\boldsymbol{Z}]=\int\boldsymbol{z}g(\boldsymbol{z})d\boldsymbol{z}=\int\boldsymbol{z}\frac{\exp(\boldsymbol{b}^{\intercal}\boldsymbol{z})}{E[\exp(\boldsymbol{b}^{\intercal}\boldsymbol{X})]}f_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}=\frac{E\left[\boldsymbol{X}\exp(\boldsymbol{b}^{\intercal}\boldsymbol{X})\right]}{E\left[\exp(\boldsymbol{b}^{\intercal}\boldsymbol{X})\right]};\text{ and}$$

$$\nabla\Lambda(\boldsymbol{r})=\int\nabla\exp(\boldsymbol{r}^{\intercal}\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}=\int\boldsymbol{x}\exp(\boldsymbol{r}^{\intercal}\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}=\int\boldsymbol{x}\frac{\exp(\boldsymbol{r}^{\intercal}\boldsymbol{x})}{E[\exp(\boldsymbol{r}^{\intercal}\boldsymbol{X})]}f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}=\frac{E\left[\boldsymbol{X}\exp(\boldsymbol{r}^{\intercal}\boldsymbol{X})\right]}{E\left[\exp(\boldsymbol{r}^{\intercal}\boldsymbol{X})\right]}.$$

Equipped with the above definition of exponentially tilted density and its properties, Example 3 below demonstrates how one may execute the two-step procedure indicated above.

Example 3. Consider the piece-wise linear loss, $\ell(\boldsymbol{x},\boldsymbol{\theta})=\max\{\boldsymbol{\theta}^{\intercal}\boldsymbol{A}_i\boldsymbol{x}:i=1,\ldots,M\}$, where $\{\boldsymbol{A}_i:i=1,\ldots,M\}$ are $p\times d$ matrices. Let $\boldsymbol{X}$ be light-tailed with log-moment generating function, $\Lambda(\boldsymbol{r})=\log E[\exp(\boldsymbol{r}^{\intercal}\boldsymbol{X})]$, for $\boldsymbol{r}\in\mathbb{R}^d$. For this example, we take $\Lambda(\cdot)$ to be strictly convex, differentiable, and finite for every $\boldsymbol{r}\in\mathbb{R}^d$. An important quantity useful for identifying a good choice of IS distribution is the convex conjugate of the log-moment generating function $\Lambda(\cdot)$ defined as follows: $\Lambda^*(\boldsymbol{x})=\sup_{\boldsymbol{r}\in\mathbb{R}^d}\{\boldsymbol{r}^{\intercal}\boldsymbol{x}-\Lambda(\boldsymbol{r})\}$. In the example of $\boldsymbol{X}$ being a multivariate normal random vector with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, we have $\Lambda(\boldsymbol{r})=\boldsymbol{m}^{\intercal}\boldsymbol{r}+\boldsymbol{r}^{\intercal}\boldsymbol{\Sigma}\boldsymbol{r}/2$ and $\Lambda^*(\boldsymbol{x})=(\boldsymbol{x}-\boldsymbol{m})^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})/2$. For $i=1,\ldots,M$, let

$$\boldsymbol{a}_i=\arg\min_{\boldsymbol{x}}\left\{\Lambda^*(\boldsymbol{x}):\boldsymbol{\theta}^{\intercal}\boldsymbol{A}_i\boldsymbol{x}\geq u\right\}.\tag{12}$$

See that the excess loss set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \theta) \geq u\}$ which is of interest to us can be seen as the union of the sub-regions $\mathscr{R}_i := \{\boldsymbol{x} : \boldsymbol{\theta}^\intercal \boldsymbol{A}_i \boldsymbol{x} \geq u\}$. We next use $\{\boldsymbol{a}_i : i = 1, \ldots, M\}$ to find the corresponding collection of roots $\{\boldsymbol{b}_i : i = 1, \ldots, M\}$ satisfying

$$\nabla \Lambda(\boldsymbol{b}_i) = \boldsymbol{a}_i, \tag{13}$$

for $i = 1, \ldots, M$. With these definitions, the following two observations are in order:

(i) Due to the property of convex conjugates, we have that the convex conjugate of $\Lambda^*(\cdot)$ is, in turn, $\Lambda(\cdot)$ itself in this example. As a consequence, we also symmetrically have $\nabla \Lambda^*(\boldsymbol{a}_i) = \boldsymbol{b}_i$, for $i = 1, \ldots, M$. Therefore, from the optimality conditions for $\boldsymbol{a}_i$ in (12), we have for $i = 1, \ldots, M$,

$$\boldsymbol{b}_i^\intercal \left(\boldsymbol{x} - \boldsymbol{a}_i\right) \geq 0, \qquad \text{for all } \boldsymbol{x} \text{ in the sub-region } \mathscr{R}_i.$$

(ii) From (13) and Lemma 1, we have $E[\boldsymbol{Z}] = \nabla \Lambda(\boldsymbol{b}_i) = \boldsymbol{a}_i$ when $\boldsymbol{Z}$ is distributed with an exponential tilting by a factor $\boldsymbol{b}_i$ as in,

$$g_i(\boldsymbol{x}) \propto e^{\boldsymbol{b}_i^\intercal \boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}). \tag{14}$$

For the above reasons, the collections $\{\boldsymbol{a}_i : i = 1, \ldots, M\}$ and $\{\boldsymbol{b}_i : i = 1, \ldots, M\}$ are called as dominating points and tilt parameters, respectively. With the tilt factors computed as roots $\{\boldsymbol{b}_i : i = 1, \ldots, M\}$ in (13), we select our IS distribution to be the mixture density

$$f_{\boldsymbol{Z}}(\boldsymbol{x}) = \sum_{i=1}^{M} p_i g_i(\boldsymbol{x}), \tag{15}$$

in which the mixture component densities g_i are obtained by exponentially tilting the original density $f_{\boldsymbol{X}}(\cdot)$ by a factor $\boldsymbol{b}_i$ (as in (14)) and the positive mixture weights satisfy $\sum_{i=1}^{m} p_i = 1$. The i-th component in the mixture density, $g_i(\cdot)$, has the dominating point $\boldsymbol{a}_i$ in the rare set as its mean, thereby placing a prominent amount of probability mass in the target rare set. Besides this property, the exponentially tilted densities g_i are such that the respective likelihood ratios $f(\boldsymbol{x})/g_i(\boldsymbol{x})$ stay controlled (small) throughout the rare sub-regions $\mathscr{R}_i = \{\boldsymbol{x} : \boldsymbol{\theta}^\intercal \boldsymbol{A}_i \boldsymbol{x}_i \geq u\}$. □

Merits and challenges in executing IS via dominating points. We shall see later in Section 3.4 that IS via dominating points, as in Example 3 above, reduces the sample requirements to a great degree, if (i) the distribution of $\boldsymbol{X}$ is light-tailed, (ii) one can get hold of the dominating points in (12) and the tilting parameters (13) with relatively low effort, and (iii) there are not too many dominating points for any given choice of $\boldsymbol{\theta} \in \Theta$. These conditions are usually not met, though, for the following reasons:

a) One may often need to solve complicated, potentially nonconvex optimization problems described in terms of the log-moment generating function $\Lambda(\cdot)$, its convex conjugate Λ^*, and the excess loss set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u\}$ to identify the dominating points. This is made further intractable by the requirements that (i) such complicated optimization problems (for identifying dominating points and tilt parameters) have to be solved repeatedly for different choices of $(u, \boldsymbol{\theta})$; and (ii) a description of log-moment generating function $\Lambda(\cdot)$ and its convex conjugate $\Lambda^*(\cdot)$ may not be available in most problems and have to be estimated additionally via Monte Carlo.

b) Further, the number of dominating points may be quite large if the objective $L(\cdot)$ is complex (or) if the ambient dimension of $\boldsymbol{X}$ is not small. In such cases where the number of dominating points are large, identifying all of them and determining how to arrive at the weighting probabilities p_i in (15) may be non-trivial.

While the above issues have somewhat limited the applicability of IS to a relatively narrow collection of instances, the last couple of years have witnessed efforts explicitly directed towards overcoming these limitations (see, example, He et al. (2023), Deo and Murthy (2023), Arief et al. (2021) and references

therein). Assuming access to the log-moment generating function $\Lambda(\cdot)$, attempts towards making IS based on dominating points applicable for more complex objectives have been undertaken in Arief et al. (2021) and Bai et al. (2023). In particular, if the number of samples to be drawn is delicately chosen to be neither too small nor too large, Bai et al. (2023) observes that it may be sufficient to choose the mixture distribution based on dominating points from the sub-collection $\text{argmin}\{\Lambda^*(\boldsymbol{a}_i) : i = 1,\ldots,M\}$. The approach presented in Section 3.3 below, based on Deo and Murthy (2023), is a radically different approach aiming to overcome the difficulties (a) - (b) above by implicitly learning a good IS distribution from the samples of $\boldsymbol{X}$.

3.3 Approach 2: IS via Self-Structuring Transformations

In this approach, the search for effective IS distributions is instead recast as follows: "Can we find a single transformation $\boldsymbol{T}(\cdot)$ whose respective push-forward distribution (i.e., the law of $\boldsymbol{T}(\boldsymbol{X})$) readily serves as an effective IS distribution when deployed across a large class of problems?" This re-framed pursuit, seeking to induce an effective IS distribution *implicitly* via a map $\boldsymbol{T}(\cdot)$, bypasses the need to explicitly tailor the IS distribution to every decision choice and to every problem. We shall see that this problem agnostic nature of the approach allows it to be simpler to use, making it closer in spirit to the SAA. The approach is sufficiently simple to render itself to be readily applicable even for the two-stage programs in Example 2.

To explain the use of this IS approach towards solving (3), suppose that ρ is a positive constant capturing the asymptotic growth rate of the loss $\ell(\boldsymbol{x},\boldsymbol{\theta})$ as a function of $\boldsymbol{x}$: that is, $\lim_{n\to\infty}\ell(n\boldsymbol{x},\boldsymbol{\theta})/n^{\rho} > 0$, for some $\boldsymbol{\theta}\in\Theta$ and $\lim_{n\to\infty}\ell(n\boldsymbol{x},\boldsymbol{\theta})/n^{\rho} < +\infty$, for all $\boldsymbol{\theta}\in\Theta$. For the piece-wise linear and two-stage losses in Examples 1 - 3, we have $\rho = 1$. For quadratic losses such as in the Delta-Gamma approximation for portfolio returns in Glasserman et al. (2000), we have $\rho = 2$. Define the $\mathbb{R}^d$-valued function

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

where for $h > 0$, we take $s_h = h\max\{\log\log(1/\beta),1\}$. The positive number s_h can be viewed as a scalar stretch factor which allows the transformation $\boldsymbol{T}_h$ to stretch the different components of $\boldsymbol{x} = (x_1,\ldots,x_d)$ differently via the vector-valued exponent $\boldsymbol{\kappa}(\boldsymbol{x}) = (\kappa_1(\boldsymbol{x}),\ldots,\kappa_d(\boldsymbol{x}))$ defined as below:

$$\kappa_i(\boldsymbol{x}) := \frac{\log(1+|x_i|)}{\rho\log(1+\|\boldsymbol{x}\|_\infty)}, \quad i = 1,\ldots,d. \tag{16}$$

The scalar stretch factor s_h, when viewed as a function of tail level β, is larger when the estimation problem is made rarer by letting β smaller. 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 (2023), Proposition 1)) and the transformed vector $\boldsymbol{Z} = \boldsymbol{T}_h(\boldsymbol{X})$ 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,

$$\mathscr{L}_h = f_{\boldsymbol{X}}(\boldsymbol{Z})/f_{\boldsymbol{Z}}(\boldsymbol{Z}) = [f_{\boldsymbol{X}}(\boldsymbol{Z})/f_{\boldsymbol{X}}(\boldsymbol{X})]\,J_h(\boldsymbol{X}) \tag{17}$$

where the Jacobian, $J_h(\cdot)$, of the transformation equals,

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

The hyper-parameter h may be selected by setting $h(\beta) = k\log(\log(1/\beta))$, and then tuning k using a line search; we refer readers to Deo and Murthy (2021) for more details.

From the i.i.d. samples $\boldsymbol{Z}_i = \boldsymbol{T}_h(\boldsymbol{X}_i), i = 1,\ldots,d$, we have the following IS estimator for the objective function in (3):

$$\hat{f}_{\mathtt{is},n}(u,\boldsymbol{\theta}) = \left[u + \frac{1}{n\beta}\sum_{i=1}^{n}(\ell(\boldsymbol{Z}_i,\boldsymbol{\theta}) - u)^+\mathscr{L}_{h,i}\right], \tag{18}$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. copies of $\boldsymbol{X}$ and $\mathscr{L}_{h,i} = J_h(\boldsymbol{X}_i) f_{\boldsymbol{X}}(\boldsymbol{Z}_i)/f_{\boldsymbol{X}}(\boldsymbol{X}_i)$ denotes the corresponding likelihood in (17).

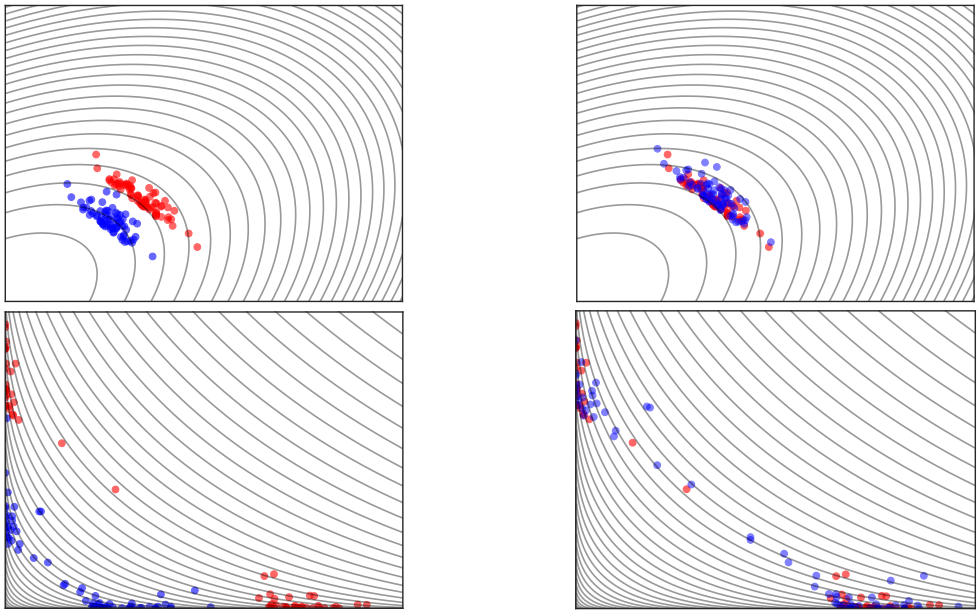

Figure 1: An illustration of how the self-structuring transformation $\boldsymbol{T}_h(\cdot)$ helps induce IS distributions with desirable concentration properties: The two panels in the left show the conditional excess loss samples of $\boldsymbol{X}$ falling in the target tail set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u\}$ (in red) and in the significantly less rare tail set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq l\}$, for a level $l \ll u$ (in blue) and for two different choices of distributions for $\boldsymbol{X}$. The respective panels in the right show how the same transformation $\boldsymbol{T}_h$, when applied to the blue samples in the top and bottom left panels, manages to replicate the concentrations of the respective target excess loss samples in red.

A main reason to use the map $\boldsymbol{T}_h(\cdot)$ is that it is capable of making use of the similarities in the distributions of the excess loss samples at different levels of rarity in order to induce a good IS distribution. We devote the rest of this section towards developing an intuitive understanding of this property with the aid of Figure 1 above. For values of l smaller than u, this approach builds on the premise that the excess loss samples falling in the target rare set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u\}$ and those in the less rare set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u\}$ exhibit a remarkable similarity in how they concentrate. The figures in the left panel in Figure 1 illustrate this property by comparing the conditional samples of $\boldsymbol{X}$ in the target tail event $\{\ell(\boldsymbol{X}, \boldsymbol{\theta}) \geq u\}$ (red points) against those in the less rare event $\{\ell(\boldsymbol{X}, \boldsymbol{\theta}) \geq l\}$ (blue points). In the upper panel in the left , $\boldsymbol{X} = (X_1, X_2)$ is drawn from a bivariate Gaussian distribution, while in the lower panel $\boldsymbol{X}$ has i.i.d. Weibull marginals satisfying $P(X_1 \leq x) = 1 - \exp(-x^{0.6})$. For this illustration, the levels u and l are chosen, respectively, to equal the $(1 - 10^{-4})$-th and $(1 - 10^{-2})$-th quantile of the loss; roughly speaking, the blue samples falling in the the latter region get observed 100 times more frequently than the target excess loss samples in red. The similarity observed at different tail levels in Figure 1 can be shown to hold quite generally, including various parametric and semiparametric distributions and copula families commonly used in practice (see Assumption 2 in Section 3.4 and Deo and Murthy (2023), Table 2 for a comprehensive collection of distribution families satisfying this tail-similarity property).

The map $\boldsymbol{T}_h(\cdot)$ is such that it preserves the concentration behaviour encoded in the less rare samples falling in sets of the form $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq l\}$, for $l \ll v_\beta(\boldsymbol{\theta})$, so that the transformed samples $\boldsymbol{T}_h(\boldsymbol{X})$ replicate this concentration in the target rare set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u\}$. The respective sub-figures in the right panel in Figure 1 show the distribution induced when the map $\boldsymbol{T}_h$ is applied to the less rare blue points in the left panel. In both the top and bottom panels in the right, observe that excess loss samples of $\boldsymbol{Z} = \boldsymbol{T}_h(\boldsymbol{X})$

falling in the target region (drawn in blue) concentrate in the same way as the target excess loss samples (drawn in red) concentrate. However, the blue samples are observed 100 times more frequently than the respective red samples, thereby providing a reduction in the overall sample requirement.

A key ingredient in allowing the map $\boldsymbol{T}_h$ to possess this concentration-preserving property is the exponent $\boldsymbol{\kappa}(\cdot)$ defined in (16). Briefly, $\boldsymbol{\kappa}$ ensures that the components of $\boldsymbol{X}$ are only magnified/stretched to the extent necessary, as informed by the samples in the less rare region $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq l\}$. For instance, if a particular component of $\boldsymbol{X}_i$ falling in this region is small, then the component remains small even after applying the map $\boldsymbol{T}_h$ (and vice versa). Please refer to Deo and Murthy (2023), Section 5 for a definition and properties of rate-function preserving transformations, a notion which lends precision to the idea discussed only at an intuitive level here with the help of Figure 1.

3.4 Reduction in Sample Requirements in Estimating the CVaR Objective Using IS

As in Sections 3.1 - 3.3, we fix a decision choice $\boldsymbol{\theta} \in \Theta$ and discuss the magnitude of sample reductions one may obtain with the presented IS approaches in the estimation of CVaR $C_\beta(\boldsymbol{\theta})$. Let $n_{\texttt{SAA}}(\beta)$ and $n_{\texttt{is}}(\beta)$ denote the number of samples required by SAA and IS, respectively, in order to guarantee that the objective $C_\beta(\boldsymbol{\theta})$ is estimated with a desired relative precision: In particular, the obtained estimate should be such that the relative error does not exceed a pre-specified level $\varepsilon \in (0, +\infty)$, with $(1-\alpha) \times 100\%$ confidence.

Theorem 3 For any $\delta > 0$, we have the sample-requirement reduction guarantee that

$$\frac{n_{\texttt{SAA}}(\beta)}{n_{\texttt{is}}(\beta)} \geq \frac{c}{\beta^{1-\delta}}, \qquad \text{for all } \beta < \beta_0,$$

for suitable constants $\beta_0 \in (0,1)$ and $c > 0$, if either

(i) IS density (15) based on exponential tilting is used and Assumption (1) below is satisfied by the loss $\ell(\cdot)$ and the distribution of $\boldsymbol{X}$; (or)
(i) IS estimator (18) based on self-structuring transformations is used and Assumption (2) below is satisfied by the loss $\ell(\cdot)$ and the distribution of $\boldsymbol{X}$.

Theorem 3 implies that the sample requirement due to IS grows only like $o((1/\beta^\delta))$, for any $\delta > 0$, as the tail-level $\beta \to 0$. This is in contrast to the steep $\tilde{O}(1/\beta)$ sample requirement that is inevitable with the use of SAA. Theorem 3 is a consequence of the variance reduction guarantees presented in (i) Arief et al. (2021), Theorem 1 for exponential tilting and (ii) Deo and Murthy (2021), Theorem 1 for self-structuring IS transformations. Assumptions 1 and 2 below, respectively, give the conditions under the above sample reduction requirements can be guaranteed for IS using the two approaches.

Assumption 1. (**Assumptions for Exponential Twisting).**

(i) The effective domain of the log-moment generating function of $\boldsymbol{X}$, $D(\Lambda) = \{\boldsymbol{r} : \Lambda(\boldsymbol{r}) < \infty\}$, has a non-empty interior. Moreover, $\Lambda(\cdot)$ is strictly convex and continuously differentiable in the interior of $D(\Lambda)$.
(ii) The set $\{\boldsymbol{x} : \ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u\}$ is orthogonally monotone for any $\boldsymbol{\theta} \in \Theta$: that is, if $\boldsymbol{x} \leq \boldsymbol{x}_1$ and $\ell(\boldsymbol{x}, \boldsymbol{\theta}) \geq u$ for $\boldsymbol{\theta} \in \Theta$, then it is necessary that $\ell(\boldsymbol{x}_1, \boldsymbol{\theta}) \geq u$.

Note that Part (i) of Assumption 1 requires that $\boldsymbol{X}$ is light-tailed. For stating the assumptions for self-structuring transformations, we introduce the following non-restrictive regularity notion: We call a function $f(\cdot)$ to be *multivariate regularly varying* if $f(n\boldsymbol{x})/f(n\boldsymbol{1})$ is uniformly convergent (as $n \to \infty$) in compact subsets of $\mathbb{R}^d$.

Assumption 2. (**Assumptions for Self-Structuring Transformations**).

(i) Either the density $f_{\boldsymbol{X}}(\cdot)$ or $\log f_{\boldsymbol{X}}(\cdot)$ is multivariate regularly varying.

ii) The loss $\ell(\cdot)$ satisfies the following for some $\rho > 0$: $\ell(n\boldsymbol{x}, \boldsymbol{\theta})/n^{\rho}$ is convergent uniformly in compact sets to a non-zero function.

Assumption 2 allows both light and heavy-tailed distributions for $\boldsymbol{X}$. A wide variety of parametric and semiparametric multivariate distributions, including normal, exponential family, elliptical, log-concave distributions and Archimedian copula models satisfy Assumption 2(i) (see, Deo and Murthy (2023), Table 2 for a comprehensive yet nonexhaustive collection of distributions satisfying Assumption 2 or its more general variants). In its presented form, Assumption 2(i) only allows for distributions whose marginals have a similar tail strength, To see how the self-structuring transformation continues to work well even if this condition is relaxed, we refer the reader to Deo and Murthy (2023). Part (ii) of Assumption 2 imposes a mild, non-parametric restriction on the how the loss function $\ell(\cdot)$ grows. It is satisfied easily in a number of losses which are of interest in stochastic optimization, including piecewise-linear losses, quadratic losses, and losses which occur in two-stage programs, such as in Example 2. While the asymptotic sample requirement reduction in Theorem 3 holds for any fixed choice of $h > 0$ in the transformation $\boldsymbol{T}_h$, an ideal choice of the hyperparameter h is made numerically as explained in Section 4 below.

4 RETROSPECTIVE APPROXIMATION AS A SOLUTION PARADIGM

We examined in the previous section examples for how one may arrive at effective IS distributions for any fixed decision choice $\boldsymbol{\theta} \in \Theta$. How to integrate the several IS distribution prescriptions we have, one for each decision choice, into a method for solving the optimization formulation (3)? The solution paradigm of retrospective approximation (RA) serves as a natural vehicle to utilize the change of measure prescriptions effectively towards solving the CVaR minimization in (3).

Retrospective approximation (Chen and Schmeiser (2001), Pasupathy (2010)) has been developed as a computationally attractive alternative to SAA in solving general stochastic optimization and root finding problems. For a fixed error tolerance, recall that SAA involves solving one large problem formulated with all the samples allowed by the computational budget. The premise behind RA is to reduce the overall computational effort relative to SAA by instead solving a *sequence* of SAA sub-problems: each sub-problem in the sequence is initialized with the solution of the previous sub-problem and is solved with a larger number of samples, and upto a smaller error tolerance, than its predecessors. This eases the overall computational burden as follows: The initial sub-problems are computationally light due to smaller sample sizes and larger error tolerances; the later sub-problems are computationally efficient as they are initialized with the solution from the previous stage, and refining them locally with the availability of more samples is less demanding than conducting an overall search.

4.1 Retrospective Approximation for Minimizing Tail Risks

Interestingly, this general RA paradigm becomes an ideal vehicle for executing importance sampling, as every new sub-problem offers an opportunity to obtain samples from an IS distribution which is most suited for the regions in which the search for a solution is being presently conducted.

With this guiding philosophy, the RA procedure outlined in Algorithm 1 below provides a template to suitably incorporate IS distribution prescriptions we have from Section 3 within a solution paradigm. In the general template presented in Algorithm 1, this scheme assumes that we have chosen a suitable family $\{g_\alpha : \alpha \in \mathscr{A}\}$ of IS probability densities, and a method for arriving at a good IS density choice g_α for any given selection of $(u, \boldsymbol{\theta})$ in solving the right hand side of (3). To capture this method's use succinctly in the algorithm, we represent it as an oracle mapping $\boldsymbol{b} : \mathbb{R} \times \Theta \to \mathscr{A}$ satisfying the following assumption.

Assumption 3. *The IS oracle $\boldsymbol{b} : \mathbb{R} \times \Theta \to \mathscr{A}$ is such that for every decision $(u, \boldsymbol{\theta})$, the resulting IS probability density g_α, with α set to $\alpha = \boldsymbol{b}(u, \boldsymbol{\theta})$, is the IS distribution of our choice for the estimation of $E(\ell(\boldsymbol{X}, \boldsymbol{\theta}) - u)^+$.*

Describing a solution paradigm conveniently via an IS oracle, as in the case of Algorithm 1 above, follows from the work of He et al. (2023). Considering the case of self-structuring importance samplers

Algorithm 1: Retrospective Approximation based CVaR Optimization

Input: Initial iterate $(u_0, \boldsymbol{\theta}_0)$, an increasing sequence $(n_k : k \geq 0)$ of sample-sizes with $n_0 = 0$, a decreasing sequence of error tolerances $\{\varepsilon_k : k \geq 1\}$, an initial IS parameter α_0, IS oracle $\boldsymbol{b}(\cdot)$.

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

1. Obtain importance samples: Draw i.i.d. samples $\boldsymbol{Z}_{n_{k-1}+1}, \ldots, \boldsymbol{Z}_{n_k}$ from the distribution $P_{\alpha_{k-1}}$.

2. Solve the IS based optimization: With the likelihood ratios set to $\mathscr{L}_i = f_{\boldsymbol{X}}(\boldsymbol{Z}_i)/g_{\alpha_{k-1}}(\boldsymbol{Z}_i)$ for $i = n_{k-1}+1, \ldots, n_k$, solve the following problem upto a tolerance of ε_k:

$$\hat{c}_{is,n_k} := \inf_{u,\boldsymbol{\theta}} \left[u + \frac{1}{n_k \beta} \sum_{i=1}^{n_k} \left(\ell(\boldsymbol{Z}_i, \boldsymbol{\theta}) - u\right)^+ \mathscr{L}_i, \right] = \inf_{u,\boldsymbol{\theta}} \hat{f}_{is,n_k}(u, \boldsymbol{\theta}) \tag{19}$$

with an initial solution $(u_{k-1}, \boldsymbol{\theta}_{k-1})$. Return $(u_k, \boldsymbol{\theta}_k)$ as the solution obtained by solving (19).

3. Update IS distribution choice: Set $\alpha_k = \boldsymbol{b}(u_k, \boldsymbol{\theta}_k)$. Set $k = k+1$.

Final Output: Return the solution $(u_k, \boldsymbol{\theta}_k)$

explained in Section 3.3, Deo et al. (2022) provides the above RA procedure for integrating IS with optimization. The above RA procedure can alternatively be interpreted as performing lazy-updates for IS distributions in the adaptive SAA scheme introduced by He et al. (2023).

Example 4 (An IS oracle $\boldsymbol{b}$ for use with exponential tilting)**.** In the case of IS based on dominating points and exponential tilting considered in Example 3, the following serve as the oracle mapping $\boldsymbol{b}(u, \boldsymbol{\theta})$: For any given $(u, \boldsymbol{\theta})$, consider the tilt parameters $(\boldsymbol{b}_i : i = 1, \ldots, M)$ obtained by solving (12) and (15), together with the mixture weights $(p_i : i = 1, \ldots, M)$. With these collections defining the parameters used in IS density in (15), we use $(\boldsymbol{b}_i : i = 1, \ldots, M)$ and $(p_i : i = 1, \ldots, M)$ as the oracle mapping $\boldsymbol{b}(u, \boldsymbol{\theta})$

Example 5 (An IS oracle $b(\cdot)$ for use with self-structuring IS in Section 3.3)**.** Recall that for the case of self-structuring IS transformations introduced in Section 3.3, a wide range of stretching hyperparameters h have been shown to offer good variance reduction asymptotically. If one wishes to update to a specific choice of h in Step 3 of Algorithm 1 above, they may do so via a simple cross-validation type one-dimensional search demonstrated in Algorithm 2 below. The iterate $(u, \boldsymbol{\theta})$ input to Algorithm 2 is an approximation of the solution of the stochastic optimisation problem (2). Recall that the standard error in stochastic optimisation depends on the variance of the objective evaluated at the optimal solution. Algorithm 2 seeks to approximate this using samples to arrive at a suitable value of parameter $\boldsymbol{b}(u, \boldsymbol{\theta})$.

Algorithm 2: An IS oracle $\boldsymbol{b}$ for use with Self-Structuring IS transformations

Input: iterate $(u, \boldsymbol{\theta})$, i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ from the distribution of $\boldsymbol{X}$, initial seed h_0.

1. The oracle objective, for any choice of stretch hyperparameter h, is evaluated to be the second moment defined below:

$$\hat{M}_2(h; u, \boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \left[(\ell(\boldsymbol{X}_i; \boldsymbol{\theta}) - u)^+ \right]^2 \mathscr{L}_{h,i},$$

where $\mathscr{L}_{h,i}$ denotes the likelihood ratio (17).

2 Update the cross validation parameter: Return a stretch parameter minimizing the oracle objective above as the output of the IS oracle:

$$\boldsymbol{b}(u, \boldsymbol{\theta}) \ \in \ \arg\min_{h} \hat{M}_2(h; u, \boldsymbol{\theta})$$

4.2 Guidance on Selecting Sample Size and Tolerance Parameter for RA

Assumption 4 below imposes a mild condition on the sample-size n_k and error tolerance ε_k one may need to use for the k-th sub-problem in Algorithm 1.

Assumption 4. Suppose that the sequence $\{(\varepsilon_k, n_k) : k \geq 1\}$ satisfies the following requirements:

1. If the optimization procedure used to solve (19) exhibits linear convergence, then $(n_k, \varepsilon_k : k \geq 1)$ is such that $\liminf_{k\to\infty} \varepsilon_{k-1}\sqrt{n_k} > 0$. If this procedure exhibits polynomial convergence, then $(n_k, \varepsilon_k : k \geq 1)$ is such that $\liminf_{k\to\infty} \log 1/\sqrt{n_{k-1}}(\log \varepsilon_k)^{-1} > 0$.
2. $\limsup_{k\to\infty}(\sum_{j=1}^{k} n_j)^2/\varepsilon_k^2 < \infty$ and $\limsup_{k\to\infty} n_k^{-1}\sum_{j=1}^{k} n_j < \infty$.

Assumption 4 imposes conditions so that the errors due to finite sample size and the errors due to solver error tolerance are balanced out, so that the cumulative work performed is kept minimal. For instance, condition 1 requires that the optimization error tolerance ε_k decays does not decay to 0 too fast. If this is not satisfied, then the error in solution due to imperfect optimization will be orders of magnitude smaller than the sampling error, and therefore lead to a wastage of computational effort. Likewise, achieving low variance with a large sample size while allowing a larger error tolerance in the solver is also computationally inefficient. Conditions 2 imposes a lower bound on rates at which (n_k, ε_k) converge to their limits, so that the solutions output by successive epochs do not get "stuck". For instance, if ε_k converges to 0 too slowly, then the tolerance condition may be too easily satisfied, and therefore lead to no improvement in solution. Conversely if n_k goes to ∞ too slowly, the difference in n_{k-1} and n_k is so small that the iterate does not move. In either case, the work done in the kth iteration of the RA procedure is wasted and leads to a computational suboptimality. The specifications in Assumption 4 are such that these two errors are balanced. A natural choice sample size is $n_k = \lceil cn_{k-1} \rceil$ for linearly converging optimization procedures, and $n_k = \lceil n_{k-1}^c \rceil$ for polynominally converging procedures; an optimisation procedure is linearly converging if the sequence of iterates it produces, call them $\boldsymbol{x}_1^*, \ldots$ satisfy $\limsup_{k\to\infty} \|\boldsymbol{x}_k^* - \boldsymbol{x}^*\|/(\|\boldsymbol{x}_{k-1}^* - \boldsymbol{x}^*\|) \in (0,1)$ and is said to be polynomially converging, if for some $p > 1$, $\limsup_{k\to\infty} \|\boldsymbol{x}_k^* - \boldsymbol{x}^*\|/(\|\boldsymbol{x}_{k-1}^* - \boldsymbol{x}^*\|^p) \in (0,1)$, where $\boldsymbol{x}^*$ is the solution of the optimisation problem. Meanwhile, for both these cases, $\varepsilon_k = K/\sqrt{n_k}$ is a good choice for the error tolerance. We refer interested readers to Pasupathy (2010), Section 5.3 for a detailed investigation.

4.3 Reduction in Sample Requirements Due to IS

Recall the IS oracle $\boldsymbol{b}(\cdot)$ mapping selections illustrated in Examples 4 - 5. For these IS oracles, we next show that the reduction in sample requirement exhibited in Theorem 3 for a fixed decision carries forward to solving the risk minimisation problem (3). As before, let $n_{\text{SAA}}(\beta)$ and $n_{\texttt{is}}(\beta)$ denote the number of samples required to optimise CVaR such that the resulting optimal values lie within ε-relative precision of the true optimal value c_β in (3) with $(1-\alpha)\times 100\%$ confidence.

Theorem 4 Suppose that Assumptions 3 - 4 are satisfied. Then for any $\delta > 0$, we have

$$\frac{n_{\text{SAA}}(\beta)}{n_{\texttt{is}}(\beta)} \geq \frac{c}{\beta^{1-\delta}}, \qquad \text{for all } \beta < \beta_0,$$

for suitable constants $\beta_0 \in (0,1)$ and $c > 0$, if either (i) Assumption 1 holds and the oracle mapping for Algorithm 1 be as in Example 4; or (ii) Assumption 2 holds and the parameter h for the self-structuring IS transformation $\boldsymbol{T}_h(\cdot)$ be selected as in Algorithm 2.

Similar to Theorem 3, Theorem 4 implies that the sample requirement to solve the more challenging CVaR optimization problem using IS also grows only like $o(1/\beta^\delta)$, for any $\delta > 0$, as the tail-level $\beta \to 0$. In particular, there is no efficiency lost in embedding the IS change of distributions using the RA solution paradigm. To get a numerical sense of the savings in sample requirement, consider the portfolio optimization task in Example 1. For $\beta = 1/100$, Deo et al. (2022) demonstrates, for instance, that while SAA takes

about 8000 samples to achieve a 1% relative error, self-structuring IS require only about 550 samples (about 15 times less). Deo et al. (2022) carries an additional analysis on the savings in the work complexity (or the total computational effort) due to the RA procedure in Algorithm 2 relative to SAA.

5 STOCHASTIC APPROXIMATION AS A SOLUTION PARADIGM

As an alternative to RA, one may also consider iterative stochastic approximation methods as a solution paradigm for solving (3). Under mild technical conditions, (3) reduces to the stochastic root finding problem

$$\nabla f(u,\boldsymbol{\theta}) = \mathbf{0} \text{ or equivalently } E[\boldsymbol{G}(\boldsymbol{X};u,\boldsymbol{\theta})] = \mathbf{0} \text{ where } \boldsymbol{G}(\boldsymbol{x};u,\boldsymbol{\theta}) = \frac{\partial F(\boldsymbol{x};u,\boldsymbol{\theta})}{\partial(u,\boldsymbol{\theta})}. \tag{20}$$

A typical approach to solving (20), without any change of distribution, is to use the following iterative scheme (see, for e.g.., Asmussen and Glynn (2007)): Given a sample $\boldsymbol{X}_n$, update the iterate recursively via

$$(u_n,\boldsymbol{\theta}_n) \leftarrow (u_{n-1},\boldsymbol{\theta}_{n-1}) - \gamma_n \boldsymbol{G}(\boldsymbol{X}_n;u_{n-1},\boldsymbol{\theta}_{n-1}) \text{ for } n = 1,2,\ldots$$

The gradients $\boldsymbol{G}(\boldsymbol{x};u,\boldsymbol{\theta}) = (\frac{\partial F}{\partial u}, \frac{\partial F}{\partial \boldsymbol{\theta}})(\boldsymbol{x};u,\boldsymbol{\theta})$ can be readily computed for the objective (3) as in,

$$\frac{\partial}{\partial u}F(\boldsymbol{x};u,\boldsymbol{\theta}) = 1-\beta^{-1}\mathbf{1}(\boldsymbol{\theta}^\top\boldsymbol{x}\geq u) \quad \frac{\partial}{\partial \boldsymbol{\theta}}F(\boldsymbol{x};u,\boldsymbol{\theta}) = \beta^{-1}\frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{x};\boldsymbol{\theta})\,\mathbf{1}(\boldsymbol{\theta}^\top\boldsymbol{x}\geq u).$$

Following the approach devised in He et al. (2023), one may readily embed the IS change of distributions into the update step as shown in Algorithm 3 below. One may also obtain the sample-requirement reduction guarantees similar to that obtained for the retrospective approximation in Theorem 4, thanks to the convergence analysis executed in He et al. (2023).

Algorithm 3: CVaR Optimization using oracle-based adaptive IS

Input: Density $f_{\boldsymbol{X}}(\cdot)$, initial iterate $(u_0,\boldsymbol{\theta}_0)$, initial IS parameter choice α_0, IS oracle $\boldsymbol{b}$, step-size parameters $c>0,\ \gamma\in(1/2,1)$

Initialise $n=1$. **While** stopping criterion not met **do**

1. Generate an independent sample $\boldsymbol{Z}_n$ from the IS density $g_{\alpha_n}(\cdot)$.
2. Set $\gamma_n = cn^{-\gamma}$ and update root estimate

$$(u_n,\boldsymbol{\theta}_n) \leftarrow (u_{n-1},\boldsymbol{\theta}_{n-1}) - \gamma_n \boldsymbol{G}(\boldsymbol{Z}_n;u_{n-1},\boldsymbol{\theta}_{n-1})\frac{f_{\boldsymbol{X}}(\boldsymbol{Z}_n)}{g_{\alpha_n}(\boldsymbol{Z}_n)}$$

3. Update IS parameter $\alpha_{n+1} = \boldsymbol{b}(u_n,\boldsymbol{\theta}_n)$. Set $n = n+1$.

Output: Return the averaged iterate $(\bar{u}_n,\bar{\boldsymbol{\theta}}_n) = n^{-1}\sum_{i=1}^n (u_n,\boldsymbol{\theta}_n)$ as an estimate of the optimal solution to (3).

ACKNOWLEDGEMENTS

The authors acknowledge support from Singapore Ministry of Education grants MOE2019-T2-2-163 and MOE-T2EP20223-0008.

REFERENCES

Arief, M., Y. Bai, W. Ding, S. He, Z. Huang, H. Lam *et al*. 2021. "Certifiable Deep Importance Sampling for Rare-Event Simulation of Black-Box Systems". *arXiv preprint arXiv:2111.02204*.

Asmussen, S. and P. W. Glynn. 2007. *"Stochastic Simulation: Algorithms and Analysis"*. First ed. New York: Springer.

Bai, Y., Z. Huang, H. Lam, and D. Zhao. 2022. "Rare-event Simulation for Neural Network and Random Forest Predictors". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 32(3):1–33.

Bai, Y., Z. Huang, H. Lam, and D. Zhao. 2023. "Over-Conservativeness of Variance-Based Efficiency Criteria and Probabilistic Efficiency in Rare-Event Simulation". *To appear in Management Science*.

Blanchet, J., J. Li, and M. K. Nakayama. 2019. "Rare-event Simulation for Distribution Networks". *Operations Research* 67(5):1383–1396.

Caccioli, F., I. Kondor, and G. Papp. 2018. "Portfolio Optimization Under Expected Shortfall: Contour Maps of Estimation Error". *Quantitative Finance* 18(8):1295–1313.

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

Deo, A. and K. Murthy. 2021. "Efficient Black-Box Importance Sampling for VaR and CVaR Estimation". In *2021 Winter Simulation Conference (WSC)*, 1–12 https://doi.org/10.1109/WSC52266.2021.9715385.

Deo, A. and K. Murthy. 2023. "Achieving Efficiency in Black Box Simulation of Distribution Tails with Self-Structuring Importance Samplers". *To appear in Operations Research*.

Deo, A., K. Murthy, and T. Sarker. 2022. "Combining Retrospective Approximation with Importance Sampling for Optimising Conditional Value at Risk". In *2022 Winter Simulation Conference (WSC)*, 891–902 https://doi.org/10.1109/WSC57314.2022.10015306.

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

He, S., G. Jiang, H. Lam, and M. C. Fu. 2023. "Adaptive Importance Sampling for Efficient Stochastic Root Finding and Quantile Estimation". *To Appear in Operations Research*.

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

Homem-de Mello, T. and G. Bayraksan. 2014. "Monte Carlo Sampling-based Methods for Stochastic Optimization". *Surveys in Operations Research and Management Science* 19(1):56–85.

Juneja, S., R. Karandikar, and P. Shahabuddin. 2007. "Asymptotics and Fast Simulation for Tail Probabilities of Maximum of Sums of Few Random Variables". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 17(2):7–43.

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.

Pasupathy, R. 2010. "On Choosing Parameters in Retrospective-Approximation Algorithms for Stochastic Root Finding and Simulation Optimization". *Operations Research* 58(4):889–901.

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

Shapiro, A. 1991. "Asymptotic Analysis of Stochastic Programs". *Annals of Operations Research* 30:169–186.

Shapiro, A., D. Dentcheva, and A. Ruszczynski. 2021. *Lectures on Stochastic Programming: Modeling and Theory*. 3rd ed. Philadelphia: Society for Industrial and Applied Mathematics.

Uryasev, S. 2000. "Conditional Value-at-Risk: Optimization Algorithms and Applications". In *Proceedings of the IEEE/IAFE/INFORMS 2000 Conference on Computational Intelligence for Financial Engineering (CIFEr) (Cat. No.00TH8520), March 2000, New York City, USA*, 49–57.

AUTHOR BIOGRAPHIES

ANAND DEO is an Assistant Professor at the Indian Institute of Management, Bangalore. His research interests are in Monte-Carlo simulation, stochastic optimisation and extreme value theory. His work focuses on development of methodologies for optimization in the presence of distribution tails. His e-mail address is anand.deo@iimb.ac.in.

KARTHYEK MURTHY is an assistant professor at the Singapore University of Technology and Design. His research interests revolve around stochastic modeling, optimization under uncertainty, and Monte Carlo methods. He focuses on developing robust and efficient data-driven algorithms for decision making under uncertainty, with applications in quantitative risk management, smart grid systems, and traffic flow management problems. His e-mail address is karthyek_murthy@sutd.edu.sg.