Learning with risk-averse feedback under potentially heavy tails

Matthew J. Holland
Institute of Scientific and Industrial Research
Osaka University

El Mehdi Haress
CentraleSupélec

Abstract

We study learning algorithms that seek to minimize the conditional value-at-risk (CVaR), when all the learner knows is that the losses (and gradients) incurred may be heavy-tailed. We begin by studying a general-purpose estimator of CVaR for potentially heavy-tailed random variables, which is easy to implement in practice, and requires nothing more than finite variance and a distribution function that does not change too fast or slow around just the quantile of interest. With this estimator in hand, we then derive a new learning algorithm which robustly chooses among candidates produced by stochastic gradient-driven sub-processes, obtain excess CVaR bounds, and finally complement the theory with a regression application.

1 INTRODUCTION

In machine learning problems, since we only have access to limited information about the underlying data-generating phenomena or goal of interest, there is significant uncertainty inherent in the learning task. As a result, any meaningful performance guarantee for a learning procedure can only be stated with some degree of confidence (e.g., a high probability "good performance" event), usually with respect to the random draw of the data used for training. Assuming some loss $L(w;z) \geq 0$ depending on parameter $w \in \mathcal{W} \subseteq \mathbb{R}^d$ and data realization $z \in \mathcal{Z}$, given random data distributed as $Z \sim \mathrm{P}$, the *de facto* standard performance metric in machine learning is the *risk*, or expected loss, defined

$$R(w) := \mathbf{E}_{\mathrm{P}}\, L(w;Z) = \int_{\mathcal{Z}} L(w;z)\, \mathrm{P}(\mathrm{d}z), \qquad w \in \mathcal{W}. \tag{1}$$

The vast majority of research done on machine learning algorithms provides performance guarantees stated in terms of the risk (Haussler, 1992; Devroye et al., 1996; Anthony and Bartlett, 1999). This risk-centric paradigm goes beyond the theory and reaches into the typical workflow of any machine learning practitioner, since "off-sample performance" is typically evaluated by using the average loss on a separate set of "test data," an empirical counterpart to the risk studied in theory. While the risk is convenient in terms of probabilistic analysis, it is merely one of countless possible descriptors of the distribution of $L(w;Z)$. When using a learning algorithm designed to minimize the risk, one makes an implicit value judgement about how the learner should be penalized for "typical" mistakes versus "atypical" but egregious errors.

As machine learning techniques are applied in increasingly diverse domains, it is important to make this value judgement more explicit, and to offer users more flexibility in controlling the ultimate *goal* of learning. One of the best-known alternatives to the risk is the *conditional value-at-risk* (CVaR), which considers the expected loss, conditioned on the event that the loss exceeds a user-specified $(1-\alpha)$-level quantile, here denoted for each $w \in \mathcal{W}$ as

$$\begin{aligned} C_\alpha(w) &:= \frac{1}{\alpha}\, \mathbf{E}_{\mathrm{P}}\, L(w;Z) I_{\{L(w;Z) \geq V_\alpha(w)\}} \\ &= \frac{1}{\alpha} \int_{L(w;z) \geq V_\alpha(w)} L(w;z)\, \mathrm{P}(\mathrm{d}z), \end{aligned} \tag{2}$$

where $V_\alpha(w) := \inf\{u \in \mathbb{R} : \mathrm{P}\{L(w;Z) \leq u\} \geq 1-\alpha\}$ (called *value-at-risk*, or VaR). Driven by influential work by Artzner et al. (1999) and Rockafellar and Uryasev (2000), under known parametric models, the problem of estimating and minimizing the CVaR reliably and efficiently has been rigorously studied, leading to a wide range of applications in finance (Krokhmal et al., 2002; Mansini et al., 2007), and even some specialized settings

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

of machine learning tasks (Takeda and Sugiyama, 2008; Chow et al., 2016). In general machine learning tasks, however, a non-parametric scenario is more typical, where virtually nothing is known about the distribution of $L(w; Z)$, adding significant challenges to both the design and analysis of procedures designed to minimize the CVaR with high confidence.

Our contributions In this work, we consider the case of potentially heavy-tailed losses, namely a learning setup in which all the learner knows is that the distribution of the loss and its gradients have finite variance, nothing more. It is unknown in advance whether the feedback received is statistically congenial in the sub-Gaussian sense, or highly susceptible to outliers with infinite higher-order moments. Our main contributions:

- New error bounds for a large class of estimators of the CVaR for potentially heavy-tailed random variables (Algorithm 1, Theorem 3).
- A general-purpose learning algorithm which runs stochastic GD sub-processes in parallel and uses the new CVaR estimators to robustly validate the strongest candidate (Algorithm 2), which enjoys sharp excess CVaR bounds (Theorem 4), when both the loss and gradients can be heavy-tailed.
- An empirical study (section 3) highlighting the potential computational advantages and robustness of the proposed approach to CVaR-based learning.

Review of related work To put the contributions stated above in context, we give an overview of the two key strands of technical literature that are closely related to our work. First, an interesting line of work has recently developed which handles risk-averse learning scenarios where the losses can be heavy-tailed, with key works due to Kolla et al. (2019), Prashanth et al. (2019), Bhat and Prashanth (2020), and Kagrecha et al. (2020). These works all consider some kind of subroutine for robustly estimating the CVaR, as we do as well. The actual estimation procedures and proof techniques differ, and we provide a detailed comparison of resulting error bounds in section 2.2.1. Furthermore, the latter three works only consider rather specialized learning algorithms in the context of bandit-like online learning problems, whereas the generic gradient-based procedures we study in section 2.3 have a much wider range of applications. Second, recent work from Cardoso and Xu (2019) and Soma and Yoshida (2020) also consider tackling the CVaR-based learning problem using general-purpose gradient-based stochastic learning algorithms. However, these works assume a bounded (and thus sub-Gaussian) loss; we discuss differences in technical assumptions in detail in Remark 5, but the most important difference is that their setup precludes the possibility of heavy-tailed losses and is thus more restrictive statistically than ours, which naturally leads to different algorithms, proof techniques, and performance guarantees.

2 THEORETICAL ANALYSIS

This section is broken into three sub-sections. First we establish notation and basic technical conditions in section 2.1. We then study pointwise CVaR estimators in section 2.2, and subsequently leverage these results to derive a new learning algorithm with performance guarantees in section 2.3.

2.1 Preliminaries

In the context of learning problems, random variable Z denotes our data, taking values in some measurable space $\mathcal{Z}$ with P the probability measure induced by Z. The set $\mathcal{W} \subseteq \mathbb{R}^d$ is a parameter set from which the learning algorithm chooses an element. We reinforce the point that the ultimate formal goal of learning here is to minimize $C_{\alpha}(\cdot)$ defined in (2) over $\mathcal{W}$, where $0 < \alpha < 1$ is a user-specified risk-level parameter. This is in contrast with the traditional risk-centric setup, which seeks to minimize $R(\cdot)$ defined in (1). For the pointwise estimation problem in section 2.2 to follow, to cut down on excess notation, we simply take $X = L(w; Z)$, re-christen P as the distribution of X, and write the distribution function as $F_{\mathrm{P}}(u) := \mathrm{P}\{X \leq u\}$ for $u \in \mathbb{R}$. Similarly, since the choice of $w \in \mathcal{W}$ is not important in section 2.2, there we shall write simply C_{α} and V_{α} for the CVaR and VaR of X, and return to the w-dependent notation $C_{\alpha}(w)$ and $V_{\alpha}(w)$ in section 2.3. For any $m \geq 1$, we denote by $[m] := \{1, \ldots, \lfloor m \rfloor\}$ all positive integers less than or equal to m. Finally, let $I_{\{\texttt{event}\}}$ denote the indicator function, returning 1 when `event` is true, and 0 otherwise.

Regarding technical assumptions, we shall henceforth assume that $F_{\mathrm{P}} : \mathbb{R} \to [0, 1]$ is continuous, which in particular implies that $F_{\mathrm{P}}(V_{\alpha}) = \mathrm{P}\{X \leq V_{\alpha}\} = 1 - \alpha$ for all α. This setup is entirely traditional; see for example the well-known work of Rockafellar and Uryasev (2000). In general, if F_{P} has flat regions, there may be infinitely many $1 - \alpha$ quantiles; here V_{α} as introduced in section 1 is simply defined to be the smallest one (see Figure 1 for an illustration). The key technical assumption that will be utilized is as follows:

A1. There exists values $0 < \gamma < \lambda < \infty$ such that for any $|u| \leq 1$, the distribution function induced by P satisfies $\gamma u \leq |F_{\mathrm{P}}(V_{\alpha} + u) - F_{\mathrm{P}}(V_{\alpha})| \leq \lambda u$.

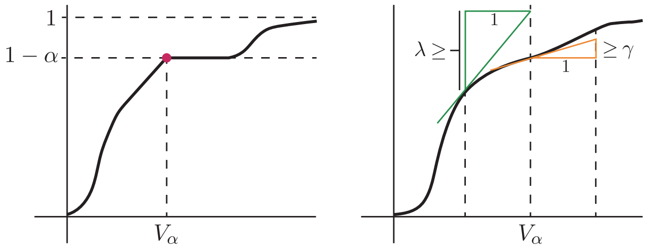

Figure 1: A simple schematic illustrating V_α and the condition A1(γ, λ).

Obviously, we are assuming that $V_\alpha \pm 1$ are within the domain of $X \sim \mathrm{P}$; this is only for notational simplicity, and the range can be taken arbitrarily small. In words, assumption A1(γ, λ) is a *local* assumption of both a λ-Lipschitz property and a γ-growth property, local in the sense that it need only hold around the particular point V_α of interest. The former property ensures that F_{P} cannot jump with arbitrary steepness in the region of interest. The latter ensures that F_{P} is not flat in this region. Finally, we remark that the property of γ-growth is utilized in key recent work done on concentration of CVaR estimators under potentially heavy-tailed data, including Kolla et al. (2019, Prop. 2) and Prashanth et al. (2019, Lem. 5.1).

2.2 Robust estimation of the CVaR criterion

We begin by considering pointwise estimates, assuming that $X \sim \mathrm{P}$ is a non-negative random variable, and that we have $2n$ independent copies of X, denoted $\boldsymbol{X}_n := \{X_1, \ldots, X_n\}$ for the first half, and $\boldsymbol{Y}_n := \{Y_1, \ldots, Y_n\}$ for the second half. The latter half will be used to construct an estimator $\widehat{V}_\alpha \approx V_\alpha$. The former half, with $\widehat{V}_\alpha$ in hand, will be used to construct an estimator $\widehat{C}_\alpha \approx C_\alpha$. As an initial approach to the problem, note that we can decompose the deviations as

$$\begin{aligned}\left|\widehat{C}_\alpha - C_\alpha\right| &= \frac{1}{\alpha}\Big|\alpha\,\widehat{C}_\alpha - \mathbf{E}_{\mathrm{P}}\, X\, I_{\{X \geq \widehat{V}_\alpha\}} \\ &\qquad + \mathbf{E}_{\mathrm{P}}\, X\, I_{\{X \geq \widehat{V}_\alpha\}} - \mathbf{E}_{\mathrm{P}}\, X\, I_{\{X \geq V_\alpha\}}\Big| \\ &\leq \frac{1}{\alpha}\Big(\left|\alpha\,\widehat{C}_\alpha - \mathbf{E}_{\mathrm{P}}\, X\, I_{\{X \geq \widehat{V}_\alpha\}}\right| \\ &\qquad + \left|\mathbf{E}_{\mathrm{P}}\, X\left(I_{\{X \geq \widehat{V}_\alpha\}} - I_{\{X \geq V_\alpha\}}\right)\right|\Big). \qquad (3)\end{aligned}$$

This gives us two terms to control. Starting with the left-most term, let us first make the notation a bit easier to manage. Conditioning on $\boldsymbol{Y}_n$ makes $\widehat{V}_\alpha \in \mathbb{R}$ a fixed value, and based on this, we define

$$X' := X\, I_{\{X \geq \widehat{V}_\alpha\}}. \qquad (4)$$

Since $\widehat{V}_\alpha$ is computed based on available data, and X is observable, it follows that X' itself is observable. Denote the corresponding sample by $\boldsymbol{X}'_n := \{X'_1, \ldots, X'_n\}$, where we set $X'_i := X_i\, I_{\{X_i \geq \widehat{V}_\alpha\}}$. The most direct approach to this problem is to simply pass this transformed dataset $\boldsymbol{X}'_n$ to a sufficiently robust sub-routine for mean estimation. More precisely, we desire a sub-routine $\texttt{RobMean}$ by which assuming only $\mathbf{E}_{\mathrm{P}}\, X^2 < \infty$, for any choice of $\delta \in (0,1)$, we can guarantee that

$$\mathbf{P}\left\{\left|\texttt{RobMean}\,[\boldsymbol{X}'_n] - \mathbf{E}_{\mathrm{P}}\, X'\right| > c\,\sigma'\sqrt{\frac{1+\log(\delta^{-1})}{n}}\right\} \leq \delta, \qquad (5)$$

where $c > 0$ is a constant depending only on the nature of $\texttt{RobMean}$, σ' is any quantity bounded as $\sigma' \leq \sqrt{\mathbf{E}_{\mathrm{P}}(X')^2}$, and probability is taken with respect to the random draw of $\boldsymbol{X}_n$. The final estimator of interest, then, using $2n$ observations in total, will simply be defined as

$$\widehat{C}_\alpha := \frac{1}{\alpha}\widehat{C}'_\alpha\,[\boldsymbol{X}_n, \boldsymbol{Y}_n]\,, \qquad (6)$$
$$\text{where } \widehat{C}'_\alpha\,[\boldsymbol{X}_n, \boldsymbol{Y}_n] := \texttt{RobMean}\,[\boldsymbol{X}'_n]\,.$$

This general procedure is summarized in Algorithm 1.

Algorithm 1 Scaled CVaR under potentially heavy-tailed data; $\widehat{C}'_\alpha\,[\boldsymbol{X}_n, \boldsymbol{Y}_n]$.

```
inputs: samples $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$, risk level $\alpha \in (0,1)$, robust sub-routine RobMean.
Sort ancillary data $Y^{*}_1 \leq Y^{*}_2 \leq \ldots \leq Y^{*}_n$.
Set threshold $\widehat{V}_\alpha = Y^{*}_{\lfloor(1-\alpha)n\rfloor}$.
Augment data $X'_i = X_i\, I_{\{X_i \geq \widehat{V}_\alpha\}}$, for $i \in [n]$.
return: $\widehat{C}'_\alpha\,[\boldsymbol{X}_n, \boldsymbol{Y}_n] = \texttt{RobMean}\,[\{X'_i : i \in [n]\}]$.
```

Deriving deviation bounds Before proceeding any further, the first question to answer is whether or not such a procedure $\texttt{RobMean}$ can be constructed. In the following lemma, we summarize the robust mean estimation performance guarantees available for these estimators.

Lemma 1 (Procedures for good $\boldsymbol{X}_n$ event)**.** *Implementing* $\texttt{RobMean}$ *using the following well-known procedures satisfies (5) at confidence level δ, as follows (details in appendix).*

- *Median of means (Lerasle and Oliveira, 2011): with $c \leq 2\sqrt{\mathrm{e}}$ and $\sigma' = \sqrt{\operatorname{var}_{\mathrm{P}} X'}$, whenever $k = \lceil \log(\delta^{-1}) \rceil$ and $n \geq 2(1+\log(\delta^{-1}))$.*
- *M-estimation (Catoni, 2012): $c \leq 2$ and $\sigma' = \sqrt{\operatorname{var}_{\mathrm{P}} X'}$, whenever $n \geq 4\log(\delta^{-1})$.*
- *Trimmed mean (Lugosi and Mendelson, 2019): with $c \leq 9\sqrt{2}$ and $\sigma' = \sqrt{\operatorname{var}_{\mathrm{P}} X'}$, whenever $n \geq (16/3)\log(8\delta^{-1})$.*

The preceding lemma settles any issues regarding a sufficiently accurate sub-routine `RobMean` under potentially heavy-tailed data. For one concrete example, the median of means sub-routine amounts to splitting the index as $[n] = \cup_{j=1}^{k} \mathcal{I}_j$ and taking the median of each subset mean, i.e.,

$$\operatorname{med}\{\overline{X}^{(1)}, \ldots, \overline{X}^{(k)}\}, \text{ where } \overline{X}^{(j)} = \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} X_i.$$

Next, note that σ' depends on $\widehat{V}_\alpha$, and thus the second sample $\boldsymbol{Y}_n$. To remove this dependence, the following lemma will be useful.

Lemma 2 (Good $\boldsymbol{Y}_n$ event)**.** *Let the observations $\boldsymbol{Y}_n$ sorted in increasing order be denoted by $\boldsymbol{Y}_n^* := \{Y_i^*\}_{i\in[n]}$, such that $Y_1^* \leq Y_2^* \leq \ldots \leq Y_n^*$. It follows that with probability no less than $1 - 2\exp(-3n\alpha/14)$ over the draw of $\boldsymbol{Y}_n$, we have that*

$$V_{2\alpha} \leq Y^*_{(1-\alpha)n} \leq V_{\alpha/2}.$$

Writing $\sigma_\alpha^2 := \mathbf{E}_{\mathrm{P}}\, X^2 I_{\{X \geq V_{2\alpha}\}} - (\mathbf{E}_{\mathrm{P}}\, X\, I_{\{X\geq V_{\alpha/2}\}})^2$, a straightforward argument (detailed derivation in appendix) yields high-probability bounds on the two terms of interest, taking the form

$$\left|\alpha\, \widehat{C}_\alpha - \mathbf{E}_{\mathrm{P}}\, X\, I_{\{X \geq \widehat{V}_\alpha\}}\right| = \left|\widehat{C}'_\alpha - \mathbf{E}_{\mathrm{P}}\, X'\right| \leq c\sigma_\alpha \sqrt{\frac{1+\log(\delta^{-1})}{n}} \tag{7}$$

$$\left|\mathbf{E}_{\mathrm{P}}\, X \left(I_{\{X\geq V_\alpha\}} - I_{\{X \geq \widehat{V}_\alpha\}}\right)\right| \leq \frac{V_{\alpha/2}\lambda}{\sqrt{2}\gamma} \sqrt{\frac{\log(\delta^{-1})}{n}}. \tag{8}$$

Taking (7) and (8) together, applied to (3), we have essentially proved the following result.

Theorem 3. *For any confidence level $\delta \in (0,1)$ and risk level $0 < \alpha < 1/2$, assume that A1(γ,λ) holds and $n \geq \log(\delta^{-1}) \max\{1/(2\gamma)^2, 14/(3\alpha)\}$. Letting $\widehat{C}'_\alpha$ be the output of Algorithm 1, and $\widehat{C}_\alpha = \widehat{C}'_\alpha/\alpha$, with probability no less than $1 - 5\delta$, we have*

$$\left|\widehat{C}_\alpha - C_\alpha\right| \leq \frac{1}{\alpha}\left(c\sigma_\alpha + \frac{V_{\alpha/2}\lambda}{\sqrt{2}\gamma}\right)\sqrt{\frac{1+\log(\delta^{-1})}{n}},$$

where c depends only on the choice of `RobMean` *(specified in Lemma 1).*

Proof of Theorem 3. To prove this result simply involves sorting out the key facts presented above. The "good" event in the theorem statement is that in which both (7) and (8) hold together. This condition can fail if even one of the following bad events takes place:

$$\mathcal{E}_1 := \{(5) \text{ fails}\}, \quad \mathcal{E}_2 := \{\text{event of Lemma 2 fails}\},$$
$$\mathcal{E}_3 := \left\{|\widehat{V}_\alpha - V_\alpha| > \sqrt{\log(\delta^{-1})/(2\gamma^2 n)}\right\}.$$

First of all, using Lemma 1 and the deviation bounds given by (5), we have

$$\mathbf{P}(\mathcal{E}_1) = \mathbf{E}_{\boldsymbol{Y}_n}\, \mathbf{P}\left[\mathcal{E}_1 \,|\, \boldsymbol{Y}_n\right] \leq \delta.$$

Next, by Lemma 2, if $n \geq 14\log(\delta^{-1})/(3\alpha)$, then we have $\mathbf{P}(\mathcal{E}_2) \leq 2\delta$. Finally, from the derivation of (8), whenever $n \geq \log(\delta^{-1})/(2\gamma^2)$, we have $\mathbf{P}(\mathcal{E}_3) \leq 2\delta$. If none of these three bad events take place, the good event holds, i.e., $(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)^c \subseteq \{(7) \text{ and } (8)\}$. A union bound implies that this holds with probability no less than $1 - 4\delta$, and via the original decomposition (3), we have

$$\left|\widehat{C}_\alpha - C_\alpha\right| \leq \frac{1}{\alpha}\left(c\sigma_\alpha\sqrt{\frac{1+\log(\delta^{-1})}{n}} + \frac{V_{\alpha/2}\lambda}{\sqrt{2}\gamma}\sqrt{\frac{\log(\delta^{-1})}{n}}\right),$$

which implies the desired result. □

2.2.1 Comparison of estimation error bounds

From the technical literature on CVaR estimation under potentially heavy-tailed data, the work of Kolla et al. (2019), Prashanth et al. (2019), and Kagrecha et al. (2020) are most closely related to our work, and in this remark we compare our results with theirs. To align our setup with theirs, we assume access to only n data points in total, meaning the two data sets used in Theorem 3 will now be $\boldsymbol{X}_{n/2}$ and $\boldsymbol{Y}_{n/2}$, for simplicity assuming that n is even. Furthermore, we convert our high-confidence interval into an exponential tail bound, which is the form taken by the main results in the cited works. First, given just n observations, our Theorem 3 implies that

$$\mathbf{P}\left\{\left|\widehat{C}_\alpha - C_\alpha\right| > \varepsilon\right\} \leq 5\exp\left(-n\left(\alpha\varepsilon/B_{\text{OURS}}\right)^2\right),$$
$$\text{with } B_{\text{OURS}} := c\sigma_\alpha + \frac{\sqrt{2}V_{\alpha/2}\lambda}{\gamma}.$$

The estimator $\widehat{C}_\alpha$ considered by Prashanth et al. (2019, Thm. 4.1), on the other hand, yields bounds of the form

$$\mathbf{P}\left\{\left|\widehat{C}_\alpha - C_\alpha\right| > \varepsilon\right\} \leq 8\exp\left(-n\left(\alpha\varepsilon/B'\right)^2\right),$$

where the factor B' is simply left as a "distribution-dependent factor." Looking at their proof, in order to obtain concentration of the VaR estimator, they also effectively require a γ-growth property and have moment dependence. Furthermore, their proof is rather specialized to an estimator borrowed from Bubeck et al. (2013), which does random truncation that is rather unintuitive when taken outside the context of online learning problems. Another closely related result published very recently is due to Kagrecha et al. (2020).

They consider a more natural estimator, which simply truncates the data to $|X_i| \leq b$ before passing it to the classical empirical CVaR estimator routine. While b is a user-specified parameter, it must be taken larger than a value which depends on the desired deviation level ε. In particular, since it must satisfy $b = \Omega(\mathbf{E}_{\mathrm{P}} X^2/(\alpha\varepsilon))$, when ε is sufficiently small, one ends up with bounds of the form

$$\mathbf{P}\left\{\left|\widehat{C}_\alpha - C_\alpha\right| > \varepsilon\right\} \leq 6\exp\left(-n\alpha^3\varepsilon^4/B''\right), \\ \text{with } B'' := 616\left(\mathbf{E}_{\mathrm{P}} X^2\right)^2.$$

Their results are obtained using very weak assumptions, the finiteness of $\mathbf{E}_{\mathrm{P}} X^2$ is all that is required. The price paid for this generality is clearly the poor dependence on α, ε, and the moments. In contrast, under mild additional assumptions on the behaviour of the distribution function around V_α (namely A1(γ,λ)), we obtain much stronger results, using a very simple proof strategy, which can be readily applied to a wide collection of estimation routines.

2.3 CVaR-driven learning algorithms

We now proceed to our main point of interest, namely learning algorithms which seek to minimize the CVaR of the loss distribution, defined in (2), given only a sample $\boldsymbol{Z}_n := \{Z_1, \ldots, Z_n\}$, independent copies of $Z \sim \mathrm{P}$. Computationally, it is convenient to introduce

$$f_\alpha(w, v; Z) := v + \frac{1}{\alpha}\left[L(w;Z) - v\right]_+, \qquad (9)$$

defined for all $w \in \mathcal{W}, v \in \mathbb{R}$. Denote the expected value denoted by $F_\alpha(w,v) := \mathbf{E}_{\mathrm{P}} f_\alpha(w,v;Z)$, not to be confused with F_{P} from the previous section. This expectation has the useful property of being convex and continuously differentiable in v, and being related to the quantities $C_\alpha(w)$ and $V_\alpha(w)$ through

$$\min\{F_\alpha(w,v) : v \in \mathbb{R}\} = F_\alpha(w, V_\alpha(w)) = C_\alpha(w),$$

which holds for any choice of $w \in \mathcal{W}$ (Rockafellar and Uryasev, 2000, Thm. 1). This implies that if we have some candidates $(\widehat{w}, \widehat{v})$ such that $F_\alpha(\widehat{w}, \widehat{v}) \leq \varepsilon$, then $C_\alpha(\widehat{w}) \leq F_\alpha(\widehat{w}, \widehat{v}) \leq \varepsilon$. Furthermore, solving the joint problem is equivalent to solving the two problems separately (Rockafellar and Uryasev, 2000, Thm. 2), meaning that $F_\alpha^* = C_\alpha^*$, where we denote $F_\alpha^* := \inf\{F_\alpha(w,v) : (w,v) \in \mathcal{W}\times\mathbb{R}\}$, $C_\alpha^* := \inf\{C_\alpha(w) : w \in \mathcal{W}\}$. When $L(w;Z)$ is convex in w, the function F_α is jointly convex in its arguments, and thus when $\mathcal{W} \subseteq \mathbb{R}^d$ is a convex set, convex optimization techniques can in principle be brought to bear on the problem.

Problems with robust objectives Recalling the analysis of the previous section 2.2, we constructed a procedure for obtaining sharp estimates of $C_\alpha(w)$, pointwise in w, under potentially heavy-tailed data. To extend the procedure given by Algorithm 1 and (6) to this setting, given an extra sample $\boldsymbol{Z}_n'$, compute

$$\begin{aligned}\widehat{C}_\alpha'(w; \boldsymbol{Z}_n') &:= \\ &\widehat{C}_\alpha'\left[\boldsymbol{X} = \{L(w;Z_i') : i \in [\lfloor n/2 \rfloor]\},\right. \\ &\quad \left.\boldsymbol{Y} = \{L(w;Z_i') : n/2 < i \leq n\}\right], \end{aligned} \qquad (10)$$

and set $\widehat{C}_\alpha(w) = \widehat{C}_\alpha'(w; \boldsymbol{Z}_n')/\alpha$. The most naive approach to this problem would be to replace the empirical mean with this robust estimator (10), namely any algorithm implementing

$$\widehat{w} \in \operatorname*{arg\,min}_{w \in \mathcal{W}} \widehat{C}_\alpha'(w; \boldsymbol{Z}_n)/\alpha.$$

The statistical properties of such an $\widehat{w}$ are naturally of interest, but the computational task of actually obtaining such a $\widehat{w}$ is highly non-trivial; for example the work of Brownlees et al. (2015) consider a similar quantity in the case of traditional risk minimization, but algorithmic considerations are left completely abstract. Indeed, even if $L(\cdot, z)$ is convex and smooth for all $z \in \mathcal{Z}$, we have no guarantee that $\widehat{C}_\alpha'(\cdot; \boldsymbol{Z}_n)$ will be. The exact same issues hold if we tackle a robustified version of the joint optimization task, namely

$$(\widehat{w}, \widehat{v}) \in \operatorname*{arg\,min}_{(w,v)\in\mathcal{W}\times\mathbb{R}} \texttt{RobMean}\left[\{f_\alpha(w,v;Z_i) : i \in [n]\}\right],$$

where `RobMean` is based on any procedure given in Lemma 1. All the robust estimates given by `RobMean` (or Algorithm 1) are easy to *compute* for any (w,v) or w, but are hard to *minimize*. It thus seems wiser to use such sub-routines for *validation*, i.e., to check that a particular candidate $\widehat{w}$ actually gets close to minimizing $C_\alpha(\cdot)$ with sufficiently high confidence.

A practical approach under heavy tails With this intuition in mind, we consider a simple divide-and-conquer procedure with independent sub-processes running stochastic gradient descent for the joint optimization of F_α, and a final robust validation step to determine a final candidate (Holland, 2021b,a). This is summarized in Algorithm 2, and we unpack the notation below.

Most of the steps in Algorithm 2 are transparent; it just remains to provide a more precise definition of the `SGD` sequence referred to in the third line. Given a sequence of observations $(Z_1, \ldots, Z_t)$ of arbitrary length $t \geq 1$, the core update is traditional projected stochastic sub-gradient descent:

$$\begin{aligned}&(\widehat{w}_t, \widehat{v}_t) = \\ &\Pi_{\mathcal{W}\times[0,V]}\left[(\widehat{w}_{t-1}, \widehat{v}_{t-1}) - \beta_t\, G_\alpha(\widehat{w}_{t-1}, \widehat{v}_{t-1}; Z_t)\right]\end{aligned} \qquad (11)$$

Algorithm 2 Fast gradient-based CVaR learning with robust verification.

```
**inputs:** samples $\boldsymbol{Z}_n$ and $\boldsymbol{Z}'_n$, initial value $(\widehat{w}_0, \widehat{v}_0)$, parameters $\alpha \in (0,1)$, $0 < V < \infty$, $1 \leq k \leq n$.
Split $\bigcup_{j=1}^{k} \mathcal{I}_j = [n]$, with $|\mathcal{I}_j| \geq \lfloor n/k \rfloor$, and $\mathcal{I}_j \cap \mathcal{I}_l = \emptyset$ when $j \neq l$.    ▷ Disjoint partition.
For each $j \in [k]$, set $(\overline{w}^{(j)}, \overline{v}^{(j)})$ to the mean of sequence $\texttt{SGD}(\widehat{w}_0, \widehat{v}_0; \boldsymbol{Z}_{\mathcal{I}_j}, \mathcal{W} \times [0,V])$.
Compute $\star = \operatorname*{arg\,min}_{j \in [k]} \widehat{C}'_\alpha\left(\overline{w}^{(j)}; \boldsymbol{Z}'_n\right)$.    ▷ Robust validation via (10), based on Algorithm 1.
**return** $\overline{w}^{(\star)}$.
```

The update direction here is $G_\alpha(w,v;Z) \in \partial f_\alpha(w,v;Z)$, namely any vector from the sub-differential of the map $(w,v) \mapsto f_\alpha(w,v;Z)$. The operator Π denotes projection in the ℓ_2 norm, and $\beta_t \geq 0$ is a step-size parameter. The recursive definition in (11) bottoms out at $t=1$, and is initialized by some pre-defined $(\widehat{w}_0, \widehat{v}_0)$, passed to the algorithm as an input. The sequence $\texttt{SGD}(\widehat{w}_0, \widehat{v}_0; \boldsymbol{Z}_{\mathcal{I}_j}, \mathcal{W} \times [0,V])$ referred to in Algorithm 2 is simply the sequence of iterates generated by (11) using data $\{Z_t : t \in \mathcal{I}_j\}$; since all Z_t are independent copies of $Z \sim \mathrm{P}$, the order does not matter. The key technical assumptions on the data are summarized below:

A2. Let A1(γ,λ) hold for $X = L(w;Z) \geq 0$, for any choice of $w \in \mathcal{W}$. Let $\mathcal{W}$ be convex, have a diameter in ℓ_2 norm of $0 < \Delta < \infty$. Let $\bar{\sigma}_\alpha := \max\{\sigma_\alpha(w) : w \in \mathcal{W}\} < \infty$ and $\bar{V}_\alpha := \max\{V_\alpha(w) : w \in \mathcal{W}\} < \infty$. Let $L(w;z)$ be a convex, differentiable function of w for all $z \in \mathcal{Z}$, and let $\mathbf{E}_{\mathrm{P}} \|\nabla L(w;Z)\|^2 \leq \lambda_L^2$ for all $w \in \mathcal{W}$.

Note $\sigma_\alpha(w)$ extends σ_α from section 2.2 to the case of $X = L(w;Z)$.

The preceding assumptions clearly allow for potentially heavy-tailed losses and gradients. As a concrete illustration of this, consider linear regression using squared error and a linear model, so that $L(w;Z) = (\langle w - w^*, X\rangle + \epsilon)^2$, where X is a d-dimensional random vector, and ϵ is additive noise. Convexity and differentiability as required by A2 are essentially immediate. As for the moment bound, noting that $\nabla L(w;Z) = 2(\langle w - w^*, X\rangle + \epsilon)X$, basic algebra and an application of Cauchy-Schwarz gives us $\mathbf{E}_{\mathrm{P}} \|\nabla L(w;Z)\|^2 \leq \lambda_L^2$, where

$$\lambda_L \leq 2\sqrt{\Delta^2\, \mathbf{E}_{\mathrm{P}} \|X\|^4 + \mathbf{E}_{\mathrm{P}} |\epsilon|^2 \|X\|^2 + 2\Delta\, \mathbf{E}_{\mathrm{P}} |\epsilon| \|X\|^3}. \tag{12}$$

In particular, the random noise ϵ and inputs X need not be bounded, nor are they required to have finite higher-order moments. As such, A2 can be satisfied on problems of practical interest when the "feedback" (CVaR loss and sub-gradients) is potentially heavy-tailed. Under this setting, the following performance guarantee holds.

Theorem 4. *Under assumption A2, run Algorithm 2 with parameters $0 < \alpha < 1/2$, $V = \bar{V}_\alpha$, $k = \lceil \log(2\lceil \log(\delta^{-1}) \rceil \delta^{-1}) \rceil$ for arbitrary choice of $\delta \in (0,1)$, and fix the step sizes in (11) to*

$$\beta_t = \alpha \sqrt{\frac{\Delta^2 + \bar{V}_\alpha}{(\lambda_L^2 + 1)|\mathcal{I}_j|}}$$

for each sub-process, indexed by $j \in [k]$. We have

$$C_\alpha(\overline{w}^{(\star)}) - C_\alpha^* \leq \frac{2\sqrt{2}}{\alpha}\left(c\bar{\sigma}_\alpha + \frac{\bar{V}_{\alpha/2}\lambda}{\sqrt{2}\gamma}\right)\sqrt{\frac{1 + \log(5\delta^{-1})}{n}} + \frac{\mathrm{e}}{\alpha}\sqrt{\frac{k(\lambda_L^2 + 1)(\Delta^2 + \bar{V}_\alpha^2)}{n}}$$

with probability no less than $1 - 3\delta$, where constant c corresponds to the relevant constant from Lemma 1.

Remark 5 (Discussion of related technical work). As far as technical conditions go, the convexity and bounded diameter assumptions align with Soma and Yoshida (2020, Thm. 3.6). The main difference is that they assume bounded and Lipschitz-continuous losses, which precludes both heavy-tailed losses and gradients. Algorithmically, they run a single averaged SGD process using a surrogate objective, for multiple passes over the data, and further assuming the losses are smooth, obtain error bounds in expectation. In contrast, as discussed above, we allow both losses and gradients to be heavy-tailed, we do not require the gradients to be Lipschitz. Our high-probability guarantees are obtained for a procedure which runs multiple SGD processes in parallel, each of which takes only a single pass over the subset of data allocated to it. Finally, we remark that since their procedure does not actually make any direct estimates of V_α, they do not use an assumption like A1. Note that it is certainly possible to modify our Algorithm 2 such that this assumption is not needed, by doing the final validation step based on an estimate of F_α instead of C_α. This would remove the need for A1, and instead result in bounds depending on the second moment of $f_\alpha(w,v;Z)$. The formal analysis goes through in a perfectly analogous fashion to our proof of Theorem 4 here. ■

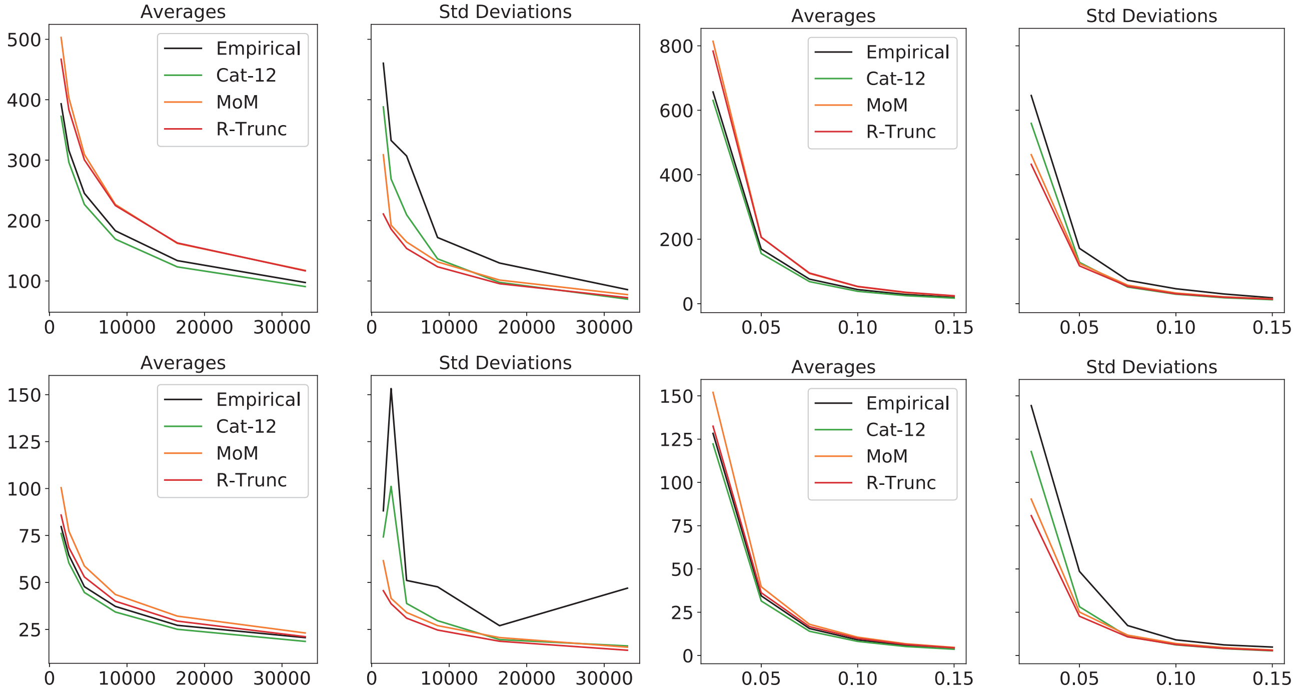

Figure 2: Average and standard deviation of $|\widehat{C}_\alpha - C_\alpha|$ over α and n. Left: fixed $\alpha = 0.05$, increasing n. Right: fixed $n = 10000$, increasing α. All methods were essentially the same for the folded-Normal case, given in the appendix. Top: log-Normal. Bottom: Pareto.

3 EMPIRICAL ANALYSIS

In this section, we start with a numerical investigation of the efficiency of pointwise CVaR estimation enabled by the analysis of section 2.2, using concrete implementations of Algorithm 1, comparing efficient robust estimators against more naive benchmarks. This is followed by an empirical analysis of the performance of CVaR-driven learning algorithms, including Algorithm 2 studied in section 2.3, under an environment in which the nature of the feedback provided to the learner is controlled to range between sub-Gaussian and heavy-tailed.

Accuracy of pointwise estimates First, we consider "static" tests looking at the accuracy of CVaR estimators newly captured by the analysis of section 2.2. Recalling the notation of section 2.2, given samples $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$, all sampled independently from $X \sim \mathrm{P}$, the objective here is to investigate the deviations $|\widehat{C}_\alpha - C_\alpha|$, in particular how these deviations change for different estimators $\widehat{C}_\alpha$, distributions P, sample sizes n, and risk levels α. We consider folded-Normal, log-Normal, and Pareto distributions for P. We study the classical empirical estimate (denoted `Empirical`), random truncation (Prashanth et al., 2019) (`R-Trunc`), and Algorithm 1 implemented using median-of-means (`MoM`) and Catoni-type M-estimation (`Cat`). Further details of the experimental setup are relegated to the supplementary materials.[1] Key results are summarized in Figure 2, where averages and standard deviations of these deviations over many trials are given. As a general take-away, we see that using a slightly more sophisticated estimation procedure can lead to clear improvements in estimation in a potentially heavy-tailed setting. The concrete procedure which tended to perform best overall (`Cat-12`) is a procedure captured by the theory of section 2.2.

Application to learning algorithms Next, we conduct "dynamic" tests which look at applications of Algorithm 2 in section 2.3 to machine learning tasks. As a natural first application, we consider linear regression in the context of CVaR-based learning. That is, random data are generated as pairs $Z = (X, Y) \sim \mathrm{P}$ following the relation $Y = \langle w^*, X \rangle + E$, where E is a zero-mean random noise term independent of X, and $w^* \in \mathcal{W}$ is some pre-fixed vector, and the goal is to minimize $C_\alpha(\cdot)$ induced by two losses, namely squared error and absolute deviations, respectively amounting to $L(w; Z) = (\langle w - w^*, X \rangle - E)^2/2$ and $L(w; Z) = |\langle w - w^*, X \rangle - E|$. The learner does not know w^* and cannot observe E directly, all it has is access to X and Y, and thus the final loss values (and

[1] Software repository:
https://github.com/feedbackward/robrisk

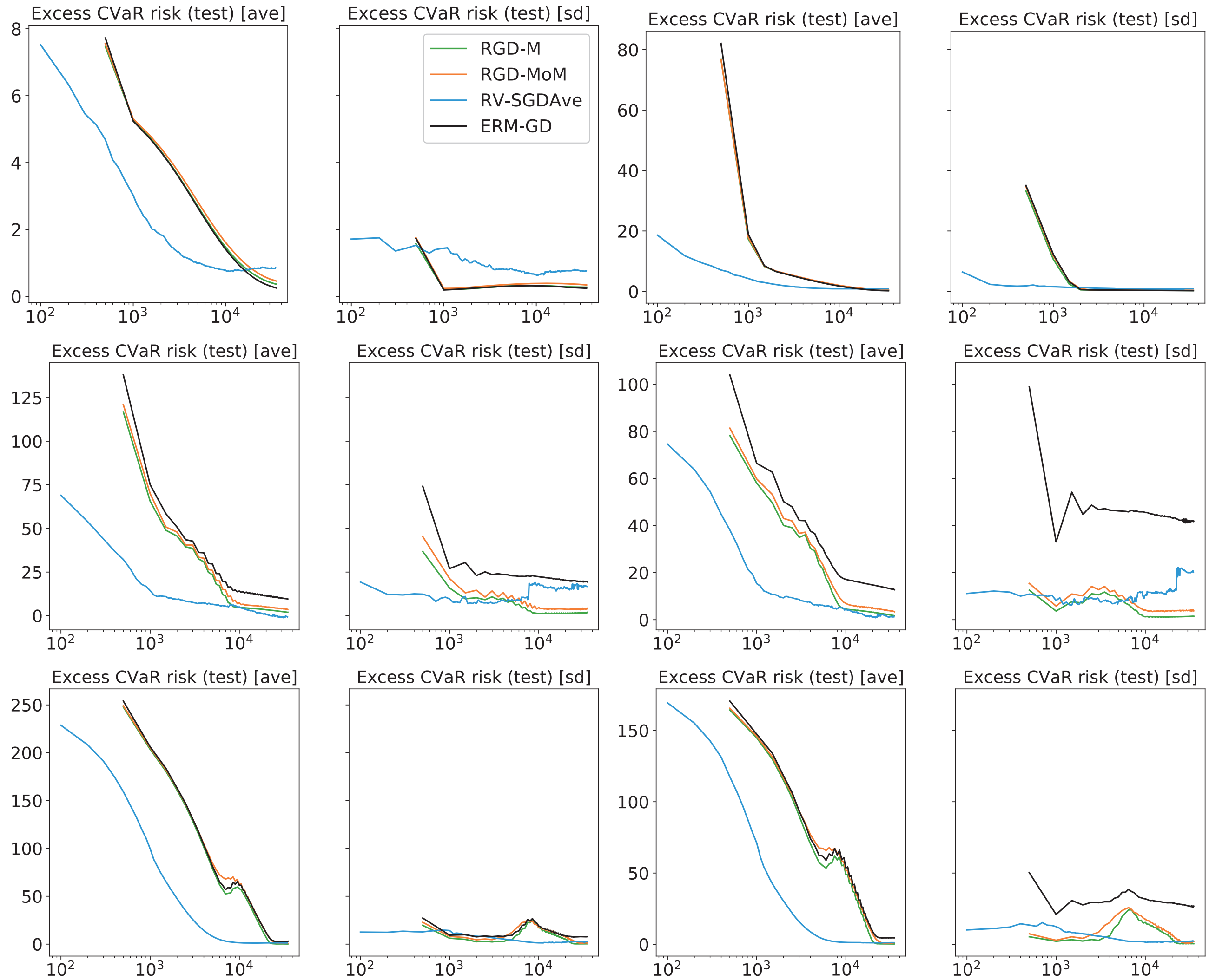

Figure 3: Average and standard deviation of excess CVaR for squared error (left-most plots) and absolute error (right-most plots). Top: Normal. Middle: log-Normal. Bottom: Pareto.

resulting partial derivatives, etc.). We consider Normal, log-Normal, and Pareto distributions for the noise E. We compare Algorithm 2 (denoted `RV-SGDAve`) with three well-known baseline methods. As a classic baseline, we run batch gradient descent empirical CVaR-risk minimization (`ERM-GD`). As modern alternatives, we run robust gradient descent using M-estimation (Holland and Ikeda, 2019) (`RGD-M`) and median-of-means (Chen et al., 2017; Prasad et al., 2018) `RGD-MoM`. Additional details are given in the supplementary materials.

Representative results are given in Figure 3. While the sample splitting leads to a small hit in performance under the Normal case, as a general take-away, we see that the proposed algorithm offers an appealing improvement in efficiency, realizing superior CVaR-risk using far less operations. Furthermore, this is robust both to the underlying distribution, and the nature of the underlying loss. That is, even when the λ_L-Lipschitz assumption on the loss breaks down (left-hand side of Figure 3), we see competitive behaviour.

4 FUTURE DIRECTIONS

One appealing future direction is to go beyond CVaR to more diverse classes of feedback, such as general coherent risks under potentially heavy-tailed data, or even extensions to completely distinct performance classes that in some sense mimic human loss/reward systems (e.g., cumulative prospect theory). Initial explorations have been made by Bhat and Prashanth (2020), but the basic theory and algorithmic analysis are still far from complete. Other notions of conditional expectation, which do not necessarily depend on quantiles, is another natural approach of interest.

Acknowledgements

This work was partially supported by the JSPS KAKENHI Grant Number 19K20342, and the Kayamori Foundation of Information Science Advancement Grant Number K30-XXIII-518.

References

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

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

Bhat, S. P. and Prashanth, L. A. (2020). Concentration of risk measures: A Wasserstein distance approach. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, 43(6):2507–2536.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Cardoso, A. R. and Xu, H. (2019). Risk-averse stochastic convex bandit. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 39–47.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

Chen, Y., Su, L., and Xu, J. (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. In *Proceedings of the ACM on Measurement and Analysis of Computing Systems*. ACM.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. (2016). Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 1522–1530.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150.

Holland, M. J. (2021a). Robustness and scalability under heavy tails, without strong convexity. In *24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*, volume 130 of *Proceedings of Machine Learning Research*.

Holland, M. J. (2021b). Scaling-up robust gradient descent techniques. In *35th AAAI Conference on Artificial Intelligence (AAAI 2021)*.

Holland, M. J. and Ikeda, K. (2019). Better generalization with less data using robust gradient descent. In *36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*.

Kagrecha, A., Nair, J., and Jagannathan, K. (2020). Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Kolla, R. K., Prashanth, L. A., Bhat, S. P., and Jagannathan, K. (2019). Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47(1):16–20.

Krokhmal, P., Palmquist, J., and Uryasev, S. (2002). Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4:43–68.

Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

Lugosi, G. and Mendelson, S. (2019). Robust multivariate mean estimation: the optimality of trimmed mean. *arXiv preprint arXiv:1907.11391v1*.

Mansini, R., Ogryczak, W., and Speranza, M. G. (2007). Conditional value at risk and related linear programming models for portfolio optimization. *Annals of Operations Research*, 152(1):227–256.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

Prashanth, L. A., Jagannathan, K., and Kolla, R. K. (2019). Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. *arXiv preprint arXiv:1901.00997v2*.

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

Soma, T. and Yoshida, Y. (2020). Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*.

Takeda, A. and Sugiyama, M. (2008). ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1056–1063.