Towards Scalable Risk Analysis for Stochastic Systems
Using Extreme Value Theory

Evan Arsenault†, Yuheng Wang†, and Margaret P. Chapman,† IEEE Member

Abstract—We aim to analyze the behaviour of a finite-time stochastic system, whose model is not available, in the context of more rare and harmful outcomes. Standard estimators are not effective in predicting the magnitude of such outcomes due to their rarity. Instead, we leverage Extreme Value Theory (EVT), the study of the long-term behaviour of normalized maxima of random variables. In this letter, we quantify risk using the upper-semideviation \( E(\max(Y - E(Y), 0)) \), which is the expected exceedance of a random variable \( Y \) above its mean \( E(Y) \). To assess more rare and harmful outcomes, we propose an EVT-based estimator for this functional in a given fraction of the worst cases. We show that our estimator enjoys a closed-form representation in terms of a popular functional called the conditional value-at-risk. In experiments, we illustrate the extrapolation power of our estimator when a small number of i.i.d. samples is available (\(<50\)). Our approach is useful for estimating the risk of finite-time systems when models are inaccessible and data collection is expensive. The numerical complexity does not grow with the size of the state space. This letter initiates a broader pathway for estimating the risk of large-scale systems.

Index Terms—Risk-aware stochastic systems, Extreme Value Theory, Model-free risk estimation and analysis

I. INTRODUCTION

The study of risk-aware systems has been growing in popularity due to its nuanced interpretations and assessments of uncertainties [1]–[10]. A risk-aware system is a dynamical system whose performance or safety is assessed in a manner that is sensitive to the possibility and severity of rare harmful outcomes. Special cases of risk-aware performance or safety criteria include:

- An expected cost \( E(Y) \), where \( Y \) is a random variable in which smaller realizations are preferred (i.e., a random cost), and
- A maximum deterministic cost \( \sup \{ c_i : i \in K \} \), where \( c_i \) is a real number indexed by \( i \in K \).

However, an expected cost is not designed to quantify more rare and harmful outcomes. Such outcomes are critical to assess when systems operate under uncertainties and have safety goals. A maximum deterministic cost may be too conservative for some applications and is not suitable when \( c_i \) is unbounded. Fortunately, one can define criteria that quantify a variety of other characteristics of a random cost \( Y \), including:

- A linear combination of the mean and variance,
- A probability of a harmful event \( \{ Y \geq \eta \} \), where \( \eta \in \mathbb{R} \) is a given threshold,
- An expected utility \( E(\nu(Y)) \), where \( \nu \) represents a decision-maker’s subjective preferences,
- A quantile at a particular probability level, and
- An expected exceedance above a given threshold, \( E(\max\{Y - \eta, 0\}) \).

A popular criterion is the conditional value-at-risk (CVaR), which represents the expectation in a given fraction of the worst cases. The CVaR has been used to define performance or safety objectives [1], [6], [9] and constraints [3], [5] for control systems. The expressive nature of risk-aware criteria suggests broad applicability to systems operating under uncertainties in practice.

However, current approaches to analyze and optimize the behaviours of risk-aware non-linear systems suffer from scalability challenges. Most approaches are based on dynamic programming (DP) [1], [2], [7], [9] or Q-learning [4], [8]. These algorithms were developed originally to minimize an expected cumulative cost for a Markov decision process (MDP). When reformulating one of the above algorithms for a risk-aware setting, the new algorithm inherits the scalability issues of the original algorithm. It is well-known that DP cannot apply to high-dimensional state spaces without function approximations. Q-learning typically involves extensive exploration, finite state spaces, and infinite time horizons [4], [8]. These conditions need not apply when sampling is moderately expensive, the state space is continuous, or analysis on a finite time horizon is needed. A recent approach for scalable risk analysis combines temporal difference learning with value function approximation [10]. The approach does not require extensive exploration but involves a finite-state infinite-time MDP, and risk is quantified using a composition of functionals [10], which may be difficult to interpret. Similar to [10], we focus on the analysis of risk rather than its optimization. However, we adopt statistical tools from Extreme Value Theory (EVT).

EVT is the study of the long-term behaviour of normalized maxima of random variables. It has been useful for examining extreme events in numerous applications, including hydrology [11], seismology [12], and disease transmission [13]. This theory offers tools to extrapolate beyond the available data to estimate the upper tail of a distribution [14], [15]. Theoretical connections between EVT and hitting time statistics for discrete-time systems have been established [16]. EVT has been applied to compute properties of chaotic systems, including the dimensions of an invariant measure [17].

1E.A., Y.W., and M.P.C. are with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering (ECE), University of Toronto, 10 King’s College Road, Toronto, Ontario M5S 3G8 Canada. Contact email: evan.arsenault@mail.utoronto.ca.
E.A., Y.W., and M.P.C. are grateful for support provided by the Edward S. Rogers Sr. Department of ECE, University of Toronto.
Statistical applications of EVT include the analysis or estimation of extreme quantiles [18], CVaR [19–23], and extreme probabilities [14, Ch. 4.4]. A distributionally robust approach for extreme quantile estimation that is sensitive to modeling errors has been developed [18]. An EVT-based formula for CVaR [19], [20] has been applied to inform the selection of controls in a multi-armed bandit problem [22]. We will find this formula useful to estimate a different risk functional. (A risk functional is a map from a space of random variables to \( \mathbb{R} \).) Deo and Murthy have estimated the CVaR and an associated gradient by combining tools from EVT and importance sampling [23]. In a classification problem, EVT has been used to estimate the probability of a sample belonging to an undetected class [24].

Contributions. We study fundamentals from EVT to estimate the risk of a random cost \( Y \), whose distribution is not known. We consider a risk functional called the upper-semideviation \( \rho(Y) := \mathbb{E}(\max\{Y - \mathbb{E}(Y), 0\}) \). \( \rho(Y) \) enjoys the intuitive interpretation of being the expected exceedance of \( Y \) above its mean. The mean-upper-semideviation \( \mathbb{E}(Y) + \lambda \rho(Y) \) with \( \lambda \geq 0 \) is a standard risk functional [10] [25, Ex. 6.20]. We consider \( \rho(Y) \) because the mean can be estimated readily. To evaluate more rare and harmful outcomes, we propose an EVT-based estimator for the upper-semideviation in a fraction of the worst cases. Our theoretical contribution is to prove that the estimator admits a closed-form representation in terms of the CVaR (Theorem 3). The proof requires the construction of a random variable \( Y_0 \) that satisfies useful properties (Lemmas 2, 4). We construct \( Y_0 \) so that its distribution resembles the distribution of \( Y \) in the upper tail under assumptions from Extreme Value Theory. In experiments, we demonstrate the ability of our estimator to perform well when the quantity of data is limited (Section IV, Figure 1). We also showcase our approach using data of total overflow volumes from combined sewer systems throughout Canada [26].

Notation. \( \mathbb{N} = \{0, 1, 2, \ldots \} \) is the set of natural numbers, \( \mathbb{R} \) is the set of real numbers, and \( \mathbb{R}_+ := (0, +\infty) \). Let (\( \Omega, F, P \)) be a probability space. The notation \( Y \sim L^1(\Omega, F, P) \), or \( Y \in L^1 \) for brevity, means that \( Y \) is a random variable on (\( \Omega, F, P \)) such that \( \mathbb{E}(Y) := \int_{\Omega} |Y| \, dP < +\infty \). The function \( 1_A : \Omega \to \{0, 1\} \) is the indicator on \( A \in \mathcal{F} \), i.e., \( 1_A(\omega) = 1 \) if \( \omega \in A \), \( 1_A(\omega) = 0 \) otherwise. If \( F \) is a distribution function, \( z^* := \sup\{z \in \mathbb{R} : F(z) < 1\} \) is the right endpoint of \( F \). For an \( \mathbb{R} \)-valued function \( f \), we define \( \tilde{f} := 1 - f \). If \( S \) is a metric space, \( B_S \) is the Borel sigma algebra on \( S \), and \( \text{int}(S) \) is the interior of \( S \). We use the following abbreviations: i.i.d. = independent and identically distributed, a.e. = almost everywhere or almost every, and w.r.t. = with respect to.

II. PRELIMINARIES

Since EVT is not well-known in control theory, it is necessary to summarize some fundamentals, which we adopt from [14] and [15]. Let \( (Z_i)_{i \in \mathbb{N}} \) be an i.i.d. sequence of random variables defined on a probability space (\( \Omega, F, P \)) with distribution function \( F \). We do not know \( F \) or \( P \). The partial maximum for \( m \in \mathbb{N} \) is defined by \( M_m := \max\{Z_1, Z_2, \ldots, Z_m\} \). Since \( M_m \) converges in probability to the right endpoint \( z^* \) of \( F \), a normalization of \( M_m \) can be useful for revealing characteristics of \( F \).

Definition 1 (\( F \in \mathcal{D}(G_{\gamma}) \)): Suppose that there exist sequences \( (a_m)_{m \in \mathbb{N}} \subseteq \mathbb{R}_+ \) and \( (b_m)_{m \in \mathbb{N}} \subseteq \mathbb{R} \) such that

\[
\lim_{m \to \infty} P \left( \left\{ \omega \in \Omega : \frac{M_m(\omega) - b_m}{a_m} \leq z \right\} \right) = G_\gamma(z)
\]

for every continuity point \( z \) of \( G_\gamma \), where \( G_\gamma \) is a non-degenerate distribution function. Then, we say that \( F \) belongs to the maximum domain of attraction of \( G_\gamma \), i.e., \( F \in \mathcal{D}(G_\gamma) \).

\( G_\gamma \) being non-degenerate means that it does not correspond to a point mass. There are many examples of \( F \) that satisfy Definition 1, including: Pareto, Burr, Fréchet, t-Student, Cauchy, Log-gamma, Uniform on (0,1), Beta, Exponential, Logistic, Gumbel, Normal, Lognormal, and Gamma [15]. The extreme value index \( \gamma \) is a qualitative measure for tail “heaviness,” i.e., how fast the tail of \( F \), provided that \( F \in \mathcal{D}(G_\gamma) \), decays to zero [15, p. 63].

The next theorem provides an equivalent characterization for \( F \in \mathcal{D}(G_\gamma) \). This characterization offers a useful approximation for the upper tail of \( F \). For convenience, we define the interval \( \mathcal{J}_\gamma \) by

\[
\mathcal{J}_\gamma := \begin{cases} 
(0, +\infty), & \text{if } \gamma \geq 0, \\
(0, -1/\gamma), & \text{if } \gamma < 0,
\end{cases}
\]

and the function \( \phi_\gamma : \mathcal{J}_\gamma \to (0, 1] \) by

\[
\phi_\gamma(z) := \begin{cases} 
(1 + \gamma z)^{-1/\gamma}, & \text{if } \gamma \neq 0, \\
\exp(-z), & \text{if } \gamma = 0.
\end{cases}
\]

\( \phi_\gamma := 1 - \phi_\gamma \) corresponds to the Generalized Pareto distribution [15, Eq. (4.6)].

Theorem 1: [14, Th. 1.1.6, Part 4]: \( F \in \mathcal{D}(G_\gamma) \) for some \( \gamma \in \mathbb{R} \) if and only if there is an \( \mathbb{R}_+ \)-valued function \( g \) s.t.

\[
\lim_{s \uparrow z^*} \frac{1 - F(s + xg(s))}{1 - F(s)} = \phi_\gamma(x), \quad x \in \mathcal{J}_\gamma,
\]

where \( s \uparrow z^* \) means that \( s \) approaches \( z^* \) from below.

For brevity, we use the notations \( g_s := g(s) \) and \( \tilde{F} := 1 - F \). Motivated by (4), the following heuristic is commonly used, e.g., see [14, pp. 65–66], to approximate the upper tail of \( F \) above some sufficiently large threshold \( s \in \mathbb{R} \):

\[
\tilde{F}(z) \approx \tilde{F}(s) \cdot \phi_\gamma((z - s)/g_s), \quad z \geq s.
\]

The subsequent lemma formalizes the heuristic \( 5 \) and uses the following definitions. Assuming that \( F \in \mathcal{D}(G_\gamma) \) for some \( \gamma \in \mathbb{R} \), let a parameter vector \( \theta \) be defined by

\[
\theta := \{k, m, \gamma, s, g_s\},
\]

where \( k \in \mathbb{N} \) and \( m \in \mathbb{N} \) with \( k < m \), \( \gamma \in \mathbb{R} \) is an extreme value index, \( s \in \mathbb{R} \) is a threshold, and \( g_s > 0 \) (\( g \) comes from Theorem 1). In addition, we define the interval \( \mathcal{I}_\theta \) by

\[
\mathcal{I}_\theta := \begin{cases} 
(s, +\infty), & \text{if } \gamma \geq 0, \\
[s, s - g_s/\gamma), & \text{if } \gamma < 0.
\end{cases}
\]

Lemma 1 (Formal tail approximation): Suppose that \( F \in \mathcal{D}(G_\gamma) \) for some \( \gamma \in \mathbb{R} \) and \( F(z) < 1 \) for all \( z \in (-\infty, z^*) \).
Let $\epsilon > 0$ and $z \in \mathcal{I}_\theta$ be given. If $z^* \in \mathbb{R}$, then there exists a $\delta_{z,z} > 0$ such that
\[
-\epsilon \tilde{F}(s) < \tilde{F}(z) - \tilde{F}(s) \cdot \phi_Y((z-s)/g_s) < \epsilon \tilde{F}(s)
\]
for all $s \in (-\delta_{z,z} + z^*, z^*)$. Otherwise, if $z^* = +\infty$, then $\exists r_{z,z} \in \mathbb{R}$ such that (8) holds for all $s \in (r_{z,z}, z^*)$.

Lemma 1 follows from applying the definition of the limit \[27, Def. 4.33, p. 98\] to (4), whose proof is omitted in the interest of space. Lemma 1 indicates how a threshold $\delta_{\alpha,Y}$ at level $\alpha$ is typically denoted by $\epsilon_{\alpha,Y}$ (25, p. 258). However, we do not know its distribution function.

### III. Estimating extremal upper-semideviation

Let $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ be a random cost. We wish to evaluate the more rare and harmful realizations of $Y$, i.e., assess the risk of $Y$. However, we do not know its distribution function $F_Y(y) := P(\{\omega \in \Omega : Y(\omega) \leq y\}), \ y \in \mathbb{R}$. (9)

Suppose that $Y$ is a random cost of a control system, which operates on a discrete time horizon of length $N \in \mathbb{N}$. Then, it is common to define the sample space $\Omega := (S \times A)^N \times S$, where $S$ is the state space and $A$ is the control space, which are metric spaces. Here, an element $\omega$ of $\Omega$ is a possible trajectory $\omega = (x_0, u_0, \ldots, x_{N-1}, u_{N-1}, x_N)$ of the system. The sigma algebra $\mathcal{F}$ of interest is $B_\mathcal{F}$. The probability measure $\mathbb{P}$ of interest is typically denoted by $P^\pi$ because it depends on an initial condition $x \in S$ and a control policy $\pi$.

While we keep the probability space generic in this letter, i.e., $(\Omega, \mathcal{F}, \mathbb{P})$, our analysis applies to the evaluation of a control system’s behaviour, where $P^\pi$ is not known.

There are different ways to assess the risk of $Y$. One approach is to use the mean-upper-semideviation functional,
\[
\varphi(Y) := \mu + \lambda \rho(Y), \quad \lambda \geq 0,
\]
where $\mu := E(Y) := \int_\Omega Y \mathrm{d}\mathbb{P}$ is the mean of $Y$, and
\[
\rho(Y) := \int_\Omega \max\{Y(\omega) - \mu, 0\} \mathrm{d}\mathbb{P}(\omega)
\]
is the upper-semideviation of $Y$. $\varphi$ has the intuitive interpretation of being a weighted sum of the mean $\mu$ and the expected exceedance above the mean $\rho(Y)$. If the domain of $\varphi$ is $L^1$ and $\lambda \in [0, 1]$, then $\varphi$ is coherent \[25, Ex. 6.20\]. That is, $\varphi$ satisfies four desirable properties: it is convex, monotonic, sub-additive, and positively homogeneous.

We aim to estimate the extremal upper-semideviation of $Y \in L^1$ at level $\alpha \in (0, 1)$, which we define by
\[
\rho_{\alpha,Y}(Y) := \int_{\{\omega \in \Omega : Y(\omega) \geq v_{\alpha,Y}\}} \max\{Y(\omega) - \mu, 0\} \mathrm{d}\mathbb{P}(\omega),
\]
where $v_{\alpha,Y} \in \mathbb{R}$ is a threshold to be described. Specifically, $v_{\alpha,Y}$ is the value-at-risk of $Y$ at level $\alpha$.

\[
v_{\alpha,Y} := \inf\{z \in \mathbb{R} : F_Y(z) \geq 1 - \alpha\}, \quad \alpha \in (0, 1),
\]
a common risk functional in finance and operations research \[25\]. In statistics, $v_{\alpha,Y}$ is called the left-side $(1 - \alpha)$-quantile of $Y$. The quantity $\rho_{\alpha,Y}(Y)$ is the expected exceedance of $Y$ above the mean in a fraction $\alpha$ of the worst cases. We focus on estimating $\rho_{\alpha,Y}(Y)$ because the former corresponds to rarer and larger realizations of $Y$. Accurately estimating these realizations is a difficult, yet important, task for safety-critical applications.

To estimate $\rho_{\alpha,Y}(Y)$, we will apply properties of the conditional value-at-risk (CVaR). The CVaR of $Y \in L^1$ at level $\alpha \in (0, 1)$ is defined by [25, Eq. (6.22)]:
\[
c_{\alpha,Y} := \inf_{\tau \in \mathbb{R}} \{\tau + \frac{1}{\alpha} E(\max\{Y - \tau, 0\})\}.
\]
It is coherent \[25, p. 261\]. A useful fact is that a minimizer of the right-hand side of (14) is $v_{\alpha,Y}$ \[25, p. 258\], and therefore,
\[
c_{\alpha,Y} = v_{\alpha,Y} + \frac{1}{\alpha} E(\max\{Y - v_{\alpha,Y}, 0\}).
\]

We will use a data set $\{y_{i,m} : i = 1, 2, \ldots, m \} \subset \mathbb{R}$ of size $m \in \mathbb{N}$ that is sampled independently from $Y$ and satisfies
\[
y_{1,m} \leq \cdots \leq y_{m-k,m} \leq \cdots \leq y_{m-1,m} \leq y_{m,m}
\]
with $k \in \mathbb{N}$ and $k < m$. The notation $\mu_m := \frac{1}{m} \sum_{i=1}^m y_{i,m}$ denotes the sample mean of the data (16) to estimate $\rho_{\alpha,Y}(Y)$.

The notation $y_{i,m}$ denotes a realization of $Y_{i,m}$, i.e., $y_{i,m} = Y_{i,m}(\omega)$ for some $\omega \in \Omega$. A data set $\{y_{i,m} : i = 1, 2, \ldots, m \}$ corresponds to some $\omega \in \Omega$ and satisfies (16).

#### A. Typical empirical estimator for $\rho_{\alpha,Y}(Y)$

A typical estimator for $\rho_{\alpha,Y}(Y)$ is computed using the $k + 1$ largest samples for some large enough $k \in \mathbb{N}$:
\[
\hat{\rho}_{\alpha,k,m} := \frac{1}{m-k} \sum_{i=m-k}^m \max\{y_{i,m} - \mu_m, 0\},
\]
where $y_{m-k,m}$ is an approximation for $v_{\alpha,Y}$ \[13\]. $\hat{\rho}_{\alpha,k,m}$ is not designed to represent the upper tail of $F_Y$ when the number $m$ of samples is limited; we will illustrate limitations of $\hat{\rho}_{\alpha,k,m}$ numerically in Section IV.

We aim to estimate $\rho_{\alpha,Y}(Y)$ using a data set (16) in a manner that is sensitive to the burden of collecting numerous samples and the challenge of observing large realizations of $Y$ in practice.

#### B. EVT-based estimator for $\rho_{\alpha,Y}(Y)$

Here, we propose an EVT-based estimator for $\rho_{\alpha,Y}(Y)$ that enjoys a closed-form representation in terms of the CVaR. Our tactic is to construct a random variable $Y_\theta$, whose extremal distribution approximates the extremal distribution of $Y$.

We suppose that an experiment has been conducted, providing a data set (16). We let $\theta$ be a parameter vector, which has been estimated using this data set. There are different ways to estimate $\theta$, and we will illustrate one way in Section IV.

Motivated by the analysis of Lemma 1, we define
\[
F_\theta(z) := \begin{cases} 
0, & \text{if } z < s, \\
1 - \frac{1 - k}{m} \phi_Y\left(\frac{z-s}{g_s}\right), & \text{if } z \in \mathcal{I}_\theta, \\
1, & \text{if } z \geq s - g_s/\gamma, \gamma < 0,
\end{cases}
\]
where $s$ is the mean-upper-semideviation of $Y$, $\phi_Y$ is the expected exceedance above the mean, and $\gamma$ is the tail index of the distribution function of $Y$. The choice $\gamma < 0$ ensures that $F_\theta(z)$ is a valid distribution function.
with $s := y_{m-k,m}$. The factor $\frac{k}{m}$ is an estimate for $\hat{F}_Y(s) = 1 - F_Y(s)$, where the empirical distribution $\hat{F}_m(x) := \frac{1}{m} \sum_{i=1}^m 1_{\{Y_i, m \leq x\}}$ is used because $F_Y$ is not known. $\hat{F}_\theta$ (19) satisfies key properties, as described below.

**Lemma 2 (F_\theta is a distribution function):** For any $\theta \in (0,1)$, $F_\theta : \mathbb{R} \to [0,1]$ (19) is non-decreasing and right-continuous with limit $F_\theta(z) = 0$ and $F_\theta(z) = 1$.

We omit the proof in the interest of brevity. As a result of Lemma 2, $F_\theta$ is the distribution function of some random variable, which we denote by $Y_\theta$ [28, p. 209]. Using a canonical construction, we define $Y_\theta$ on the probability space $(\mathbb{R}, \mathcal{B}_\mathbb{R}, P_\theta)$, where $P_\theta$ is the Lebesgue-Stieltjes measure corresponding to $F_\theta$ and $Y_\theta(z) := z$ for all $z \in \mathbb{R}$ [28, p. 209].

By Lemma 1 under some assumptions on $F_Y$, in particular, $F_Y \in \mathcal{D}(G_\theta)$ for some $\gamma \in \mathbb{R}$, the approximation $F_\theta \approx F_Y$ is valid on $\mathcal{I}_\theta$ (7). Then, $Y_\theta$ and $Y$ have similar distributions in the upper tails. Next, we provide a sufficient condition for $Y_\theta \in L^1$, which will be useful for evaluating the CVaR of $Y_\theta$.

**Lemma 3 (Y_\theta \in L^1):** Let $Y_\theta$ be a random variable with distribution function $F_\theta$ (19). If $\gamma < 1$, then $Y_\theta \in L^1(\mathbb{R}, \mathcal{B}_\mathbb{R}, P_\theta)$.

**Proof:** It suffices to show that $\int_{-\infty}^\infty |Y_\theta(z)| \, dP_\theta(z) < +\infty$. Take $b \in \mathbb{R}$ such that $b < s$, and define $Y_\theta := Y_\theta - b$. Therefore,

$$F_\theta(\{z \in \mathbb{R} : Y_\theta(z) > x\}) = 1 - F_\theta(x + b), \quad x \in \mathbb{R}. \quad (20)$$

Since $F_\theta(b) = 0$ by (19), $Y_\theta$ is positive a.e. w.r.t. $P_\theta$. By a generalized tail integral formula [29, Prop. 6.24], we have that

$$\int_{\mathbb{R}} |Y_\theta(z)| \, dP_\theta(z) = \int_0^{+\infty} 1 - F_\theta(x + b) \, dx. \quad (21)$$

Since $\gamma < 1$, the integral in the right-hand side of (21) is finite by lengthy but standard calculus calculations. Then, we apply Minkowski’s Inequality [29, p. 183] to $Y_\theta = Y_\theta + b$ to conclude that $\int_{\mathbb{R}} |Y_\theta(z)| \, dP_\theta(z) \leq \int_{\mathbb{R}} |Y_\theta(z)| \, dP_\theta(z) + |b| < +\infty$. −

Given a data set $\{y_{i,m} : i = 1,2,\ldots,m\}$ (16) be given. Suppose that $\theta \in (0,1)$ has been estimated from the data with $s := y_{m-k,m}$, and let $0 < \alpha < \frac{k}{m} < 1$. If $v_{a,Y_\theta} \geq \mu_m$, then $\theta \equiv \alpha = \alpha(c_{a,Y_\theta} - \mu_m)$.

**Proof:** By Lemma 3 we have that $Y_\theta \in L^1(\mathbb{R}, \mathcal{B}_\mathbb{R}, P_\theta)$.

Thus, the CVaR of $Y_\theta$ at level $\alpha \in (0,1)$ satisfies

$$c_{a,Y_\theta} = v_{a,Y_\theta} + \frac{1}{\alpha} \int_{\mathbb{R}} \max\{z - v_{a,Y_\theta}, 0\} \, dP_\theta(z) \quad (25)$$

by utilizing the argument underlying (15). For brevity, we use the notation $(x)^+ := \max\{x,0\}$ for all $x \in \mathbb{R}$. By applying the definition of $A_{a,\theta}$ (23) and the assumption $v_{a,Y_\theta} \geq \mu_m$, we have that

$$1_{A_{a,\theta}}(z) \cdot (z - \mu_m)^+ = 1_{A_{a,\theta}}(z) \cdot (z - v_{a,Y_\theta})^+ + 1_{A_{a,\theta}}(z) \cdot (v_{a,Y_\theta} - \mu_m). \quad (26)$$

for all $z \in \mathbb{R}$. By re-expressing $\hat{\rho}_{a,\theta}$ (22) and noting that a sum of non-negative Borel-measurable functions can be integrated term by term [28, Cor. 1.6.4], it holds that

$$\hat{\rho}_{a,\theta} = \int_{\mathbb{R}} 1_{A_{a,\theta}}(z) \cdot (z - \mu_m)^+ \, dP_\theta(z) = \psi_1 + \psi_2, \quad (27)$$

$$\psi_1 := \int_{\mathbb{R}} 1_{A_{a,\theta}}(z) \cdot (z - v_{a,Y_\theta})^+ \, dP_\theta(z), \quad (28)$$

Since $A_{a,\theta} = [v_{a,Y_\theta}, +\infty)$ (23), we have that

$$1_{A_{a,\theta}}(z) \cdot (z - v_{a,Y_\theta})^+ = (z - v_{a,Y_\theta})^+, \quad z \in \mathbb{R}, \quad (29)$$

and therefore, $\psi_1$ (27) simplifies as follows:

$$\psi_1 = \int_{\mathbb{R}} (z - v_{a,Y_\theta})^+ \, dP_\theta(z). \quad (30)$$

Next, we simplify $\psi_2$ (28). Using $A_{a,\theta} = [v_{a,Y_\theta}, +\infty)$, $F_\theta$ being a distribution function (Lemma 2), and $F_\theta$ being the corresponding Lebesgue-Stieltjes measure, we have that

$$P_\theta(A_{a,\theta}) = 1 - \lim_{z \uparrow v_{a,Y_\theta}} F_\theta(z) \quad (31)$$

by [28, 1.4.5 (9), p. 25]. Since $F_\theta$ (19) is continuous on $\text{int}(\mathcal{I}_\theta)$, $v_{a,Y_\theta} \in \text{int}(\mathcal{I}_\theta)$, and $F_\theta(v_{a,Y_\theta}) = 1 - \alpha$ (Lemma 4), we find that

$$\lim_{z \uparrow v_{a,Y_\theta}} F_\theta(z) = F_\theta(v_{a,Y_\theta}) = 1 - \alpha. \quad (32)$$
Using (31–32), we simplify $\psi_{2,\theta}$ (28) as follows:

$$\psi_{2,\theta} = (v_{\alpha, Y} - \mu_m) \cdot (1 - (1 - \alpha)) = \alpha \cdot (v_{\alpha, Y} - \mu_m).$$  \hspace{1cm} (33)

Combining (30) and (33), we conclude that

$$\hat{\rho}_{\alpha, \theta} = \alpha \left( \frac{1}{\alpha} \int_{\mathbb{R}} (z - v_{\alpha, Y})^+ \ dP_0(z) + v_{\alpha, Y} - \mu_m \right) = \alpha (c_{\alpha, Y} - \mu_m),$$  \hspace{1cm} (34)

where we use (25) in the final line.

**Remark 2 (Theorem 2 assumptions):** Typically, $\alpha$ is small, e.g., less than 0.05, to emphasize rare high-consequence outcomes. Thus, we anticipate the value-at-risk of $\hat{\rho}$ at level $\alpha$ to exceed the sample mean, i.e., $v_{\alpha, Y} \geq \mu_m$, in applications. We will present an example in Section IV-B.

The next remark uses techniques from [19], [20], [25] to provide analytical expressions for $v_{\alpha, Y}$ and $c_{\alpha, Y}$ (25).

**Remark 3 (Expressions for $v_{\alpha, Y}$ and $c_{\alpha, Y}$):** Assume the conditions of Theorem 2. Since $0 < \alpha, \frac{k}{m} < 1$, it holds that $v_{\alpha, Y} \in \text{int}(\mathcal{I}_0$) by Lemma 4. Since $F_\theta$ (19) is continuous and strictly increasing on $\text{int}(\mathcal{I}_0)$, we invert $F_\theta$ on $\text{int}(\mathcal{I}_0)$ to derive the following expression for $v_{\alpha, Y}$:

$$v_{\alpha, Y} = \begin{cases} s + g_s \left( \left( \frac{m - \alpha}{k} \right)^{-1} - 1 \right), & \text{if } \gamma \neq 0, \\ s - g_s \cdot \log \left( \frac{m - \alpha}{k} \right), & \text{if } \gamma = 0. \end{cases}$$  \hspace{1cm} (35)

As well as $\alpha \in (0, 1)$, we have that $Y_\theta \in L^1(\mathbb{R}, B_\mathbb{R}, P_\theta)$ by Lemma 3. The CVaR can be expressed as an average of the value-at-risk, $c_{\alpha, Y} = \frac{1}{\alpha} \int_{\text{int}(\mathcal{I}_0)} v_{1-\gamma, Y} d\tau$ [25, Th. 6.2]. We use this expression with $v_{1-\gamma, Y}$ as per (35) and the assumption $\gamma < 1$ from Theorem 2 to derive

$$c_{\alpha, Y} = (v_{\alpha, Y} + g_s - \gamma \cdot s)(1 - \gamma)^{-1}.$$  \hspace{1cm} (36)

**Theorem 2 and Remark 3** together provide a closed-form expression for $\hat{\rho}_{\alpha, \theta}$ (22). This expression is useful for our numerical experiments, which we present subsequently.

**IV. NUMERICAL EXPERIMENTS**

We consider two experiments regarding the estimation of $\rho(Y)$ (12). First, we compare the performance of a typical estimator $\hat{\rho}_{\alpha, k, m}$ (18) to our EVT-based estimator $\hat{\rho}_{\alpha, \theta}$ (22) using six benchmark distributions. The second experiment uses data of combined sewer overflows in Canada from 2013–2017 using six benchmark distributions. The numerical experiments, which we present subsequently.

**A. Benchmark distributions**

We compare the performance of the estimators $\hat{\rho}_{\alpha, k, m}$ (18) and $\hat{\rho}_{\alpha, \theta}$ (22) for $m \in \{20, 21, 22, \ldots, 99\}$ samples. For each $m$, we have conducted 10,000 runs of the following: 1) draw $m$ i.i.d. samples from a random variable $Y$; 2) estimate $\theta$ as described previously; and 3) compute the errors $\hat{\rho}_{\alpha, k, m} - \hat{\rho}_{\alpha, \theta}$ and $\hat{\rho}_{\alpha, k, m} - \hat{\rho}_{\alpha, \theta}$. We approximate the ground-truth value of $\rho_{\alpha, \theta}(Y)$ using a Monte Carlo simulation with over 4 million samples.

We consider six distributions for $Y$: Pareto(2), t-Student with 5 degrees of freedom (d.o.f.), Exponential(1), Gumbel, Uniform(0,1), and Beta(1,2). The Pareto(2) and t-Student distributions have $\gamma > 0$; the Exponential(1) and Gumbel distributions have $\gamma = 0$; the Uniform(0,1) and Beta(1,2) distributions have $\gamma < 0$ [15]. Figure 1 presents the results. The EVT-based estimator $\hat{\rho}_{\alpha, \theta}$ has lower average error compared to the typical estimator $\hat{\rho}_{\alpha, k, m}$ for smaller values of $m$. The error decreases as $m$ increases, and for sufficiently large $m$, the typical estimator is superior. These results support the use of the EVT-based estimator for risk analysis when a small number of samples is available (which is our focus). We illustrate one application next.

**B. System example**

Monthly combined sewer overflow volumes in Canada during the years of 2013–2017 are available from [26]. Each data point is the total volume of water (a mixture of stormwater and untreated wastewater) that overflowed into the environment during a particular month. The combined sewer network that underlies this data set is a vast hard-to-model dynamical system with safety and performance requirements to protect environmental and public health. We can view the data for a fixed month as being i.i.d., but this would provide only $m = 5$, as there are 5 years of data. Hence, for the purpose of illustration, we consider the data during the spring and early summer (March–June), providing $m = 20$.

We have repeated the previous procedures to compute $\hat{\rho}_{\alpha, k, m} = 0.59 \times 10^6$ m$^3$ and $\hat{\rho}_{\alpha, \theta} = 0.23 \times 10^6$ m$^3$, where $k = 2$. We verify some assumptions from Theorem 2:

$$v_{\alpha, Y} = 26 \times 10^6 \text{ m}^3 > \mu_m = 15 \times 10^6 \text{ m}^3 \text{ and } \alpha = 0.01 < \frac{2}{m} = \frac{k}{m}.$$  \hspace{1cm} (37)

We find that $\gamma = 0.87$, suggesting a heavy-tailed distribution. Interestingly, $\hat{\rho}_{\alpha, k, m} > \hat{\rho}_{\alpha, \theta}$, which also occurs for smaller values of $m$ in the previous experiment (Figure 1). Hence, the typical estimator may over-approximate the risk compared to the EVT-based estimator.

**V. CONCLUSIONS**

We have shown the ability of a new EVT-based estimator for a risk functional to perform well when data is limited. We have developed our estimator from measure-theoretic first principles, initiating a pathway for broader EVT-based risk analysis. We are deriving a consistency proof by applying theory from extreme quantile estimation [14, Th. 4.3.1]. We are excited about exploring automated selection of a threshold $s$, e.g., [31], and developing controllers that are sensitive to more rare and harmful outcomes when data sets are sparse.

1 Combined sewers are present in older cities throughout North America.
Fig. 1. The plots depict the errors, $\hat{\rho}_{0.01,k,m} - \rho_{0.01}(Y)$ (blue) and $\hat{\rho}_{0.01,\theta} - \rho_{0.01}(Y)$ (orange), versus the number of samples $m$. $\hat{\rho}_{0.01,k,m}$ is a typical estimator, and $\hat{\rho}_{0.01,\theta}$ is our EVT-based estimator. The average error across 10,000 trials is plotted using the solid lines, and the coloured bands represent 50% confidence intervals. The EVT-based estimator outperforms the typical estimator for smaller values of $m$.

ACKNOWLEDGMENT

M.P.C. gratefully acknowledges Mr. Changrui Liu for insightful discussions.

REFERENCES

[1] N. Bäuerle and J. Ott, “Markov decision processes with Average-Value-at-Risk criteria,” Math. Methods Oper. Res., vol. 74, no. 3, pp. 361–379, 2011.
[2] N. Bäuerle and U. Rieder, “More risk-sensitive Markov decision processes,” Math. Oper. Res., vol. 39, no. 1, pp. 105–120, 2014.
[3] V. Borkar and R. Jain, “Risk-constrained Markov decision processes,” IEEE Trans. Autom. Control, vol. 59, no. 9, pp. 2574–2579, 2014.
[4] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, “Risk-sensitive reinforcement learning,” Neural Comput., vol. 26, no. 7, pp. 1298–1328, 2014.
[5] B. P. van Parys, D. Kuhn, P. J. Goulart, and M. Morari, “Distributionally robust control of constrained stochastic systems,” IEEE Trans. Autom. Control, vol. 61, no. 2, pp. 430–442, 2015.
[6] C. W. Miller and I. Yang, “Optimal control of Conditional Value-at-Risk in continuous time,” SIAM J. Control Optim., vol. 55, no. 2, pp. 856–884, 2017.
[7] N. Bäuerle and A. Glauener, “Minimizing spectral risk measures applied to Markov decision processes,” Math. Methods Oper. Res., pp. 1–35, 2021.
[8] W. Huang and W. B. Haskett, “Stochastic approximation for risk-aware Markov decision processes,” IEEE Trans. Autom. Control, vol. 66, no. 3, pp. 1314–1320, 2021.
[9] M. P. Chapman, R. Bonalli, K. M. Smith, I. Yang, M. Pavone, and C. J. Tomlin, “Risk-sensitive safety analysis using Conditional Value-at-Risk,” IEEE Trans. Autom. Control, in press, 2021.
[10] U. Köse and A. Ruszczyński, “Risk-averse learning by temporal difference methods with Markov risk measures,” J. Mach. Learn. Res., vol. 22, pp. 1–34, 2021.
[11] E. Towler, B. Rajagopalan, E. Gilleland, R. S. Summers, D. Yates, and R. W. Katz, “Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory,” Water Resour. Res., vol. 46, no. 11, 2010.
[12] R. Shcherbakov, J. Zhuang, G. Zoller, and Y. Ogata, “Forecasting the magnitude of the largest expected earthquake,” Nature Commun., vol. 10, no. 1, pp. 1–11, 2019.
[13] J. T. Lim, Y. Han, B. Sue Lee Dickens, L. C. Ng, and A. R. Cook, “Time varying methods to infer extremes in dengue transmission dynamics,” PLoS Comput. Biol., vol. 16, no. 10, p. e1008279, 2020.
[14] L. de Haan and A. Ferreira, Extreme Value Theory: An Introduction. New York, NY, USA: Springer, 2006, vol. 21.
[15] J. F. Alves and C. Neves, Extreme Events in Finance: A Handbook of Extreme Value Theory and its Applications, Chapter 4, F. Longin, Ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017.
[16] A. C. M. Freitas, J. M. Freitas, and M. Todd, “Hitting time statistics and extreme value theory,” Probab. Theory Relat. Fields, vol. 147, no. 3, pp. 675–710, 2010.
[17] T. Caby, “Extreme value theory for dynamical systems, with applications in climate and neuroscience,” Ph.D. dissertation, Université de Toulon; Università degli studi dell’Insubria (Como, Italie), 2020.
[18] J. Blanchet, F. He, and K. Murthy, “On distributionally robust extreme value analysis,” Extremes, vol. 23, no. 2, pp. 317–347, 2020.
[19] A. J. McNeil, “Extreme value theory for risk managers,” Departement Mathematik ETH Zentrum, vol. 12, no. 5, pp. 217–37, 1999.
[20] M. Gilli and E. Kellezi, “An application of extreme value theory for measuring financial risk,” Comput. Econ., vol. 27, no. 2, pp. 207–228, 2006.
[21] C. Martins-Filho, F. Yao, and M. Torero, “Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory,” Econom. Theory, vol. 34, no. 1, pp. 23–67, 2018.
[22] D. Troop, F. Godin, and J. Y. Yu, “Risk-averse action selection using Extreme Value Theory estimates of the CVaR,” arXiv preprint arXiv:1912.01718, 2019.
[23] A. Deo and K. Murthy, “Optimizing tail risks using an importance sampling based extrapolation for heavy-tailed objectives,” in Proc. Conf. Decis. Control. IEEE, 2020, pp. 1070–1077.
[24] X. Yu, Z. Zhao, X. Zhang, Q. Zhang, Y. Liu, C. Sun, and X. Chen, “Deep-learning-based open set fault diagnosis by extreme value theory,” IEEE Trans. Ind. Inform., vol. 18, no. 1, pp. 185–196, 2021.
[25] A. Shapiro, D. Dentcheva, and A. Ruszczyński, Lectures on Stochastic Programming: Modeling and Theory. Philadelphia, PA, USA: SIAM, 2009.
[26] “Monthly combined sewer overflow volumes, Canada,” https://www150.statcan.gc.ca/n1/daily-quotidien/190625/cg-c003-eng.htm, Statistics Canada, 2019, accessed: March 10, 2022.
[27] W. Rudin, Principles of Mathematical Analysis. New York, NY, USA: McGraw Hill, 1964, vol. 3.
[28] W. Rudin, Real and Complex Analysis. New York, NY, USA: Academic Press, Inc., 1972.
[29] G. B. Folland, Real Analysis: Modern Techniques and Their Applications, 2nd ed. New York, NY, USA: John Wiley & Sons, 1999, vol. 40.
[30] A. W. van der Vaart, Asymptotic Statistics. Cambridge, U.K.: Cambridge University Press, 1998.
[31] B. Bader, J. Yan, and X. Zhang, “Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate,” Ann. Appl. Stat., vol. 12, no. 1, pp. 310–329, 2018.