Conditional Value-at-risk Constrained Optimization

ANISH SENAPATI, JOSE BLANCHET, FAN ZHANG, BERT ZWART

Abstract. We consider chance constrained optimization problems with heavy-tailed distributions within the risk factors. The usual chance constrained optimization problem with value-at-risk constrains has limitations in modeling and tractability motivating a conditional-value-at-risk (CVaR) risk measure which prevails in many real-world applications. In this project, we aim to look at the generic CVaR constrained optimization problem. We transform the optimization problem into its corresponding Lagrangian relaxation problem and design an algorithm to solve this relaxation minimization using stochastic gradient descent. To find the optimal solution as the chance constraints become tighter, rare event simulation techniques were developed and used to improve our algorithm efficiency in our Monte Carlo evaluation beyond that of a naive Monte Carlo approach. A scaled importance sampling gradient descent method was developed to create a optimization technique that has bounded relative error as the risk constrains become stricter. Numerical results show a significant improvement in this method over the traditional naive optimization descent in both time and algorithm efficiency with constant runtime for scaled importance sampling gradient descent.

1. Background

1.1. Introduction.

Many systems face the challenge of achieving optimal utility while also satisfying risk constraints with high probability. Such problems can be formulated into chance constrined optimizations problems whose objective is to satisfy and solve the following optimization problem as $\delta \to 0$.

$$\begin{aligned} &\text{minimize} && \boldsymbol{c}^T\boldsymbol{x} \\ &\text{subject to} && \text{Prob}\{\phi(\boldsymbol{x},\boldsymbol{\xi}) > 0\} \leq \delta. \end{aligned} \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^m$ is an m-dimensional decision variable, and $\boldsymbol{\xi} \in \mathbb{R}^n$ is an n-dimensional random vector. The elements of $\boldsymbol{\xi}$ are often referred to as risk factors; the function $\phi : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is often assumed to be convex in $\boldsymbol{x}$ and often models a cost constraint; the parameter $\delta > 0$ is the risk level of the tolerance.

Chance constrained optimization formulations of problems have a wide array of applications in modeling and decision making in a large number of settings. In power networks and energy management, the applications of chance constrained optimization was reviewed in [Bienstock et al., 2012]. The work of [Bonami and Lejeune, 2009] demonstrates using chance constrained optimization formulation in the context of portfolio selection. More recently, chance constrained optimizations have become a pivotal part in robotics and autonomous vehicles used extensively in model Predictive Control (MPC) frameworks [Vitus and Tomlin, 2013, Lenz et al., 2015]. The wide range of applications for chance constrained formulations make it an pivotal field of research to benefit from in the coming future.

Date: August 6, 2022.

The probabilistic constraint in (1) can be reformulated using a risk measure called value-at-risk (VaR). The VaR at level $\alpha \in (0,1)$ for a loss random variable X is defined as

$$\mathrm{VaR}_\alpha\{X\} = \min\{z \in \mathbb{R} : F_X(z) \geq \alpha\},$$

where $F_X : \mathbb{R} \to [0,1]$ is the cumulative distribution function of X. Consequently, the constraint in problem (1) is equivalent to: $\mathrm{VaR}_{1-\delta}\{\phi(\boldsymbol{x},\boldsymbol{\xi})\} \leq 0$ resulting in the optimization problem:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & \mathrm{VaR}_{1-\delta}\{\phi(\boldsymbol{x},\boldsymbol{\xi})\} \leq 0. \end{array} \tag{2}$$

There has been a significant research on finding the solutions of such chance constrained optimization problems/optimization problems with VaR constraints. However, these problems have been proven to be intractable and NP-hard in the worst case [Luedtke and Ahmed, 2008]. In this paper, we will concentrate on a convex approximation of (1) which does not possess the same limitations as the VaR constraints.

1.2. **Optimization problem with conditional value-at-risk.**

In addition to the intractability of (2), VaR constraints have other modeling limitations which make it a less appealing risk measure:

- VaR does not include on scenarios exceeding $\mathrm{VaR}_\alpha\{X\}$.
- VaR fails to meet the subadditivity axiom, so it is not "coherent". "coherent" is a desirable properties for risk measures proposed by [Artzner et al., 1999].
- Evaluation of $\mathrm{VaR}_{1-\delta}\{\phi(\boldsymbol{x},\boldsymbol{\xi})\}$ often involves integration over $\boldsymbol{\xi}$, which is computationally intractable.
- For fixed δ, the mapping $\boldsymbol{x} \mapsto \mathrm{VaR}_{1-\delta}\{\phi(\boldsymbol{x},\boldsymbol{\xi})\}$ is usually non-convex increasing the difficulty of optimization.

For these reasons, a second risk measure called conditional value at risk (CVar) is motivated for such risk-constrained optimization problems. Introduced by [Rockafellar and Uryasev, 2000], CVaR is a risk measure used in many real world applications, such as portfolio management [Rockafellar and Uryasev, 2000], supply chain management [Heckmann et al., 2015], and power system analysis [Morales et al., 2010, Conejo et al., 2010].

The CVaR at level $\alpha \in (0,1)$ for a loss random variable X is defined as

$$\mathrm{CVaR}_\alpha\{X\} = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\beta\{X\}d\beta.$$

We concentrate on the following CVaR constrained optimization problem:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & \mathrm{CVaR}_{1-\gamma}\{\phi(\boldsymbol{x},\boldsymbol{\xi})\} \leq 0. \end{array} \tag{3}$$

Initially introduced by [Nemirovski and Shapiro, 2007], they showed CVaR is a tight convex approximation of VaR and (3) is a convex approximation of (2). Thus, the CVaR constrained optimization problem accounts for many of the limitations VaR constraints possess making it an appealing approximation to optimize.

1.3. **Properties of Conditional Value-at-risk.**
We now review properties of CVaR which are useful in analyzing (3). First, CVaR calculations can be alternatively represented using the following lemma:

Lemma 1.1 (Alternative representation)**.** *Let X be a random variable with cumulative distribution function $F_X : \mathbb{R} \to [0,1]$. For $\alpha \in (0,1)$, define X^α be random variable with cumulative distribution function $F_{X^\alpha} : \mathbb{R} \to [0,1]$ defined as*

$$F_{X^\alpha}(z) = \begin{cases} 0 & \text{if } z < \mathrm{VaR}_\alpha\{X\}, \\ \frac{F_X(z)-\alpha}{1-\alpha} & \text{if } z \geq \mathrm{VaR}_\alpha\{X\}. \end{cases}$$

Then we have $\mathrm{CVaR}_\alpha\{X\} = \mathbb{E}[X^\alpha]$. *In particular, if X has density, then*

$$\mathrm{CVaR}_\alpha\{X\} = \mathbb{E}[X|X \geq \mathrm{VaR}_\alpha\{X\}] = \mathbb{E}[X|X > \mathrm{VaR}_\alpha\{X\}].$$

This theorem proven by [Rockafellar and Uryasev, 2002] is widely used to reformulate CVaR optimization problems:

Theorem 1.2 (Fundamental minimization formula)**.** *For $\alpha \in (0,1)$, define $h_\alpha : \mathbb{R} \to \mathbb{R}$ as*

$$h_\alpha(z) = z + \frac{1}{1-\alpha}\mathbb{E}\left[(X-z)^+\right], \qquad \text{where } (t)^+ = \max(0,t).$$

Then we have h_α is finite and convex (hence continuous), and

$$\mathrm{CVaR}_\alpha\{X\} = \min_{z\in\mathbb{R}} h_\alpha(z), \qquad \mathrm{VaR}_\alpha\{X\} = \min\{z \in \mathbb{R} : h_\alpha(z) = \mathrm{CVaR}_\alpha\{X\}\}.$$

The convexity of CVaR can also be proven [Pflug, 2000].

1.4. **Goals.**
Assuming that there is a regularization parameter $\lambda = \lambda(\delta)$ which can create an equivalence (same optimal solution) between the Lagrangian relaxation problem and the original optimization problem (3), then we need to optimize the problem:

$$\underset{x}{\text{minimize}} \quad \boldsymbol{c}^T\boldsymbol{x} + \lambda \cdot \mathrm{CVaR}_{1-\delta}\{\phi(\boldsymbol{x},\boldsymbol{\xi})\} \tag{4}$$

We can apply theorem 1.2 (setting $\gamma = 1-\alpha$) to reduce the (4) to

$$\underset{x,z}{\text{minimize}} \quad \boldsymbol{c}^T\boldsymbol{x} + \lambda z + \lambda\gamma^{-1} \cdot \mathbb{E}\left[(\phi(\boldsymbol{x},\boldsymbol{\xi}) - z)^+\right]. \tag{5}$$

Our goal in this paper is to solve optimization problem (5) efficiently as $\gamma \to 0$ or $\alpha \to 1$. The solution to the optimization problem is unbounded as the limit is approached ($z \to \infty$) so rare event simulation techniques must be used to improve the efficiency of the minimization problem. Specifically, we seek to improve the efficiency of the evaluation of $(\phi(\boldsymbol{x},\boldsymbol{\xi}) - z)^+$ for heavy-tailed $\boldsymbol{\xi}$ using rare event simulation techniques. Within this paper, we explore two different expressions for $\phi(\boldsymbol{x},\boldsymbol{\xi})$ and find rare event techniques to optimize (5) in both cases for heavy tailed random vectors $\boldsymbol{\xi}$.

2. Problem Formulation

In this section, we will describe the two different forms of $\phi(\boldsymbol{x}, \boldsymbol{\xi})$ that were optimized for and provide a scaling argument which transforms the two respective problems into

2.1. **Scaling Reformulation.**

While the original optimization problem can be solved, the inefficiencies of finding the solution arise from the fact that $\boldsymbol{x}$ and $\boldsymbol{z}$ approach infinity as $\gamma \to 0$. Intuitively, there is a critical scaling that can be applied to the solution of $\boldsymbol{x}$ and $\boldsymbol{z}$ which can be applied to simplify the calculations in the optimizations. Since we are considering heavy tailed distributions, we expect that $\mathbb{E}[z > t] \approx ct^{-a} \approx \gamma$. This subsequently gives the estimation that $x \approx O(\gamma^{-1/a})$ and $z \approx O(\gamma^{-1/a})$. So, applying the transformations $x = \frac{\bar{x}}{\gamma^{1/a}}$ and $z = \frac{\bar{z}}{\gamma^{1/a}}$ and minimizing with respect to $\bar{x}$ and $\bar{z}$ gives a scaled optimization problem which can help ease the divergence of x and z in the solutions. This technique is applied in the next sections.

2.2. **Maximization Problem.**

In this case, we can set $\phi(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{i=1}^{d}(\boldsymbol{\xi}_i - \boldsymbol{x}_i)$. Additionally, for simplicity, we will set $\boldsymbol{c}$ to 1 in (5). In this case, we can rewrite our problem as

$$\min_{z, \boldsymbol{x} \geq 0} \boldsymbol{x}^T \mathbf{1} + \lambda \mathbb{E}\left[\frac{\max(\boldsymbol{\xi}_i - \boldsymbol{x}_i) - z)^+}{1 - \alpha} + z\right]$$

for fixed parameters α, λ. We will assume that $\boldsymbol{\xi}$ is drawn from a regularly varying distribution with index a. Specifically, a Pareto distribution with fixed shape parameter a is used to draw from $\boldsymbol{\xi}$. We can approximate the maximum in the problem with the logsum max function, a smoothed version of the max function, $\delta - \max_{i=1}^{d} a_i = \delta \log\left(\sum_{j=1}^{d} \exp\left(\frac{a_i}{\delta}\right)\right)$. This smoothing will give the maximum function much more appealing properties for the optimization procedure that follow. After applying the scaling argument, our problem reduces to:

$$\max_{\bar{z}, \bar{\boldsymbol{x}}} \bar{\boldsymbol{x}}^T 1 + \lambda \mathbb{E}\left[\frac{(\delta - \max_{i=1}^{d}(\gamma^{\frac{1}{a}} \boldsymbol{\xi}_i - \bar{\boldsymbol{x}}) - \bar{z})^+}{\gamma} + \bar{z}\right] \tag{6}$$

2.3. **Salvage Fund Problem.**

In this case, let $L = (L_1, L_2, ..., L_d)$ be drawn from a regularly varying distribution (i.i.d Pareto for instance). We can qualify L_i as the total incurred loss that entity i is responsible to pay. Let $Q = (Q_{i,j} : i, j \in \{1, ..., d\})$ be a deterministic matrix where $Q_{i,j}$ denotes the amount of money received by entity j when entity i pays one dollar. As defined in the salvage fund problem in [Blanchet et al., 2020], $\phi(\boldsymbol{x}, \boldsymbol{L})$ is defined to be the linear programming problem of

$$\phi(\boldsymbol{x}, \boldsymbol{L}) = \begin{array}{ll} \text{minimize}_{b,y} & \boldsymbol{b} - m \\ \text{subject to} & (\boldsymbol{L} - \boldsymbol{y}) \leq \boldsymbol{b} * \mathbf{1}, (\boldsymbol{I} - \boldsymbol{Q}^T)\boldsymbol{y} \leq \boldsymbol{x}, \boldsymbol{y} \geq 0 \end{array} \tag{7}$$

We define bankruptcy for an entity if its deficit is greater than a constant m. Then, the problem of

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}} & 1^T \boldsymbol{x} \\ \text{subject to} & P(\phi(\boldsymbol{x}, \boldsymbol{L}) \geq 0) \leq \delta \end{array}$$

is equivalent to finding the minimum amount of salvage fund such that no bankruptcy occurs for all entity with probability $1-\delta$.

Applying the CVaR dual representation and Lagrangian relaxation, we then reduce this problem to the familiar form:

$$\text{minimize}_{x,z} \quad \mathbf{1}^T\boldsymbol{x} + \lambda\left[z + \tfrac{1}{\delta}\mathbb{E}[(\phi(\boldsymbol{x},\boldsymbol{L}) - z)^+]\right] \tag{8}$$

Applying the scaling arguments of $x = \frac{\bar{x}}{\delta^{1/a}}$ and $z = \frac{\bar{z}}{\delta^{1/a}}$, we then arrive at

$$\begin{aligned} &\text{minimize} \quad \mathbf{1}^T\tfrac{\bar{x}}{\delta^{1/a}} + \lambda\left[\tfrac{\bar{z}}{\delta^{1/a}} + \tfrac{1}{\delta}\mathbb{E}[(\phi(\tfrac{\bar{x}}{\delta^{1/a}},\boldsymbol{L}) - \tfrac{\bar{\boldsymbol{z}}}{\delta^{1/a}})^+]\right] \\ &\text{minimize} \quad \mathbf{1}^T\bar{x} + \lambda\left[\bar{z} + \tfrac{1}{\delta}\mathbb{E}[(\delta^{1/a}\phi(\tfrac{\bar{x}}{\delta^{1/a}},\boldsymbol{L}) - \bar{z})^+]\right] \end{aligned} \tag{9}$$

where a factor of $\delta^{1/a}$ is factored out in the last expression.

3. Algorithms for Optimization Problem

3.1. **Maximization Problem Optimization.**

With the scaled optimization problem in 6, we can apply stochastic gradient descent to minimize this function by grabbing samples of $\boldsymbol{\xi}_i$ from a Pareto distribution with scale parameter a and going in the opposite direction of the gradients for that iterations $\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}$. The gradients of the function can be calculated to be

$$D_{\bar{\boldsymbol{x}}}g(\bar{\boldsymbol{x}},\bar{z}) = 1 + \lambda\mathbb{E}\left[\frac{I(\delta - \max_{i=1}^d(\gamma^{\frac{1}{a}}\boldsymbol{\xi}_i - \bar{\boldsymbol{x}}) > \bar{\boldsymbol{z}}) * D_{\bar{\boldsymbol{x}}}\delta - \max_{i=1}^d(\gamma^{\frac{1}{a}}\boldsymbol{\xi}_i - \bar{\boldsymbol{x}}_i)}{\gamma}\right] \tag{10}$$

$$D_{\bar{z}}g(\bar{\boldsymbol{x}},\bar{z}) = 1 - \frac{\mathbb{E}\left[I(\delta - \max_{i=1}^d(\gamma^{\frac{1}{a}}\boldsymbol{\xi}_i - \bar{\boldsymbol{x}}) > \bar{\boldsymbol{z}})\right]}{\gamma} \tag{11}$$

where $g(\bar{\boldsymbol{x}},\bar{z})$ is the function we are minimizing in 6. The derivation of these derivatives comes from applying the chain rule to our original function.

While a simple SGD algorithm with a crude Monte Carlo (MC) sampling of $\boldsymbol{\xi}$ from a Pareto distribution will eventually converge to the correct minimum, the rate of convergence as $\gamma \to 0$ is slow as the indicator function in the gradients is 0 more often as $\gamma \to 0$, resulting in less correct movement per gradient. Figure 1 shows values of the function $g(\bar{\boldsymbol{x}},\bar{z})$ for $\bar{\boldsymbol{x}},\bar{z}$ through iterations of a simple MC SGD over time for a 1-dimensional Pareto function (d=1) with scale parameter $\alpha = 3$. For a 1-dimensional Pareto function, the optimal solution of the maximization problem for a fixed α with $\lambda < 1$ is $\theta = 0$ and $\beta = VaR_\alpha(X)$, a value is known for a Pareto distribution. This gives the expected minimum values in the dotted lines in Figures 1 and 2. As $\gamma \to 0 \implies \alpha \to 1$, the failure of convergence over 300 iterations is evident for a crude MC sampling method. Additionally, for the SGD to even reach values of $\boldsymbol{\xi}, z$ that are within 5% of the expected value requires more iterations as $\alpha \to 1$.

Thus, we need to sample $\boldsymbol{\xi}_i$ in more efficient way, namely such that $\mathbb{E}\left[I(\delta - \max_{i=1}^d(\gamma^{\frac{1}{a}}\boldsymbol{\xi}_i - \bar{\boldsymbol{x}}) > \bar{\boldsymbol{z}})\right]$ can be calculated with bounded relative error as $\gamma \to 0$. This requirement can be satisfied through

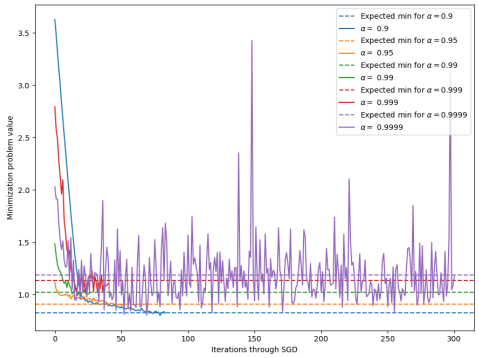

FIGURE 1. A graph of the value of $g(\bar{\boldsymbol{x}}, \bar{z})$ when drawing $\boldsymbol{\xi}_i$ from the Pareto distribution using a crude MC scheme along with the expected minimum values using $\lambda = .8, \delta = .1$. The failure of convergence as $\alpha \to 1$ is evident, especially at $\alpha = .9999$.

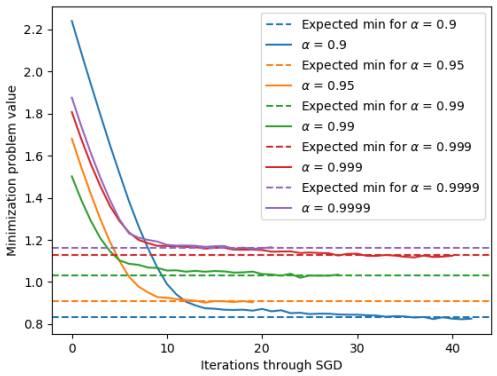

FIGURE 2. A graph of the value of $g(\bar{\boldsymbol{x}}, \bar{z})$ when drawing $\boldsymbol{\xi}_i$ from the importance distribution described along with the expected minimum values using $\lambda = .8, \delta = .1$. The constant amount of time of convergence as $\alpha \to 1$ is evident along with the smooth convergence

importance sampling, a sampling method to sample from another distribution which more consistently draws from the area of importance in the probability distribution.

As a review of importance sampling, consider calculating an expectation of function $f(x)$, where $x \sim p(x)$, subjected to some distribution. We have the following estimation of $\mathbb{E}(f(x))$:

$$\mathbb{E}(f(x)) = \int f(x)p(x)dx \approx \frac{1}{n}\sum_i f(x_i)$$

where x_i. is drawn from the probability distribution x_i. However, we can also sample x_i from another distribution $q(x)$ and calculate $\mathbb{E}(f(x))$ using this transformation:

$$\mathbb{E}[f(x)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x) \approx \frac{1}{n}\sum_i f(x_i)\frac{p(x_i)}{q(x_i)}$$

where x_i is now drawn from q(x). The expression $Z = \frac{p(x_i)}{q(x_i)}$ is the likelihood ratio. Properly tuning $q(x)$ can lead to a reduction in the variance of our expectation measurement and leads to more accurate results of such calulations.

In our case, we want to draw samples such that $\max(\gamma^{1/a}\boldsymbol{\xi}_i - \bar{\boldsymbol{x}}_i) > \bar{z} \implies \boldsymbol{\xi}_i > \frac{\bar{z}+\bar{\boldsymbol{x}}_i}{\gamma^{1/a}}$ for at least one i to satisfy our indicator function. We can bound this event by the sum of the events $X_i > \frac{\beta+\theta_i}{\gamma^{1/a}}$ which can individually be sampled from using inverse transform sampling since the CDF of the Pareto distribution is used. Thus, we can use a mixture of individual probability distributions to sample in the area of events we desire. Specifically, consider the probability distribution where

with probability $\frac{P(\boldsymbol{\xi}_i > \frac{\bar{z}+\bar{\boldsymbol{x}}_i}{\gamma^{1/a}})}{\sum_{i=1}^{d} P(\boldsymbol{\xi}_i > \frac{\bar{z}+\bar{\boldsymbol{x}}_i}{\gamma^{1/a}})}$, we sample with the assumption of $X_i > \frac{\beta+\theta_i}{\gamma^{1/a}}$ (done with inverse sampling) with all other components drawn normally from the Pareto distribution. In order to apply importance sampling, the likelihood factor for a sample X would be set to $Z = \frac{\sum_{i=1}^{d} P(X_i > \frac{\beta+\theta_i}{\gamma^{1/a}})}{\sum_{i=1}^{d} I(X_i > \frac{\beta+\theta_i}{\gamma^{1/a}})}$. Drawing samples (X, Z) from this probability distribution will assure that the indicator function in the gradients is necessarily nonzero, as desired. It can also be shown that such a probability distribution has bounded relative error as $\gamma \to 0$.

Figure 2 shows the value of the minimization problem using the importance sampling technique described above with the same parameters as in Figure 1. Comparing the two graphs, the convergence with importance sampling is signifcantly smoother since the SGD does not take many incorrect jumps in the wrong direction by pulling out many useless samples. Additionally, using importance sampling takes approximately a constant number of iterations (40 iterations) for the differing values of α which is what we expect for a sampling method with bounded relative error.

3.2. **Salvage Fund Problem Optimization.**

With the same idea as the maximization problem, we hope to perform a SGD algorithm to optimize 9. We compute the derivatives of the objective function $g(\bar{x}, \bar{z})$:

$$D_{\bar{x}} g(\bar{x}, \bar{z}) = 1 + \frac{\lambda}{\delta} \mathbb{E}\left[I\left(\delta^{1/a} \phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{\xi}) > \bar{z} \right) * D_{\bar{x}} (\delta^{1/a} \phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{\xi})) \right] \tag{12}$$

$$D_{\bar{z}} g(\bar{x}, \bar{z}) = \lambda - \frac{\lambda}{\delta} \mathbb{E}\left[I(\delta^{1/a} \phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{\xi})) > \bar{z}) \right] \tag{13}$$

So, we wish to draw ξ with an importance sampling method that chooses $\boldsymbol{\xi}$ such that $\mathbb{E}[I(\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{\xi}) > \frac{\bar{z}}{\delta^{1/a}})]$ has bounded relative error as $\delta \to 0$. Since $\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{\xi})$ is a minimization linear programming problem, it is evident that any feasible solution to the constraints given in $\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{\xi})$ would be an upper bound to the function itself. Thus, a necessary condition for the inequality $\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{\xi}) > \frac{\bar{z}}{\delta^{1/a}}$ is that any feasible solution to the linear programming problem is also greater than $\frac{\bar{z}}{\delta^{1/a}}$.

We can construct an importance sampling method based off this necessary condition. Specifically, it is easy to see that $\boldsymbol{b} = \max \xi_i$ and $y = 0$ will always be a feasible solution to the constraints of $\phi(\boldsymbol{x}, \boldsymbol{\xi})$ with a value of $\max \xi_i - m$. So, we would want to sample such that $\max \boldsymbol{\xi}_i - m \geq \frac{\bar{z}}{\delta^{1/a}}$ to ensure all samples satisfy the necessary condition.

Similar to the maximization problem, we can bound the necessary condition by the union of the events $\xi_i \geq \frac{\bar{z}}{\delta^{1/a}} + m$ which can individually be sampled from using inverse transform sampling since the CDF of the Pareto distribution is known. Thus, we can use a mixture of individual probability distributions to sample in the area of events we desire. Specifically, consider the probability distribution where with probability $\frac{P(\xi_i > \frac{\bar{z}}{\delta^{1/a}})}{\sum_{i=1}^{d} P(\xi_i > \frac{\bar{z}}{\delta^{1/a}})}$, we sample with the assumption of $\xi_i > \frac{\bar{z}}{\delta^{1/a}}$ (done with inverse sampling on the Pareto distribution) with all other components drawn normally from the Pareto distribution. In this case, the likelihood parameter is $Z = \frac{\sum_{i=1}^{d} P(\xi_i > \frac{\bar{z}}{\delta^{1/a}})}{\sum_{i=1}^{d} I(\xi_i > \frac{\bar{z}}{\delta^{1/a}})}$. Drawing samples (X, Z) from this probability distribution will assure that the necessary conditions for the indicator

function in our gradients would be satisfied increasing the likelihoods of useful samples. It can also be shown that such a probability distribution has bounded relative error as $\gamma \to 0$.

Along with the importance sampling method, a method needed to be developed to estimate $D_{\bar{x}}(\delta^{1/a}\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{L}))$. This can be approximated by the subgradient of $\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{L})$ which can be derived from the dual problem of $\phi(x, \boldsymbol{L})$. Specifically, the dual problem of $\phi(x, \boldsymbol{\xi})$ is:

$$\text{dual-}\phi(\boldsymbol{x}, \boldsymbol{L}) = \begin{array}{ll} \text{maximize}_{\boldsymbol{\kappa},\boldsymbol{\beta}} & \boldsymbol{\kappa}^T\boldsymbol{L} - \boldsymbol{\beta}^T\boldsymbol{x} - m \\ \text{subject to} & (\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{\beta} - \boldsymbol{\kappa} \geq 0, \boldsymbol{\kappa}^T\mathbf{1} = 1, \boldsymbol{\kappa}, \boldsymbol{\beta} \geq 0 \end{array} \tag{14}$$

It can be shown that the subgradient with respect to x of $\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{L}))$ is $\frac{\kappa^*}{\delta^{1/a}}$ where κ^* is the optimal value of κ in dual-$\phi(\frac{\bar{x}}{\delta^{1/a}}, \boldsymbol{L})$ [Bertsimas and Tsitsiklis, 1997]. Thus, solving the linear optimization dual problem would give us an approximation for this derivative completing our expressions for the gradients and allow a SGD algorithm to be made with the importance sampling.

4. Numerical Results

The importance sampling SGD method described for the salvage fund problem was coded up in Python using NumPy and SciPy to draw from the Pareto distribution as described. The linear optimization problem of $\phi(\boldsymbol{x}, \boldsymbol{L})$ and dual-$\phi(\boldsymbol{x}, \boldsymbol{L})$ were coded up and solved using CVXPY. The SGD was run starting x and z to be $\mathbf{1}$ and was set to converge when $|x(n) - x(n-1)| \leq .01$ and $|z(n) - z(n-1)| \leq .005$ where n is the iteration number of the SGD. Figure 3 shows the convergence of the SGD to the minimal value of the objective function for different values of δ. For this graph, λ was set to .8, a was set to 3, and Q was defined to be a $10 * 10$ matrix where $Q_{ij} = 0$ if $i == j$ and 1/10 otherwise. It is important to note that the values of the objective function that are plotted in the graph below used the importance samples that were used in that iteration. Since our goal is the efficiency of our SGD algorithm, a small number of samples were used per iteration (5000) which sometimes resulted in a variance in the objective function value. However, once the values of x and z stabilized, 30000 importance samples were drawn to get a more accurate representation of the minimal objective function. These values are recorded in Table 4.

δ	Objective function value
.2	3.3914
.1	5.9812
.01	12.577
.001	27.813
.0001	62.938
1e-5	139.377

It is immediately obvious through the graph that the importance sampling scaled SGD convergence takes a constant number of iterations to converge. In order to gain an idea of the advantage of the importance sampling SGD, an experiment was run to compare it to an unscaled naive SGD (without importance sampling).

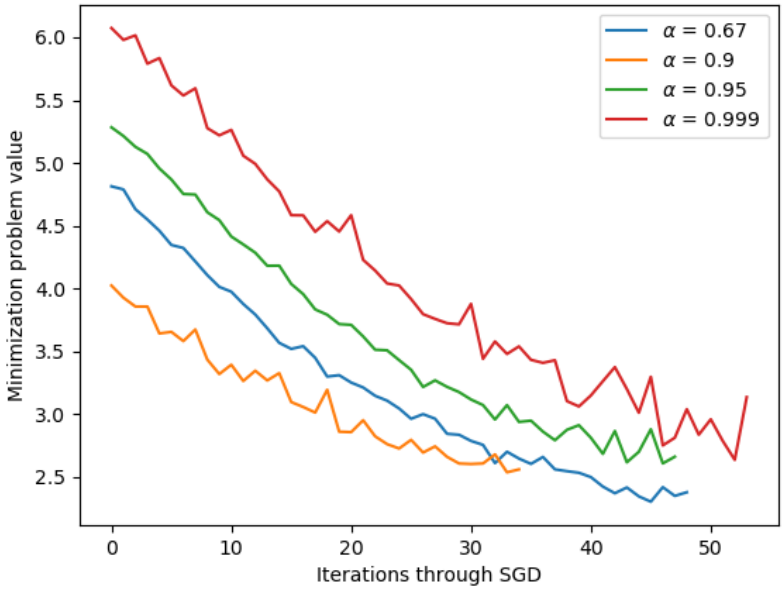

FIGURE 3. A convergence of the importance smapling SGD algorithm for the salvage fund with varying values of α where $\alpha = 1 - \delta$. For all the values of alpha, the algorithm managed to converge well within 40 iterations.

Both algorithms were run until the problems reach within 5% minimum of the original value and the number of iterations and time per iteration were recorded for both methods. The results are shown in table.

δ	Samples per Iteration	Iterations (Importance)	Time per Iterations (Importance)	Iterations (Naive)	Time per Iterations (Naive)
10^{-2}	2000	38	36.2 sec	2	49.3 sec
10^{-3}	4000	33	82.3 sec	156	79.4 sec
10^{-4}	7500	32	100.6 sec	3681	94.5 sec
10^{-5}	15000	21	243.7 sec	22421	231.7 sec

The samples needed per iteration was increased in order for the naive SGD method to properly converge. Convergence was defined as when the Pareto average of SGD iterations was found to be within 5% of the minimum value for that specific δ. The constant number of iterations in the convergence of the importance sampling scaled SGD method is evident from the table unlike the naive unscaled method which appears to have a exponential level of growth in the iterations needed as $\delta \to 0$. Further, both SGD methods take similar levels of time per iteration indicating no disadvantage in efficiency in using the importance sampling scaled method. It is also important to note that though the importance sampling and scaled method were performed using the same number of iterations for a specific δ, the importance sampling method could have been performed for different delta using considerably less iterations such as in Figure 3 where all SGD runs were done using 4000 samples per iteration. This numerical results are promising in efficiently optimizing the CVaR minimization problem since it presents an algorithm which requires a constant number

of iterations and samples per iterations due to the bounded relative error method developed with the importance sampling.

Acknowledgements

I would like to thank Dr. Jose Blanchet, Fan Zhang, and Dr. Bert Zwart for supporting me through this project.

References

[Artzner et al., 1999] Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3):203–228.

[Bertsimas and Tsitsiklis, 1997] Bertsimas, D. and Tsitsiklis, J. (1997). *Introduction to Linear Optimization*. Athena Scientific, 1st edition.

[Bienstock et al., 2012] Bienstock, D., Chertkov, M., and Harnett, S. (2012). Chance constrained optimal power flow: Risk-aware network control under uncertainty. *SIAM Review*, 56.

[Blanchet et al., 2020] Blanchet, J., Zhang, F., and Zwart, B. (2020). Optimal scenario generation for heavy-tailed chance constrained optimization. *arXiv preprint arXiv:2002.02149*.

[Bonami and Lejeune, 2009] Bonami, P. and Lejeune, M. A. (2009). An exact solution approach for portfolio optimization problems under stochastic and integer constraints. *Operations Research*, 57(3):650–670.

[Conejo et al., 2010] Conejo, A. J., Carrión, M., Morales, J. M., et al. (2010). *Decision making under uncertainty in electricity markets*, volume 1. Springer.

[Heckmann et al., 2015] Heckmann, I., Comes, T., and Nickel, S. (2015). A critical review on supply chain risk–definition, measure and modeling. *Omega*, 52:119–132.

[Lenz et al., 2015] Lenz, D., Kessler, T., and Knoll, A. (2015). Stochastic model predictive controller with chance constraints for comfortable and safe driving behavior of autonomous vehicles. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 292–297.

[Luedtke and Ahmed, 2008] Luedtke, J. and Ahmed, S. (2008). A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699.

[Morales et al., 2010] Morales, J. M., Conejo, A. J., and Pérez-Ruiz, J. (2010). Short-term trading for a wind power producer. *IEEE Transactions on Power Systems*, 25(1):554–564.

[Nemirovski and Shapiro, 2007] Nemirovski, A. and Shapiro, A. (2007). Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996.

[Pflug, 2000] Pflug, G. C. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic constrained optimization*, pages 272–281. Springer.

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

[Rockafellar and Uryasev, 2002] Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471.

[Vitus and Tomlin, 2013] Vitus, M. P. and Tomlin, C. J. (2013). A probabilistic approach to planning and control in autonomous urban driving. In *52nd IEEE Conference on Decision and Control*, pages 2459–2464.