PAC-Bayesian Bound for the Conditional Value at Risk

Zakaria Mhammedi
The Australian National University and Data61
zak.mhammedi@anu.edu.au

Benjamin Guedj
Inria and University College London
benjamin.guedj@inria.fr

Robert C. Williamson
bobwilliamsonoz@icloud.com

Abstract

Conditional Value at Risk (CVAR) is a family of "coherent risk measures" which generalize the traditional mathematical expectation. Widely used in mathematical finance, it is garnering increasing interest in machine learning, e.g., as an alternate approach to regularization, and as a means for ensuring fairness. This paper presents a generalization bound for learning algorithms that minimize the CVAR of the empirical loss. The bound is of PAC-Bayesian type and is guaranteed to be small when the empirical CVAR is small. We achieve this by reducing the problem of estimating CVAR to that of merely estimating an expectation. This then enables us, as a by-product, to obtain concentration inequalities for CVAR even when the random variable in question is unbounded.

1   Introduction

The goal in statistical learning is to learn hypotheses that generalize well, which is typically formalized by seeking to minimize the *expected* risk associated with a given loss function. In general, a loss function is a map $\ell \colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, where $\mathcal{X}$ is a feature space and $\mathcal{H}$ is an hypotheses space. In this case, the *expected risk* associated with a given hypothesis $h \in \mathcal{H}$ is given by $\mathrm{R}[\ell(h, X)] \coloneqq \mathrm{E}[\ell(h, X)]$. Since the data-generating distribution is typically unknown, the expected risk is approximated using observed i.i.d. samples $X_1, \ldots, X_n$ of X, and an hypothesis is then chosen to minimize the empirical risk $\widehat{\mathrm{R}}[\ell(h, X)] \coloneqq \sum_{i=1}^{n} \ell(h, X_i)/n$. When choosing an hypothesis $\hat{h}$ based on the empirical risk $\widehat{\mathrm{R}}$, one would like to know how close $\widehat{\mathrm{R}}[\ell(\hat{h}, X)]$ is to the actual risk $\mathrm{R}[\ell(\hat{h}, X)]$; only then can one infer something about the *generalization* property of the learned hypothesis $\hat{h}$.

Generalization bounds—in which the expected risk is bounded in terms of its empirical version up to some error—are at the heart of many machine learning problems. The main techniques leading to such bounds comprise uniform convergence arguments (often involving the Rademacher complexity of the set $\mathcal{H}$), algorithmic stability arguments (see e.g. [Bousquet and Elisseeff, 2002] and more recently the work from [Abou-Moustafa and Szepesvári, 2019, Bousquet et al., 2019, Celisse and Guedj, 2016]), and the PAC-Bayesian analysis for non-degenerate randomized estimators [McAllester, 2003]. Behind these techniques lies *concentration inequalities*, such as Chernoff's inequality (for the PAC-Bayesian analysis) and McDiarmid's inequality (for algorithmic stability and the uniform convergence analysis), which control the deviation between population and empirical averages [see Boucheron et al., 2003, 2013, McDiarmid, 1998, among others].

Standard concentration inequalities are well suited for learning problems where the goal is to minimize the expected risk $\mathrm{E}[\ell(h, X)]$. However, the expected risk—the mean performance of an algorithm—might fail to capture the underlying phenomenon of interest. For example, when

dealing with medical (responsitivity to a specific drug with grave side effects, etc.), environmental (such as pollution, exposure to toxic compounds, etc.), or sensitive engineering tasks (trajectory evaluation for autonomous vehicles, etc.), the mean performance is not necessarily the best objective to optimize as it will cover potentially disastrous mistakes (*e.g.*, a few extra centimeters when crossing another vehicle, a slightly too large dose of a lethal compound, etc.) while possibly improving on average. There is thus a growing interest to work with alternative measures of risk (other than the expectation) for which standard concentration inequalities do not apply directly. Of special interest are *coherent risk measures* [Artzner et al., 1999] which possess properties that make them desirable in mathematical finance and portfolio optimization [Allais, 1953, Ellsberg, 1961, Rockafellar et al., 2000], with a focus on optimizing for the worst outcomes rather than on average. Coherent risk measures have also been recently connected to fairness, and appear as a promising framework to control the fairness of an algorithm's solution [Williamson and Menon, 2019].

A popular coherent risk measure is the Conditional Value at Risk (CVAR; see Pflug, 2000); for $\alpha \in (0,1)$ and random variable Z, $\mathrm{CVAR}_\alpha[Z]$ measures the expectation of Z conditioned on the event that Z is greater than its $(1-\alpha)$-th quantile. CVAR has been shown to underlie the classical SVM [Takeda and Sugiyama, 2008], and has in general attracted a large interest in machine learning over the past two decades [Bhat and Prashanth, 2019, Chen et al., 2009, Chow and Ghavamzadeh, 2014, Huo and Fu, 2017, Morimura et al., 2010, Pinto et al., 2017, Prashanth and Ghavamzadeh, 2013, Takeda and Kanamori, 2009, Tamar et al., 2015, Williamson and Menon, 2019].

Various concentration inequalities have been derived for $\mathrm{CVAR}_\alpha[Z]$, under different assumptions on Z, which bound the difference between $\mathrm{CVAR}_\alpha[Z]$ and its standard estimator $\widehat{\mathrm{CVAR}}_\alpha[Z]$ with high probability [Bhat and Prashanth, 2019, Brown, 2007, Kolla et al., 2019, Prashanth and Ghavamzadeh, 2013, Wang and Gao, 2010]. However, none of these works extend their results to the statistical learning setting where the goal is to learn an hypothesis from data to minimize the conditional value at risk. In this paper, we fill this gap by presenting a sharp PAC-Bayesian generalization bound when the objective is to minimize the conditional value at risk.

Related Works. Deviation bounds for CVAR were first presented by Brown [2007]. However, their approach only applies to bounded continuous random variables, and their lower deviation bound has a sub-optimal dependence on the level α. Wang and Gao [2010] later refined their analysis to recover the "correct" dependence in α, albeit their technique still requires a two-sided bound on the random variable Z. Thomas and Learned-Miller [2019] derived new concentration inequalities for CVAR with a very sharp empirical performance, even though the dependence on α in their bound is sub-optimal. Further, they only require a one-sided bound on Z, without a continuity assumption.

Kolla et al. [2019] were the first to provide concentration bounds for CVAR when the random variable Z is unbounded, but is either sub-Gaussian or sub-exponential. Bhat and Prashanth [2019] used a bound on the Wasserstein distance between true and empirical cumulative distribution functions to substantially tighten the bounds of Kolla et al. [2019] when Z has finite exponential or kth-order moments; they also apply their results to other coherent risk measures. However, when instantiated with bounded random variables, their concentration inequalities have sub-optimal dependence in α.

On the statistical learning side, Duchi and Namkoong [2018] present generalization bounds for a class of coherent risk measures that technically includes CVAR. However, their bounds are based on uniform convergence arguments which lead to looser bounds compared with ours. Another bound based on a uniform convergence argument was also presented in a concurrent work by Curi et al. [2020].

Contributions. Given a learning algorithm that outputs a posterior distribution $\widehat{\rho}$ on $\mathcal{H}$, our main contribution is a *PAC-Bayesian generalization bound for the conditional value at risk*, where we bound the difference $\mathrm{CVAR}_\alpha[Z] - \widehat{\mathrm{CVAR}}_\alpha[Z]$, for $\alpha \in (0,1)$ and $Z \coloneqq \mathrm{E}_{h\sim\widehat{\rho}}[\ell(h,X)]$, by a term of order $\sqrt{\widehat{\mathrm{CVAR}}_\alpha[Z] \cdot \mathcal{K}_n/(n\alpha)}$, with $\mathcal{K}_n$ representing a complexity term which depends on $\mathcal{H}$. Due to the presence of $\widehat{\mathrm{CVAR}}_\alpha[Z]$ inside the square-root, our generalization bound has the desirable property that it becomes small whenever the empirical conditional value at risk is small. For the standard expected risk, only state-of-the-art PAC-Bayesian bounds share this property (see *e.g.* Catoni [2007], Langford and Shawe-Taylor [2003], Maurer [2004] or more recently in Mhammedi et al. [2019], Tolstikhin and Seldin [2013]). We refer to [Guedj, 2019] for a recent survey on PAC-Bayes.

As a by-product of our analysis, we derive a new way of obtaining *concentration bounds for the conditional value at risk* by reducing the problem to estimating expectations using empirical means. This reduction then makes it easy to obtain concentration bounds for $\text{CVAR}_\alpha[Z]$ even when the random variable Z is unbounded (Z may be sub-Gaussian or sub-exponential). Our bounds have explicit constants and are sharper than existing ones due to Bhat and Prashanth [2019], Kolla et al. [2019] which deal with the unbounded case.

Outline. In Section 2, we define the conditional value at risk along with its standard estimator. In Section 3, we recall the statistical learning setting and present our PAC-Bayesian bound for CVAR. The proof of our main bound is in Section 4. In Section 5, we present a new way of deriving concentration bounds for CVAR which stems from our analysis in Section 4. Section 6 concludes and suggests future directions.

2 Preliminaries

Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $p \in \mathbb{N}$, we denote by $\mathcal{L}^p(\Omega) := \mathcal{L}^p(\Omega, \mathcal{F}, P)$ the space of p-integrable functions, and we let $\mathcal{M}_P(\Omega)$ be the set of probability measures on Ω which are absolutely continuous with respect to P. We reserve the notation E for the expectation under the reference measure P, although we sometimes write E_P for clarity. For random variables $Z_1, \ldots, Z_n$, we denote $\widehat{P}_n := \sum_{i=1}^n \delta_{Z_i}/n$ the empirical distribution, and we let $Z_{1:n} := (Z_1, \ldots, Z_n)$. Furthermore, we let $\pi := (1, \ldots, 1)^\top/n \in \mathbb{R}^n$ be the uniform distribution on the simplex. Finally, we use the notation $\widetilde{\mathcal{O}}$ to hide log-factors in the sample size n.

Coherent Risk Measures (CRM). A CRM [Artzner et al., 1999] is a functional $\text{R} \colon \mathcal{L}^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ that is simultaneously, positive homogeneous, monotonic, translation equivariant, and sub-additive[1] (see Appendix B for a formal definition). For $\alpha \in (0, 1)$ and a real random variable $Z \in \mathcal{L}_1(\Omega)$, the conditional value at risk $\text{CVAR}_\alpha[Z]$ is a CRM and is defined as the mean of the random variable Z conditioned on the event that Z is greater than its $(1 - \alpha)$-th quantile[2]. This is equivalent to the following expression, which is more convenient for our analysis:

$$\text{CVAR}_\alpha[Z] = \text{C}_\alpha[Z] := \inf_{\mu \in \mathbb{R}} \left\{ \mu + \frac{\text{E}[Z - \mu]_+}{\alpha} \right\}.$$

Key to our analysis is the *dual representation* of CRMs. It is known that any CRM $\text{R} \colon \mathcal{L}^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ can be expressed as the *support function* of some closed convex set $\mathcal{Q} \subseteq \mathcal{L}^1(\Omega)$ [Rockafellar and Uryasev, 2013]; that is, for any real random variable $Z \in \mathcal{L}^1(\Omega)$, we have

$$\text{R}[Z] = \sup_{q \in \mathcal{Q}} \text{E}_P[Zq] = \sup_{q \in \mathcal{Q}} \int_\Omega Z(\omega) q(\omega) \mathrm{d}P(\omega). \quad \text{(dual representation)} \tag{1}$$

In this case, the set $\mathcal{Q}$ is called the *risk envelope* associated with the risk measure R. The risk envelope $\mathcal{Q}_\alpha$ of $\text{CVAR}_\alpha[Z]$ is given by

$$\mathcal{Q}_\alpha := \left\{ q \in \mathcal{L}^1(\Omega) \;\middle|\; \exists Q \in \mathcal{M}_P(\Omega), \; q = \frac{\mathrm{d}Q}{\mathrm{d}P} \leq \frac{1}{\alpha} \right\}, \tag{2}$$

and so substituting $\mathcal{Q}_\alpha$ for $\mathcal{Q}$ in (1) yields $\text{CVAR}_\alpha[Z]$.[3] Though the overall approach we take in this paper may be generalizable to other popular CRMs, (see Appendix B) we focus our attention on CVAR for which we derive new PAC-Bayesian and concentration bounds in terms of its natural estimator $\widehat{\text{CVAR}}_\alpha[Z]$; given i.i.d. copies of $Z_1, \ldots, Z_n$ of Z, we define

$$\widehat{\text{CVAR}}_\alpha[Z] := \widehat{\text{C}}_\alpha[Z] := \inf_{\mu \in \mathbb{R}} \left\{ \mu + \sum_{i=1}^n \frac{[Z_i - \mu]_+}{n\alpha} \right\}. \tag{3}$$

From now on, we write $\text{C}_\alpha[Z]$ and $\widehat{\text{C}}_\alpha[Z]$ for $\text{CVAR}_\alpha[Z]$ and $\widehat{\text{CVAR}}_\alpha[Z]$, respectively.

---

[1]These are precisely the properties which make coherent risk measures excellent candidates in some machine learning applications (see e.g. [Williamson and Menon, 2019] for an application to fairness)

[2]We use the convention in Brown [2007], Prashanth and Ghavamzadeh [2013], Wang and Gao [2010].

[3]The dual representation of CVAR was also leveraged in a concurrent work by Curi et al. [2020] for the purpose of efficiently optimizing CVAR objectives.

3 PAC-Bayesian Bound for the Conditional Value at Risk

In this section, we briefly describe the statistical learning setting, formulate our goal, and present our main results.

In the statistical learning setting, Z is a *loss* random variable which can be written as $Z = \ell(h, X)$, where $\ell \colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a loss function and $\mathcal{X}$ [resp. $\mathcal{H}$] is a feature [resp. hypotheses] space. The aim is to learn an hypothesis $\hat{h} = \hat{h}(X_{1:n}) \in \mathcal{H}$, or more generally a distribution $\hat{\rho} = \hat{\rho}(X_{1:n})$ over $\mathcal{H}$ (also referred to as randomized estimator), based on i.i.d. samples $X_1, \ldots, X_n$ of X which minimizes some measure of risk—typically, the expected risk $\mathrm{E}_P[\ell(\hat{\rho}, X)]$, where $\ell(\hat{\rho}, X) \coloneqq \mathrm{E}_{h \sim \hat{\rho}}[\ell(h, X)]$.

Our work is motivated by the idea of replacing this expected risk by any coherent risk measure R. In particular, if $\mathcal{Q}$ is the risk envelope associated with R, then our quantity of interest is

$$\mathrm{R}[\ell(\hat{\rho}, X)] \coloneqq \sup_{q \in \mathcal{Q}} \int_{\Omega} \ell(\hat{\rho}, X(\omega)) q(\omega) \mathrm{d}P(\omega).$$

Thus, given a consistent estimator $\widehat{\mathrm{R}}[\ell(\hat{\rho}, X)]$ of $\mathrm{R}[\ell(\hat{\rho}, X)]$ and some prior distribution ρ_0 on $\mathcal{H}$, our grand goal (which goes beyond the scope of this paper) is to bound the risk $\mathrm{R}[\ell(\hat{\rho}, X)]$ as

$$\mathrm{R}[\ell(\hat{\rho}, X)] \leq \widehat{\mathrm{R}}[\ell(\hat{\rho}, X)] + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathrm{KL}(\hat{\rho} \| \rho_0)}{n}}\right), \tag{4}$$

with high probability. Based on (3), the consistent estimator we use for $\mathrm{C}_\alpha[\ell(\hat{\rho}, X)]$ is

$$\widehat{\mathrm{C}}_\alpha[\ell(\hat{\rho}, X)] \coloneqq \inf_{\mu \in \mathbb{R}} \left\{ \mu + \sum_{i=1}^{n} \frac{[\ell(\hat{\rho}, X_i) - \mu]_+}{n\alpha} \right\}, \quad \alpha \in (0, 1). \tag{5}$$

This is in fact a consistent estimator (see e.g. [Duchi and Namkoong, 2018, Proposition 9]). As a first step towards the goal in (4), we derive a sharp PAC-Bayesian bound for the conditional value at risk, which we state now as our main theorem:

Theorem 1. *Let $\alpha \in (0, 1)$, $\delta \in (0, 1/2)$, $n \geq 2$, and $N \coloneqq \lceil \log_2(n/\alpha) \rceil$. Further, let ρ_0 be any distribution on an hypothesis set $\mathcal{H}$, $\ell \colon \mathcal{H} \times \mathcal{X} \to [0, 1]$ be a loss, and $X_1, \ldots, X_n$ be i.i.d. copies of X. Then, for any "posterior" distribution $\hat{\rho} = \hat{\rho}(X_{1:n})$ over $\mathcal{H}$, $\varepsilon_n \coloneqq \sqrt{\frac{2 \ln(N/\delta)}{\alpha n}} + \frac{\ln(N/\delta)}{3\alpha n}$, and $\mathcal{K}_n \coloneqq \mathrm{KL}(\hat{\rho} \| \rho_0) + \ln(N/\delta)$, we have, with probability at least $1 - 2\delta$.*

$$\mathrm{E}_{h \sim \hat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]] \leq \widehat{\mathrm{C}}_\alpha[\ell(\hat{\rho}, X)] + \sqrt{\frac{27 \widehat{\mathrm{C}}_\alpha[\ell(\hat{\rho}, X)] \mathcal{K}_n}{5\alpha n}} + 2\varepsilon_n \widehat{\mathrm{C}}_\alpha[\ell(\hat{\rho}, X)] + \frac{27 \mathcal{K}_n}{5n\alpha}. \tag{6}$$

Discussion of the bound. Although we present the bound for the bounded loss case, our result easily generalizes to the case where $\ell(h, X)$ is sub-Gaussian or sub-exponential, for all $h \in \mathcal{H}$. We discuss this in Section 5. Our second observation is that since $\mathrm{C}_\alpha[Z]$ is a coherent risk measure, it is convex in Z [Rockafellar and Uryasev, 2013], and so we can further bound the term $\mathrm{E}_{h \sim \hat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]]$ on the LHS of (6) from below by $\mathrm{C}_\alpha[\ell(\hat{\rho}, X)] = \mathrm{C}_\alpha[\mathrm{E}_{h \sim \hat{\rho}}[\ell(h, X)]]$. This shows that the type of guarantee we have in (6) is in general tighter than the one in (4).

Even though not explicitly done before, a PAC-Bayesian bound of the form (4) can be derived for a risk measure R using an existing technique due to McAllester [2003] as soon as, for any fixed hypothesis h, the difference $\mathrm{R}[\ell(h, X)] - \widehat{\mathrm{R}}[\ell(h, X)]$ is sub-exponential with a sufficiently fast tail decay as a function of n (see the proof of Theorem 1 in [McAllester, 2003]). While it has been shown that the difference $\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z]$ also satisfies this condition for bounded i.i.d. random variables $Z, Z_1, \ldots, Z_n$ (see *e.g.* Brown [2007], Wang and Gao [2010]), applying the technique of McAllester [2003] yields a bound on $\mathrm{E}_{h \sim \hat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]]$ (*i.e.* the LHS of (6)) of the form

$$\mathrm{E}_{h \sim \hat{\rho}}[\widehat{\mathrm{C}}_\alpha[\ell(h, X)]] + \sqrt{\frac{\mathrm{KL}(\hat{\rho} \| \rho_0) + \ln \frac{n}{\delta}}{\alpha n}}. \tag{7}$$

Such a bound is weaker than ours in two ways: **(I)** by Jensen's inequality the term $\widehat{\mathrm{C}}_\alpha[\ell(\hat{\rho}, X)]$ in our bound (defined in (5)) is always smaller than the term $\mathrm{E}_{h \sim \hat{\rho}}[\widehat{\mathrm{C}}_\alpha[\ell(h, X)]]$ in (7); and **(II)** unlike in our bound in (6), the complexity term inside the square-root in (7) does not multiply the

empirical conditional value at risk $\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]$. This means that our bound can be much smaller whenever $\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]$ is small—this is to be expected in the statistical learning setting since $\widehat{\rho}$ will typically be picked by an algorithm to minimize the empirical value $\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]$. This type of improved PAC-Bayesian bound, where the empirical error appears multiplying the complexity term inside the square-root, has been derived for the expected risk in works such as [Catoni, 2007, Langford and Shawe-Taylor, 2003, Maurer, 2004, Seeger, 2002]; these are arguably the state-of-the-art generalization bounds.

A reduction to the expected risk. A key step in the proof of Theorem 1 is to show that for a real random variable Z (not necessarily bounded) and $\alpha \in (0, 1)$, one can construct a function $g \colon \mathbb{R} \to \mathbb{R}$ such that the auxiliary variable $Y = g(Z)$ satisfies **(I)**

$$\mathrm{E}[Y] = \mathrm{E}[g(Z)] = \mathrm{C}_\alpha[Z]; \tag{8}$$

and **(II)** for i.i.d. copies $Z_{1:n}$ of Z, the i.i.d. random variables $Y_1 \coloneqq g(Z_1), \ldots, Y_n \coloneqq g(Z_n)$ satisfy

$$\frac{1}{n} \sum_{i=1}^n Y_i \le \widehat{\mathrm{C}}_\alpha[Z](1 + \epsilon_n), \quad \text{where} \ \ \epsilon_n = \widetilde{\mathcal{O}}(\alpha^{-1/2} n^{-1/2}), \tag{9}$$

with high probability. Thus, due to (8) and (9), bounding the difference

$$\mathrm{E}[Y] - \frac{1}{n} \sum_{i=1}^n Y_i, \tag{10}$$

is sufficient to obtain a concentration bound for CVAR. Since $Y_1, \ldots, Y_n$ are i.i.d., one can apply standard concentration inequalities, which are available whenever Y is sub-Gaussian or sub-exponential, to bound the difference in (10). Further, we show that whenever Z is sub-Gaussian or sub-exponential, then essentially so is Y. Thus, our method allows us to obtain concentration inequalities for $\widehat{\mathrm{C}}_\alpha[Z]$, even when Z is unbounded. We discuss this in Section 5.

4 Proof Sketch for Theorem 1

In this section, we present the key steps taken to prove the bound in Theorem 1. We organize the proof in three subsections. In Subsection 4.1, we introduce an auxiliary estimator $\widetilde{\mathrm{C}}_\alpha[Z]$ for $\mathrm{C}_\alpha[Z]$, $\alpha \in (0, 1)$, which will be useful in our analysis; in particular, we bound this estimator in terms of $\widehat{\mathrm{C}}_\alpha[Z]$ (as in (9) above, but with the LHS replaced by $\widetilde{\mathrm{C}}_\alpha[Z]$). In Subsection 4.2, we introduce an auxiliary random variable Y whose expectation equals $\mathrm{C}_\alpha[Z]$ (as in (8)) and whose empirical mean is bounded from above by the estimator $\widetilde{\mathrm{C}}_\alpha[Z]$ introduced in Subsection 4.1—this enables the reduction described at the end of Section 3. In Subsection 4.3, we conclude the argument by applying the classical Donsker-Varadhan variational formula [Csiszár, 1975, Donsker and Varadhan, 1976].

4.1 An Auxiliary Estimator for CVAR

In this subsection, we introduce an auxiliary estimator $\widetilde{\mathrm{C}}_\alpha[Z]$ of $\mathrm{C}_\alpha[Z]$ and show that it is not much larger than $\widehat{\mathrm{C}}_\alpha[Z]$. For $\alpha, \delta \in (0, 1)$, $n \in \mathbb{N}$, and $\pi \coloneqq (1, \ldots, 1)^\top / n \in \mathbb{R}^n$, define:

$$\widetilde{\mathcal{Q}}_\alpha \coloneqq \left\{ \boldsymbol{q} \in [0, 1/\alpha]^n \ : \ |\mathrm{E}_{i \sim \pi}[q_i] - 1| \le \epsilon_n \right\}, \quad \text{where} \quad \epsilon_n \coloneqq \sqrt{\frac{2 \ln \frac{1}{\delta}}{\alpha n}} + \frac{\ln \frac{1}{\delta}}{3 \alpha n}. \tag{11}$$

Using the set $\widetilde{\mathcal{Q}}_\alpha$, and given i.i.d. copies $Z_1, \ldots, Z_n$ of Z, let

$$\widetilde{\mathrm{C}}_\alpha[Z] \coloneqq \sup_{\boldsymbol{q} \in \widetilde{\mathcal{Q}}_\alpha} \frac{1}{n} \sum_{i=1}^n Z_i q_i. \tag{12}$$

In the next lemma, we give a "variational formulation" of $\widetilde{\mathrm{C}}_\alpha[Z]$, which will be key in our results:

Lemma 2. *Let* $\alpha, \delta \in (0, 1)$, $n \in \mathbb{N}$, *and* $\widetilde{\mathrm{C}}_\alpha[Z]$ *be as in* (12). *Then, for any* $Z_1, \ldots, Z_n \in \mathbb{R}$,

$$\widetilde{\mathrm{C}}_\alpha[Z] = \inf_{\mu \in \mathbb{R}} \left\{ \mu + |\mu| \epsilon_n + \frac{\mathrm{E}_{i \sim \pi}[Z_i - \mu]_+}{\alpha} \right\}, \quad \text{where } \epsilon_n \text{ as in } (11). \tag{13}$$

The proof of Lemma 2 (which is in Appendix A.1) is similar to that of the generalized Donsker-Varadhan variational formula considered in [Beck and Teboulle, 2003]. The "variational formulation" on the RHS of (13) reveals some similarity between $\widetilde{C}_\alpha[Z]$ and the standard estimator $\widehat{C}_\alpha[Z]$ defined in (3). In fact, thanks to Lemma 2, we have the following relationship between the two:

Lemma 3. *Let $\alpha, \delta \in (0,1)$, $n \in \mathbb{N}$, and $Z_1, \ldots, Z_n \in \mathbb{R}_{\geq 0}$. Further, let $Z_{(1)}, \ldots, Z_{(n)}$ be the decreasing order statistics of $Z_1, \ldots, Z_n$. Then, for ϵ_n as in (11), we have*

$$\widetilde{C}_\alpha[Z] \leq \widehat{C}_\alpha[Z] \cdot (1 + \epsilon_n); \tag{14}$$

and if $Z_1, \ldots, Z_n \in \mathbb{R}$ (not necessarily positive), then

$$\widetilde{C}_\alpha[Z] \leq \widehat{C}_\alpha[Z] + |Z_{(\lceil n\alpha \rceil)}| \cdot \epsilon_n. \tag{15}$$

The inequality in (15) will only be relevant to us in the case where Z maybe negative, which we deal with in Section 5 when we derive new concentration bounds for CVAR.

4.2 An Auxiliary Random Variable

In this subsection, we introduce a random variable Y which satisfies the properties in (8) and (9), where $Y_1, \ldots, Y_n$ are i.i.d. copies of Y (this is where we leverage the dual representation in (1)). This allows us to the reduce the problem of estimating CVAR to that of estimating an expectation.

Let $\mathcal{X}$ be an arbitrary set, and $f\colon \mathcal{X} \to \mathbb{R}$ be some fixed measurable function (we will later set f to a specific function depending on whether we want a new concentration inequality or a PAC-Bayesian bound for CVAR). Given a random variable X in $\mathcal{X}$ (arbitrary for now), we define

$$Z \coloneqq f(X) \tag{16}$$

and the auxiliary random variable:

$$Y \coloneqq Z \cdot \mathrm{E}[q_\star \mid X] = f(X) \cdot \mathrm{E}[q_\star \mid X], \quad \text{where} \quad q_\star \in \operatorname*{argmax}_{q \in \mathcal{Q}_\alpha} \mathrm{E}[Zq], \tag{17}$$

and $\mathcal{Q}_\alpha$ as in (2). In the next lemma, we show two crucial properties of the random variable Y—these will enable the reduction mentioned at the end of Section 3:

Lemma 4. *Let $\alpha, \delta \in (0,1)$ and $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathcal{X}$. Then, (I) the random variable Y in (17) and $Y_i \coloneqq Z_i \cdot \mathrm{E}[q_\star \mid X_i], i \in [n]$, where $Z_i \coloneqq f(X_i)$, are i.i.d. and satisfy $\mathrm{E}[Y] = \mathrm{E}[Y_i] = C_\alpha[Z]$, for all $i \in [n]$; and (II) with probability at least $1 - \delta$,*

$$(\mathrm{E}[q_\star \mid X_1], \ldots, \mathrm{E}[q_\star \mid X_n])^\top \in \widetilde{\mathcal{Q}}_\alpha, \quad \text{where } \widetilde{\mathcal{Q}}_\alpha \text{ is as in (11)}. \tag{18}$$

The random variable Y introduced in (17) will now be useful since due to (18) in Lemma 4, we have, for $\alpha, \delta \in (0,1)$; Z as in (16); and i.i.d. random variables $X, X_1, \ldots, X_n \in \mathcal{X}$,

$$P\left[\frac{1}{n} \sum_{i=1}^n Y_i \leq \widetilde{C}_\alpha[Z]\right] \geq 1 - \delta, \quad \text{where } Y_i = Z_i \cdot \mathrm{E}[q_\star \mid X_i] \tag{19}$$

and $\widetilde{C}_\alpha[Z]$ as in (12). We now present a concentration inequality for the random variable Y in (17); the proof, which can be found in Appendix A, is based on a version of the standard Bernstein's moment inequality [Cesa-Bianchi and Lugosi, 2006, Lemma A.5]:

Lemma 5. *Let $X, (X_i)_{i \in [n]}$ be i.i.d. random variables in $\mathcal{X}$. Further, let Y be as in (17), and $Y_i = f(X_i) \cdot \mathrm{E}[q_\star \mid X_i], i \in [n]$, with $q_\star$ as in (17). If $\{f(x) \mid x \in \mathcal{X}\} \subseteq [0,1]$, then for all $\eta \in [0, \alpha]$,*

$$\mathrm{E}\left[\exp\left(n\eta \cdot \left(\mathrm{E}_P[Y] - \frac{1}{n} \sum_{i=1}^n Y_i - \frac{\eta \kappa(\eta/\alpha)}{\alpha} C_\alpha[Z]\right)\right)\right] \leq 1, \quad \text{where } Z = f(X),$$

and $\kappa(x) \coloneqq (e^x - 1 - x)/x^2$, for $x \in \mathbb{R}$.

Lemma 5 will be our starting point for deriving the PAC-Bayesian bound in Theorem 1.

6

4.3 Exploiting the Donsker-Varadhan Formula

In this subsection, we instantiate the results of the previous subsections with $f(\cdot) \coloneqq \ell(\cdot, h)$, $h \in \mathcal{H}$, for some loss function $\ell \colon \mathcal{H} \times \mathcal{X} \to [0, 1]$; in this case, the results of Lemmas 3 and 5 hold for

$$Z = Z_h \coloneqq \ell(h, X), \tag{20}$$

for any hypothesis $h \in \mathcal{H}$. Next, we will need the following result which follows from the classical Donsker-Varadhan variational formula [Csiszár, 1975, Donsker and Varadhan, 1976]:

Lemma 6. *Let $\delta \in (0, 1)$, $\gamma > 0$ and ρ_0 be any fixed (prior) distribution over $\mathcal{H}$. Further, let $\{R_h : h \in \mathcal{H}\}$ be any family of random variables such that $\mathrm{E}[\exp(\gamma R_h)] \le 1$, for all $h \in \mathcal{H}$. Then, for any (posterior) distribution $\widehat{\rho}$ over $\mathcal{H}$, we have*

$$P\left[\mathrm{E}_{h \sim \widehat{\rho}}[R_h] \le \frac{\mathrm{KL}(\widehat{\rho} \| \rho_0) + \ln \frac{1}{\delta}}{\gamma} \right] \ge 1 - \delta.$$

In addition to Z_h in (20), define $Y_h \coloneqq \ell(h, X) \cdot \mathrm{E}[q_\star \mid X]$ and $Y_{h,i} \coloneqq \ell(h, X_i) \cdot \mathrm{E}[q_\star \mid X_i]$, for $i \in [n]$. Then, if we set $\gamma = \eta n$ and $R_h = \mathrm{E}_P[Y_h] - \sum_{i=1}^n Y_{h,i}/n - \eta \kappa(\eta/\alpha) \mathrm{C}_\alpha[Z_h]/\alpha$, Lemma 5 guarantees that $\mathrm{E}[\exp(\gamma R_h)] \le 1$, and so by Lemma 6 we get the following result:

Theorem 7. *Let $\alpha, \delta \in (0, 1)$, and $\eta \in [0, \alpha]$. Further, let $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathcal{X}$. Then, for any randomized estimator $\widehat{\rho} = \widehat{\rho}(X_{1:n})$ over $\mathcal{H}$, we have, with $\widehat{Z} \coloneqq \mathrm{E}_{h \sim \widehat{\rho}}[\ell(h, X)]$,*

$$\mathrm{E}_{h \sim \widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]] \le \widehat{\mathrm{C}}_\alpha[\widehat{Z}](1 + \epsilon_n) + \frac{\eta \kappa(\eta/\alpha) \, \mathrm{E}_{h \sim \widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]]}{\alpha} + \frac{\mathrm{KL}(\widehat{\rho} \| \rho_0) + \ln \frac{1}{\delta}}{\eta n}, \tag{21}$$

with probability at least $1 - 2\delta$ on the samples $X_1, \ldots, X_n$, where ϵ_n is defined in (11).

If we could optimize the RHS of (21) over $\eta \in [0, \alpha]$, this would lead to our desired bound in Theorem 1 (after some rearranging). However, this is not directly possible since the optimal η depends on the sample $X_{1:n}$ through the term $\mathrm{KL}(\widehat{\rho} \| \rho_0)$. The solution is to apply the result of Theorem 7 with a union bound, so that (21) holds for any estimator $\hat{\eta} = \hat{\eta}(X_{1:n})$ taking values in a carefully chosen grid $\mathcal{G}$; to derive our bound, we will use the grid $\mathcal{G} \coloneqq \{\alpha 2^{-1}, \ldots, \alpha 2^{-N} \mid N \coloneqq \lceil 1/2 \log_2(n/\alpha) \rceil\}$. From this point, the proof of Theorem 1 is merely a mechanical exercise of rearranging (21) and optimizing $\hat{\eta}$ over $\mathcal{G}$, and so we postpone the details to Appendix A.

5 New Concentration Bounds for CVAR

In this section, we show how some of the results of the previous section can be used to reduce the problem of estimating $\mathrm{C}_\alpha[Z]$ to that of estimating a standard expectation. This will then enable us to easily obtain concentration inequalities for $\widehat{\mathrm{C}}_\alpha[Z]$ even when Z is unbounded. We note that previous works [Bhat and Prashanth, 2019, Kolla et al., 2019] used sophisticated techniques to deal with the unbounded case (sometimes achieving only sub-optimal rates), whereas we simply invoke existing concentration inequalities for empirical means thanks to our reduction.

The key results we will use are Lemmas 3 and 4, where we instantiate the latter with $\mathcal{X} = \mathbb{R}$ and $f \equiv \mathrm{id}$, in which case:

$$Y = Z \cdot \mathrm{E}[q_\star \mid Z], \quad \text{and } q_\star \in \operatorname*{argmax}_{q \in \mathcal{Q}_\alpha} \mathrm{E}[Zq]. \tag{22}$$

Together, these two lemmas imply that, for any $\alpha, \delta \in (0, 1)$, i.i.d. random variables $Z_1, \ldots, Z_n$,

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] - |Z_{(\lceil n\alpha \rceil)}| \epsilon_n \le \mathrm{E}[Y] - \frac{1}{n} \sum_{i=1}^n Y_i, \tag{23}$$

with probability at least $1 - \delta$, where ϵ_n is as in (11) and $Z_{(1)}, \ldots, Z_{(n)}$ are the decreasing order statistics of $Z_1, \ldots, Z_n \in \mathbb{R}$. Thus, getting a concentration inequality for $\widehat{\mathrm{C}}_\alpha[Z]$ can be reduced to getting one for the empirical mean $\sum_{i=1}^n Y_i/n$ of the i.i.d. random variables $Y_1, \ldots, Y_n$. Next, we show that whenever Z is a sub-exponential [resp. sub-Gaussian] random variable, essentially so is Y. But first we define what this means:

Definition 8. *Let $\mathcal{I} \subseteq \mathbb{R}$, $b > 0$, and Z be a random variable such that, for some $\sigma > 0$,*

$$\mathrm{E}[\exp\left(\eta \cdot (Z - \mathrm{E}[Z])\right)] \leq \exp\left(\eta^2 \sigma^2/2\right), \quad \forall \eta \in \mathcal{I},$$

Then, Z is (σ, b)-sub-exponential [resp. σ-sub-Gaussian] if $\mathcal{I} = (-1/b, 1/b)$ [resp. $\mathcal{I} = \mathbb{R}$].

Lemma 9. *Let $\sigma, b > 0$ and $\alpha \in (0, 1)$. Let Z be a zero-mean real random variable and let Y be as in (22). If Z is (σ, b)-sub-exponential [resp. σ-sub-Gaussian], then*

$$\mathrm{E}[\exp(\eta Y)] \leq 2 \exp(\eta^2 \sigma^2/(2\alpha^2)), \quad \forall \eta \in (-\alpha/b, \alpha/b) \quad [\text{resp. } \eta \in \mathbb{R}]. \tag{24}$$

Note that in Lemma 9 we have assumed that Z is a zero-mean random variable, and so we still need to do some work to derive a concentration inequality for $\widehat{\mathrm{C}}_\alpha[Z]$. In particular, we will use the fact that $\mathrm{C}_\alpha[Z - \mathrm{E}[Z]] = \mathrm{C}_\alpha[Z] - \mathrm{E}[Z]$ and $\widehat{\mathrm{C}}_\alpha[Z - \mathrm{E}[Z]] = \widehat{\mathrm{C}}_\alpha[Z] - \mathrm{E}[Z]$, which holds since C_α and $\widehat{\mathrm{C}}_\alpha$ are coherent risk measures, and thus translation equivariant (see Definition 14). We use this in the proof of the next theorem (which is in Appendix A):

Theorem 10. *Let $\sigma, b > 0$, $\alpha, \delta \in (0, 1)$, and ϵ_n be as in (11). If Z is a σ-sub-Gaussian random variable, then with $\mathrm{G}[Z] \coloneqq \mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z]$ and $t_n \coloneqq \left|Z_{(\lceil n\alpha \rceil)} - \mathrm{E}[Z]\right| \cdot \epsilon_n$, we have*

$$P\left[\mathrm{G}[Z] \geq t + t_n\right] \leq \delta + 2\exp(-n\alpha^2 t^2/(2\sigma^2)), \quad \forall t \geq 0; \tag{25}$$

otherwise, if Z is (σ, b)-sub-exponential random variable, then

$$P\left[\mathrm{G}[Z] \geq t + t_n\right] \leq \delta + \begin{cases} 2\exp\left(-n\alpha^2 t^2/(2\sigma^2)\right), & \text{if } 0 \leq t \leq \sigma^2/(b\alpha); \\ 2\exp\left(-n\alpha t/(2b)\right), & \text{if } t > \sigma^2/(b\alpha). \end{cases}$$

We note that unlike the recent results due to Bhat and Prashanth [2019] which also deal with the unbounded case, the constants in our concentration inequalities in Theorem 10 are explicit.

When Z is a σ-sub-Gaussian random variable with $\sigma > 0$, an immediate consequence of Theorem 10 is that by setting $t = \sqrt{2\sigma^2 \ln(2/\delta)/(n\alpha^2)}$ in (25), we get that, with probability at least $1 - 2\delta$,

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \leq \frac{\sigma}{\alpha}\sqrt{\frac{2\ln\frac{1}{\delta}}{n}} + \left|Z_{(\lceil n\alpha \rceil)} - \mathrm{E}[Z]\right| \cdot \left(\sqrt{\frac{2\ln\frac{1}{\delta}}{\alpha n}} + \frac{\ln\frac{1}{\delta}}{3\alpha n}\right). \tag{26}$$

A similar inequality holds for the sub-exponential case. We note that the term $\left|Z_{(\lceil n\alpha \rceil)} - \mathrm{E}[Z]\right|$ in (26) can be further bounded from above by

$$\frac{n\alpha}{\lfloor n\alpha \rfloor}\widehat{\mathrm{C}}_\alpha[Z] - \mathrm{E}_{\widehat{P}_n}[Z] + \left|\mathrm{E}[Z] - \mathrm{E}_{\widehat{P}_n}[Z]\right|. \tag{27}$$

This follows from the triangular inequality and facts that $\widehat{\mathrm{C}}_\alpha[Z] \geq \frac{1}{n\alpha}\sum_{i=1}^{\lfloor n\alpha \rfloor} Z_{(i)} \geq \frac{\lfloor n\alpha \rfloor}{n\alpha}Z_{(\lceil n\alpha \rceil)}$ (see *e.g.* Lemma 4.1 in Brown [2007]), and $\widehat{\mathrm{C}}_\alpha[Z] \geq \mathrm{E}_{\widehat{P}_n}[Z]$ [Ahmadi-Javid, 2012]. The remaining term $\left|\mathrm{E}_P[Z] - \mathrm{E}_{\widehat{P}_n}[Z]\right|$ in (27) which depends on the unknown P can be bounded from above using another concentration inequality.

Generalization bounds of the form (4) for unbounded but sub-Gaussian or sub-exponential $\ell(h, X)$, $h \in \mathcal{H}$, can be obtained using the PAC-Bayesian analysis of [McAllester, 2003, Theorem 1] and our concentration inequalities in Theorem 10. However, due to the fact that α is squared in the argument of the exponentials in these inequalities (which is also the case in the bounds of Bhat and Prashanth [2019], Kolla et al. [2019]) the generalization bounds obtained this way will have the α outside the square-root "complexity term"—unlike our bound in Theorem 1.

We conjecture that the dependence on α in the concentration bounds of Theorem 10 can be improved by swapping α^2 for α in the argument of the exponentials; in the sub-Gaussian case, this would move α inside the square-root on the RHS of (26). We know that this is at least possible for bounded random variables as shown in Brown [2007], Wang and Gao [2010]. We now recover this fact by presenting a new concentration inequality for $\widehat{\mathrm{C}}_\alpha[Z]$ when Z is bounded using the reduction described at the beginning of this section.

Theorem 11. *Let $\alpha, \delta \in (0, 1)$, and $Z_{1:n}$ be i.i.d. rvs in $[0, 1]$. Then, with probability at least $1 - 2\delta$,*

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \leq \sqrt{\frac{12\mathrm{C}_\alpha[Z]\ln\frac{1}{\delta}}{5\alpha n}} \vee \frac{3\ln\frac{1}{\delta}}{\alpha n} + \mathrm{C}_\alpha[Z]\left(\sqrt{\frac{2\ln\frac{1}{\delta}}{\alpha n}} + \frac{\ln\frac{1}{\delta}}{3\alpha n}\right). \tag{28}$$

The proof is in Appendix A. The inequality in (28) essentially replaces the range of the random variable Z typically present under the square-root in other concentration bounds [Brown, 2007, Wang and Gao, 2010] by the smaller quantity $C_\alpha[Z]$. The concentration bound (28) is not immediately useful for computational purposes since its RHS depends on $C_\alpha[Z]$. However, it is possible to rearrange this bound so that only the empirical quantity $\widehat{C}_\alpha[Z]$ appears on the RHS of (28) instead of $C_\alpha[Z]$; we provide the means to do this in Lemma 13 in the appendix.

6 Conclusion and Future Work

In this paper, we derived a first PAC-Bayesian bound for CVAR by reducing the task of estimating CVAR to that of merely estimating an expectation (see Section 4). This reduction then made it easy to obtain concentration inequalities for CVAR (with explicit constants) even when the random variable in question is unbounded (see Section 5).

We note that the only steps in the proof of our main bound in Theorem 1 that are specific to CVAR are Lemmas 2 and 3, and so the question is whether our overall approach can be extended to other coherent risk measures to achieve (4).

In Appendix B, we discuss how our results may be extended to a rich class of coherent risk measures known as φ-entropic risk measures. These CRMs are often used in the context of robust optimization Namkoong and Duchi [2017], and are perfect candidates to consider next in the context of this paper.

Broader Impact

Coherent risk measures (including conditional value at risk) have been gaining significant traction in the machine learning community recently, as they allow for capturing in a much richer way the behaviour and performance of algorithms' outputs. This comes at the expense of a much harder theoretical analysis and such measures are not supported by as many guarantees than the traditional mean risk (expectation of the loss). We provide in this paper one of the few generalisation bounds for CVAR and we believe this will shed light on the advantages of using CVAR in machine learning. We intend our contributions to be of prime interest to theoreticians, but also to practitioners.

Acknowledgments and Disclosure of Funding

References

Karim T. Abou-Moustafa and Csaba Szepesvári. An exponential tail bound for lq stable learning rules. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pages 31–63. PMLR, 2019. URL http://proceedings.mlr.press/v98/abou-moustafa19a.html.

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546, 1953.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Sanjay P Bhat and LA Prashanth. Concentration of risk measures: A Wasserstein distance approach. In *Advances in Neural Information Processing Systems*, pages 11739–11748, 2019.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer School on Machine Learning*, pages 208–240. Springer, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *CoRR*, abs/1910.07833, 2019. URL http://arxiv.org/abs/1910.07833.

David B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Lecture Notes-Monograph Series. IMS, 2007.

Alain Celisse and Benjamin Guedj. Stability revisited: new generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412*, 2016.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Youhua (Frank) Chen, Minghui Xu, and Zhe George Zhang. A Risk-Averse Newsvendor Model Under the CVaR Criterion. *Operations Research*, 57(4):1040–1044, 2009. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/25614814.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR Optimization in MDPs. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3509–3517. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5246-algorithms-for-cvar-optimization-in-mdps.pdf.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

Sebastian Curi, Kfir. Y. Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning, 2020.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time — III. *Communications on pure and applied Mathematics*, 29(4):389–461, 1976.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, pages 643–669, 1961.

Benjamin Guedj. A Primer on PAC-Bayesian Learning. *arXiv*, 2019. URL https://arxiv.org/abs/1901.05353.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science*, 4(11), 2017.

Ravi Kumar Kolla, LA Prashanth, Sanjay P Bhat, and Krishna Jagannathan. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47(1): 16–20, 2019.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings COLT 2009*, 2009.

David A. McAllester. PAC-Bayesian Stochastic Model Selection. In *Machine Learning*, volume 51, pages 5–21, 2003.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected Bernstein Inequality. In *Advances in Neural Information Processing Systems*, pages 12180–12191, 2019.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 799–806, 2010.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980, 2017.

Georg Ch. Pflug. *Some Remarks on the Value-at-Risk and the Conditional Value-at-Risk*, pages 272–281. Springer US, Boston, MA, 2000. ISBN 978-1-4757-3150-7. doi: 10.1007/978-1-4757-3150-7_15. URL https://doi.org/10.1007/978-1-4757-3150-7_15.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org, 2017.

LA Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in neural information processing systems*, pages 252–260, 2013.

R Tyrrell Rockafellar and Stan Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1-2):33–53, 2013.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

Akiko Takeda and Takafumi Kanamori. A robust approach based on conditional value-at-risk measure to statistical learning problems. *European Journal of Operational Research*, 198(1): 287 – 296, 2009. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2008.07.027. URL http://www.sciencedirect.com/science/article/pii/S0377221708005614.

Akiko Takeda and Masashi Sugiyama. ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on Machine learning*, pages 1056–1063, 2008.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Philip Thomas and Erik Learned-Miller. Concentration inequalities for conditional value at risk. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6225–6233, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Ilya O. Tolstikhin and Yevgeny Seldin. Pac-bayes-empirical-bernstein inequality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 109–117, 2013. URL http://papers.nips.cc/paper/4903-pac-bayes-empirical-bernstein-inequality.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Ying Wang and Fuqing Gao. Deviation inequalities for an estimator of the conditional value-at-risk. *Operations Research Letters*, 38(3):236–239, 2010.

Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797, 2019.

A Proofs

A.1 Proof of Lemma 2

Proof Let $\varphi(\cdot) \coloneqq \iota_{[0,1/\alpha]}(\cdot)$, where for a set $\mathcal{C} \subseteq \mathbb{R}$, $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$; and $+\infty$ otherwise. From (12), we have that $\widetilde{\mathrm{C}}_\alpha[Z]$ is equal to

$$\mathcal{P} \coloneqq \sup_{\boldsymbol{q}:|\mathrm{E}_{i\sim\pi}[q_i]-1|\leq\epsilon_n} \mathrm{E}_{i\sim\pi}[Z_i q_i - \varphi(q_i)], \tag{29}$$

where we recall that $\pi = (1,\ldots,1)^\top/n \in \mathbb{R}^n$. The Lagrangian dual $\mathcal{D}$ of (29) is given by

$$
\begin{aligned}
\mathcal{D} &\coloneqq \inf_{\eta,\gamma\geq 0} \left\{ \eta - \gamma + (\eta+\gamma)\epsilon_n + \sup_{\boldsymbol{q}:0\leq q_i\leq 1/\alpha, i\in[n]} \{\mathrm{E}_{i\sim\pi}[(Z_i-\eta+\gamma)q_i - \varphi(q_i)]\} \right\}, \\
&= \inf_{\eta,\gamma\geq 0} \left\{ \eta - \gamma + (\eta+\gamma)\epsilon_n + \mathrm{E}_{i\sim\pi}\left[\sup_{0\leq x\leq 1/\alpha} \{(Z_i-\eta+\gamma)x - \varphi(x)\} \right] \right\}, \\
&= \inf_{\eta,\gamma\geq 0} \left\{ \eta - \gamma + (\eta+\gamma)\epsilon_n + \mathrm{E}_{i\sim\pi}[\varphi^\star(Z_i-\eta+\gamma)] \right\}, \tag{30} \\
&= \inf_{\mu\in\mathbb{R}} \left\{ \mu + |\mu|\epsilon_n + \mathrm{E}_{i\sim\pi}[\varphi^\star(Z_i-\mu)] \right\}, \tag{31}
\end{aligned}
$$

where (30) is due to $\{x \in \mathbb{R} \mid \varphi(x) < +\infty\} = [0,1/\alpha]$, and (31) follows by setting $\mu \coloneqq \eta - \gamma$ and noting that the inf in (30) is always attained at a point $(\eta,\gamma) \in \mathbb{R}^2_{\geq 0}$ satisfying $\eta \cdot \gamma = 0$, in which case $\eta + \gamma = |\mu|$; this is true because by the positivity of ϵ_n, if $\eta,\gamma > 0$, then $(\eta+\gamma)\epsilon_n$ can always be made smaller while keeping the difference $\eta - \gamma$ fixed. Finally, since the primal problem is feasible—$\boldsymbol{q} = \pi$ is a feasible solution—there is no duality gap (see the proof of [Beck and Teboulle, 2003, Theorem 4.2]), and thus the RHS of (31) is equal to $\mathcal{P}$ in (29). The proof is concluded by noting that the Fenchel dual of φ satisfies $\varphi^\star(z) = 0 \vee (z/\alpha)$, for all $z \in \mathbb{R}$. ∎

A.2 Proof of Lemma 3

Proof Let $\widehat{\mu}$ be the argmin in $\mu \in \mathbb{R}$ of the RHS of (3). By Lemma 2, we have

$$
\begin{aligned}
\widetilde{\mathrm{C}}_\alpha[Z] &= \inf_{\mu\in\mathbb{R}} \left\{ \mu + |\mu|\epsilon_n + \frac{\mathrm{E}_{i\sim\pi}[Z_i-\mu]_+}{\alpha} \right\}, \\
&\leq \widehat{\mu} + |\widehat{\mu}|\epsilon_n + \frac{\mathrm{E}_{i\sim\pi}[Z_i-\widehat{\mu}]_+}{\alpha}, \\
&= \widehat{\mathrm{C}}_\alpha[Z] + |\widehat{\mu}|\epsilon_n. \quad \text{(by definition of } \widehat{\mu}) \tag{32}
\end{aligned}
$$

The inequality in (15) follows from (32) and the fact that $\widehat{\mu} = Z_{(\lceil n\alpha\rceil)}$ (see proof of [Brown, 2007, Proposition 4.1]).

Now we show (14) under the assumption that $Z_i \geq 0$, for all $i \in [n]$. Note that by definition $\widehat{\mathrm{C}}_\alpha[Z] = \widehat{\mu} + \frac{1}{\alpha}\mathrm{E}_{i\sim\pi}[Z_i-\widehat{\mu}]_+$, and so $\widehat{\mu} \leq \widehat{\mathrm{C}}_\alpha[Z]$. Furthermore, since $\alpha \in (0,1)$ and $Z_i \geq 0$, for $i \in [n]$, the RHS of (3) is a decreasing function of μ on $]-\infty,0]$, and thus $\widehat{\mu} \geq 0$ (since $\widehat{\mu}$ is the minimizer of (3)). Combining the fact that $0 \leq \widehat{\mu} \leq \widehat{\mathrm{C}}_\alpha[Z]$ with (32) completes the proof. ∎

A.3 Proof of Lemma 4

Proof The first claim follows by the fact that $X_i, i \in [n]$, are i.i.d., and an application of the total expectation theorem. Now for the second claim, let $\Delta \coloneqq |\mathrm{E}_{\widehat{P}_n}[q_\star \mid X] - 1|$. Since $q_\star$ is a density, the total expectation theorem implies

$$\Delta = |\mathrm{E}_{\widehat{P}_n}[q_\star \mid X] - \mathrm{E}[\mathrm{E}[q_\star \mid X]]|,$$

13

and so by Bennett's inequality (see *e.g.* Theorem 3 in Maurer and Pontil [2009]) applied to the random variable $\mathrm{E}[q_\star \mid X]$, we get that, with probability at least $1 - \delta$,

$$\Delta \le \sqrt{\frac{2\mathrm{VAR}[\mathrm{E}[q_\star \mid X]]\ln\frac{1}{\delta}}{n}} + \frac{\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{3n},$$

$$\le \sqrt{\frac{2\,\mathrm{E}[\mathrm{E}[q_\star \mid X]^2]\ln\frac{1}{\delta}}{n}} + \frac{\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{3n},$$

$$\le \sqrt{\frac{2\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{n}} + \frac{\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{3n},$$

where the last inequality follows by the fact that $\mathrm{E}[\mathrm{E}[q_\star \mid X]^2] \le \mathrm{E}[\mathrm{E}[q_\star \mid X]] \cdot \|\mathrm{E}[q_\star \mid X]\|_\infty = \|\mathrm{E}[q_\star \mid X]\|_\infty$, which holds since $\mathrm{E}[q_\star \mid X] \ge 0$ and $\mathrm{E}[\mathrm{E}[q_\star \mid X]] = \mathrm{E}[q_\star] = 1$. The proof is concluded by the facts that $\|\mathrm{E}[q_\star \mid X]\|_\infty \le \|q_\star\|_\infty$ (by Jensen's inequality); $\|q\|_\infty \le 1/\alpha$, for all $q \in \mathcal{Q}_\alpha$ by definition; and $q_\star \in \mathcal{Q}_\alpha$. ■

A.4 Proof of Lemma 5

We need the following lemma in the proof of Lemma 5:

Lemma 12. *Let* $S, S_1, \ldots, S_n$ *be i.i.d. random variable such that* $S \in [0, B]$, $B > 0$. *We have,*

$$\mathrm{E}_P\left[\exp\left(n\eta\,\mathrm{E}_P[S] - \eta\sum_{i=1}^n S_i - n\eta^2\kappa(\eta B) \cdot \mathrm{E}_P[S^2]\right)\right] \le 1, \tag{33}$$

for all $\eta \in [0, 1/B]$, *where* $\kappa(\eta) \coloneqq (e^\eta - \eta - 1)/\eta^2$.

Proof The desired bound follows by the version of Bernstein's moment inequality in [Cesa-Bianchi and Lugosi, 2006, Lemma A.5] and [Mhammedi et al., 2019, Proposition 10-(b)]. ■

Proof of Lemma 5 By Lemma 4, the random variables $Y, Y_1, \ldots, Y_n$ are i.i.d., and so the result of Lemma 12 applies; this means that (33) holds for $(S, S_1, \ldots, S_n) = (Y, Y_1, \ldots, Y_n)$ and $B = b \ge \|Y\|_\infty$. Thus, to complete the proof it suffices to bound $\|Y\|_\infty$ and $\|Y\|_2^2 = \mathrm{E}[Y^2]$ from above. Starting with $\mathrm{E}[Y^2]$, and recalling that $Z = f(X) \in [0, 1]$ by assumption, we have:

$$\begin{aligned}
\mathrm{E}[Y^2] &= \mathrm{E}[Z^2 \cdot \mathrm{E}[q_\star \mid X]^2], \\
&\le \mathrm{E}[Z \cdot \mathrm{E}[q_\star \mid X]] \cdot \|Z \cdot \mathrm{E}[q_\star \mid X]\|_\infty, &\text{(Hölder)} \\
&\le \mathrm{C}_\alpha[Z] \cdot \|Z \cdot \mathrm{E}[q_\star \mid X]\|_\infty, &\text{(Lemma 4)} \\
&\le \mathrm{C}_\alpha[Z]/\alpha, &(Z \le 1,\ q_\star \le 1/\alpha)
\end{aligned}$$

where the fact that $q_\star \le 1/\alpha$ follows simply from $q_\star \in \mathcal{Q}_\alpha$ and the definition of $\mathcal{Q}_\alpha$. We also have

$$\begin{aligned}
\|Y\|_\infty = \|Z \cdot \mathrm{E}[q_\star \mid X]\|_\infty &\le \|Z\|_\infty \cdot \|\mathrm{E}[q_\star \mid X]\|_\infty, \\
&\le \|q_\star\|_\infty, &(Z \le 1\ \&\ \text{Jensen}) \\
&\le 1/\alpha,
\end{aligned}$$

again the last inequality follows from $q_\star \in \mathcal{Q}_\alpha$ and the definition of $\mathcal{Q}_\alpha$. ■

A.5 Proof of Theorem 7

Proof Let $h \in \mathcal{H}$ and $\alpha, \delta \in (0, 1)$, and define

$$R_h \coloneqq \mathrm{C}_\alpha[Z_h] - \frac{1}{n}\sum_{i=1}^n Y_i - \frac{\eta\kappa(\eta/\alpha)}{\alpha}\mathrm{C}_\alpha[Z_h], \tag{34}$$

where $Y_i \coloneqq \ell(h, X_i) \cdot \mathrm{E}[q_\star \mid X_i], i \in [n]$, where $q_\star$ is as in (17) with Z as in (20). By Lemma 4, $\mathrm{C}_\alpha[Z_h] = \mathrm{E}_P[Y]$, where $Y \coloneqq \ell(h, X) \cdot \mathrm{E}[q_\star \mid X]$. Thus, by Lemma 5 with $Z = Z_h$, we have $\mathrm{E}_P[\exp(n\eta R_h)] \le 1$. Applying Lemma 6 with R_h as in (34) and $\gamma = n\eta$, yields,

$$\begin{aligned}
\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]] \le{} &\frac{1}{n}\sum_{i=1}^n \ell(\widehat{\rho}, X_i) \cdot \mathrm{E}[q_\star \mid X_i] + \frac{\eta\kappa(\eta/\alpha)\,\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]]}{\alpha} \\
&+ \frac{\mathrm{KL}(\widehat{\rho}\|\rho_0) + \ln\frac{1}{\delta}}{\eta n},
\end{aligned} \tag{35}$$

with probability at least $1 - \delta$. Now invoking Lemmas 3 and 4 (in particular (18)), yields

$$\frac{1}{n} \sum_{i=1}^{n} \ell(\widehat{\rho}, X_i) \cdot \mathrm{E}[q_\star \mid X_i] \geq \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \cdot (1 + \epsilon_n).$$

with probability at least $1 - \delta$, where $\widehat{Z} := \mathrm{E}_{h \sim \widehat{\rho}}[\ell(h, X)]$. Combining this with (35) via a union bound yields the desired bound. ∎

A.6 Proof of Theorem 1

To prove Theorem 1, we will need the following lemma:

Lemma 13. *Let* $R, \widehat{R}, A, B > 0$. *If* $R \leq \widehat{R} + \sqrt{RA} + B$, *then*

$$R \leq \widehat{R} + \sqrt{\widehat{R}A} + 2B + A.$$

Proof If $R \leq \widehat{R} + \sqrt{RA} + B$, then for all $\eta > 0$,

$$R \leq \widehat{R} + \frac{\eta}{2} R + \frac{A}{2\eta} + B,$$

which after rearranging, becomes,

$$R \leq \frac{\widehat{R}}{1 - \eta/2} + \frac{A}{2\eta \cdot (1 - \eta/2)} + \frac{B}{1 - \eta/2}, \quad \text{for } \eta \notin \{0, 2\}. \tag{36}$$

The minimizer of the RHS of (36) is given by

$$\eta = \eta_\star := \frac{-A + \sqrt{A^2 + 4AB + 4A\widehat{R}}}{2(B + \widehat{R})}.$$

Plugging this η into (36), yields,

$$R \leq \widehat{R} + \frac{A}{2} + B + \frac{1}{2} \sqrt{4A\widehat{R} + A^2 + 4AB},$$

$$\leq \widehat{R} + A + 2B + \sqrt{A\widehat{R}}, \tag{37}$$

where (37) follows by the facts that $A^2 + 4AB \leq (A + 2B)^2$ and $\sqrt{4\widehat{R}A + (A + 2B)^2} \leq \sqrt{4\widehat{R}A} + A + 2B$. ∎

Proof of Theorem 1 Define the grid $\mathcal{G}$ by

$$\mathcal{G} := \left\{ 2^{-1}\alpha, \ldots, 2^{-N}\alpha \mid N := \lceil 1/2 \cdot \log_2 \tfrac{n}{\alpha} \rceil \right\},$$

and let $\hat{\eta} = \hat{\eta}(Z_{1:n}) \in \mathcal{G}$ be any estimator. Then, using the fact that $\kappa(x) \leq 3/5$, for all $x \leq 1/2$, and invoking Theorem 7 with a union bound over $\eta \in \mathcal{G}$, and $\varepsilon_n := \sqrt{\frac{2\ln\frac{N}{\delta}}{\alpha n}} + \frac{\ln\frac{N}{\delta}}{3\alpha n}$, we get that

$$\mathrm{E}_{h \sim \widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]] - \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \cdot (1 + \varepsilon_n) \leq \frac{\mathrm{KL}(\widehat{\rho} \| \rho_0) + \ln\frac{N}{\delta}}{\hat{\eta} n} + \frac{3\hat{\eta}}{5\alpha} \mathrm{E}_{h \sim \widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]], \tag{38}$$

with probability at least $1 - 2\delta$, where we recall that $\widehat{Z} = \mathrm{E}_{h \sim \widehat{\rho}}[\ell(h, X)]$. Let $\hat{\eta}$ be an estimator which satisfies

$$\hat{\eta} \in [\eta_\star \wedge (\alpha/2), \, 2\eta_\star] \cap \mathcal{G}, \quad \text{where} \quad \eta_\star := \sqrt{\frac{5\alpha \cdot (\mathrm{KL}(\widehat{\rho} \| \rho_0) + \ln\frac{N}{\delta})}{3n \, \mathrm{E}_{h \sim \widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]]}} \tag{39}$$

is the unconstrained minimizer $\hat{\eta}$ of the RHS of (38). Since the loss ℓ has range in $[0, 1]$, $\mathrm{KL}(\widehat{\rho} \| \rho_0) \geq 0$, and $(\delta, n) \in]0, 1/2[\times [2, +\infty[$, we have $\eta_\star \geq \sqrt{\alpha/n} \geq \min \mathcal{G}$. This, with the fact that $\mathcal{G}$ is in the form of a geometric progression with common ratio 2 and $\max \mathcal{G} = \alpha/2$, ensures the existence (and in fact the uniqueness) of $\hat{\eta}$ satisfying (39).

15

Case 1. Suppose that $\eta_\star \le \alpha/2$. In this case, the estimator $\hat{\eta}$ in (39) satisfies $\eta_\star \le \hat{\eta} \le 2\eta_\star$. Plugging $\hat{\eta}$ into (38) yields

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] - \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \le 3\sqrt{\frac{3\,\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]\cdot(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5\alpha n}} + \widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n.$$

By applying Lemma 13 with $R = \mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]$, $\widehat{R} = \widehat{\mathrm{C}}_\alpha[\widehat{Z}]$, $A = \frac{27(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5\alpha n}$, and $B = \widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n$, we get

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] - \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \le \sqrt{\frac{27\widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5\alpha n}} + 2\widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n$$
$$+ \frac{27(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5n\alpha}. \tag{40}$$

Case 2. Suppose now that $\eta_\star > \alpha/2$. In this case, $\hat{\eta} = \alpha/2$. Plugging this into (38) and using the fact that $\eta_\star > \alpha/2$, yields

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] - \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \le \frac{4(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{\alpha n} + \widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n. \tag{41}$$

Since $\widehat{\mathrm{C}}_\alpha[\widehat{Z}] \ge 0$ and $4 \le 27/5$, the RHS of (41) is less than the RHS of (40), which completes the proof. ∎

A.7 Proof of Lemma 9

Proof Suppose that Z is (σ, b)-sub-exponential. Then,

$$\mathrm{E}[e^{\eta Z}] \le e^{\frac{\eta^2\sigma^2}{2}}, \quad \forall|\eta| \le 1/b. \tag{42}$$

Using that $\mathrm{E}[q_\star \mid Z] \le 1/\alpha$, we get

$$|\eta Y| \le |\eta Z|/\alpha, \quad \forall\eta \in \mathbb{R}, \tag{43}$$

and so, for all $|\eta| \le \alpha/b$, we have

$$\mathrm{E}[e^{\eta Y}] \le \mathrm{E}[e^{|\eta Y|}] \overset{(43)}{\le} \mathrm{E}[e^{\frac{\eta Z}{\alpha}}] + \mathrm{E}[e^{-\frac{\eta Z}{\alpha}}] \overset{(42)}{\le} 2e^{\frac{\eta^2\sigma^2}{2\alpha^2}}.$$

When Z is σ-sub-Gaussian case, the proof is the same, except that we replace b by 0. ∎

A.8 Proof of Theorem 11

Proof Let $\mathcal{X} = [0,1]$ and $f \equiv \mathrm{id}$ be the identity map. By invoking Lemmas 3 and 5 with $Z = f(X) = X$; and using (19) (which is a consequence of Lemma 4), we get, for all $\eta \in [0, \alpha]$,

$$\mathrm{E}_P\left[\exp\left(n\eta\cdot\left(\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z](1+\epsilon_n) - \frac{\eta\kappa(\eta/\alpha)\mathrm{C}_\alpha[Z]}{\alpha}\right)\right)\right] \le 1, \tag{44}$$

with probability at least $1 - \delta$, where ϵ_n is as in (11). By adding $\mathrm{C}_\alpha[Z]\cdot\epsilon_n$ to both sides of (44) and using the fact that $\kappa(x) \le 3/5$, for all $x \le 1/2$, we get, for all $\eta \in [0, \alpha/2]$,

$$\mathrm{E}_P\left[\exp\left(n\eta\cdot\left(\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] - \left(\frac{3\eta}{5\alpha}+\epsilon_n\right)\mathrm{C}_\alpha[Z]\right)\right)\right] \le 1, \tag{45}$$

with probability at least $1 - \delta$. Let $W \coloneqq \mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] - \left(\frac{3\eta}{5\alpha}+\epsilon_n\right)\mathrm{C}_\alpha[Z]$, and note that by (45), we have

$$P[\mathrm{E}_P[\exp(n\eta W)] \le 1] \ge 1 - \delta.$$

Let $\mathcal{E}$ be the event that $\mathrm{E}_P[\exp(n\eta W)] \le 1$. With this, we have, for any $\delta \in (0, 1)$ and all $\eta \in [0, \alpha/2]$,

$$P\left[\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \ge \left(\frac{3\eta}{5\alpha} + \epsilon_n\right)\mathrm{C}_\alpha[Z] + \frac{\ln\frac{1}{\delta}}{\eta n}\right] = P\left[e^{n\eta W} \ge \frac{1}{\delta}\right]$$

$$= P\left[e^{n\eta W} \ge \frac{1}{\delta}\,\middle|\,\mathcal{E}\right] \cdot P[\mathcal{E}]$$

$$+ P\left[e^{n\eta W} \ge \frac{1}{\delta}\,\middle|\,\mathcal{E}^{\mathrm{c}}\right] \cdot (1 - P[\mathcal{E}]),$$

$$\le \delta\,\mathrm{E}[e^{n\eta W} \mid \mathcal{E}] + \delta, \tag{46}$$

$$\le 2\delta, \qquad \text{(by definition of } \mathcal{E}) \tag{47}$$

where (46) follows by Markov's inequality and (45). Now, we can re-express (47) as

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \le \left(\frac{3\eta}{5\alpha} + \epsilon_n\right)\mathrm{C}_\alpha[Z] + \frac{\ln\frac{1}{\delta}}{\eta n},$$

with probability at least $1 - 2\delta$. By setting $\eta = \sqrt{\frac{5\alpha \ln\frac{1}{\delta}}{3n\mathrm{C}_\alpha[Z]}} \wedge \alpha/2$ (which does not depend on the samples), we get

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \le \epsilon_n \mathrm{C}_\alpha[Z] + \sqrt{\frac{12\mathrm{C}_\alpha[Z] \ln\frac{1}{\delta}}{5\alpha n}} \vee \frac{3\ln\frac{1}{\delta}}{\alpha n},$$

with probability at least $1 - 2\delta$. $\blacksquare$

A.9   Proof of Theorem 10

Proof Let $\bar{Z} = Z - \mathrm{E}[Z]$. Suppose that Z is (σ, b)-sub-exponential. In this case, by Lemma 9 the random variable $Y := \bar{Z} \cdot \mathrm{E}[q_\star \mid \bar{Z}]$ satisfies (24), and so by [Wainwright, 2019, Theorem 2.19], we have

$$P\left[\mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i \ge t\right] \le \begin{cases} 2e^{-\frac{n\alpha^2 t^2}{2\sigma^2}}, & \text{if } 0 \le t \le \frac{\sigma^2}{b\alpha}; \\ 2e^{-\frac{n\alpha t}{2b}}, & \text{if } t > \frac{\sigma^2}{b\alpha}. \end{cases} \tag{48}$$

For any real random variables A, B, and C, we have $[A \ge C] \implies [A \ge B \text{ or } B \ge C]$, and so $P[A \ge C] \le P[A \ge B] + P[B \ge C]$. Applying this with $A = \mathrm{C}_\alpha[\bar{Z}] - \widehat{\mathrm{C}}_\alpha[\bar{Z}] - |\bar{Z}_{(\lceil n\alpha\rceil)}|\epsilon_n$, $B = \mathrm{E}[Y] - \sum_{i=1}^n Y_i/n$, and $C = t \in \mathbb{R}$. we get:

$$P\left[\mathrm{C}_\alpha[\bar{Z}] - \widehat{\mathrm{C}}_\alpha[\bar{Z}] - |\bar{Z}_{(\lceil n\alpha\rceil)}| \cdot \epsilon_n \ge t\right] \le P\left[\mathrm{C}_\alpha[\bar{Z}] - \widehat{\mathrm{C}}_\alpha[\bar{Z}] - |\bar{Z}_{(\lceil n\alpha\rceil)}| \cdot \epsilon_n \ge \mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i\right]$$

$$+ P\left[\mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i \ge t\right],$$

$$\le \delta + \begin{cases} 2e^{-\frac{n\alpha^2 t^2}{2\sigma^2}}, & \text{if } 0 \le t \le \frac{\sigma^2}{b\alpha}; \\ 2e^{-\frac{n\alpha t}{2b}}, & \text{if } t > \frac{\sigma^2}{b\alpha}, \end{cases} \tag{49}$$

where the last inequality follows by (48) and the fact that (23) (with Z replaced by $\bar{Z}$) holds with probability at least $1 - \delta$. Since $\mathrm{C}_\alpha[Z]$ [resp. $\widehat{\mathrm{C}}_\alpha[Z]$] is a coherent risk measure, we have $\mathrm{C}_\alpha[\bar{Z}] = \mathrm{C}_\alpha[Z] - \mathrm{E}[Z]$ [resp. $\widehat{\mathrm{C}}_\alpha[\bar{Z}] = \widehat{\mathrm{C}}_\alpha[Z] - \mathrm{E}[Z]$], and so the LHS of (49) is equal to

$$P\left[\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \ge t + |\bar{Z}_{(\lceil n\alpha\rceil)}| \cdot \epsilon_n\right].$$

This with the fact that $\bar{Z}_{(\lceil n\alpha\rceil)} = Z_{(\lceil n\alpha\rceil)} - \mathrm{E}[Z]$ completes the proof for the sub-exponential case.

When Z is σ-sub-Gaussian case, the proof is the same, except that we replace b by 0 and use the convention that $0/0 = +\infty$. $\blacksquare$

B Beyond CVAR

First, we give a formal definition of a coherent risk measure (CRM):

Definition 14. *We say that* $R: \mathcal{L}^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ *is a coherent risk measure if, for any* $Z, Z' \in \mathcal{L}^1(\Omega)$ *and* $c \in \mathbb{R}$*, it satisfies the following axioms: (Positive Homogeneity)* $R[\lambda Z] = \lambda R[Z]$*, for all* $\lambda \in (0,1)$*; (Monotonicity)* $R[Z] \leq R[Z']$ *if* $Z \leq Z'$ *a.s.; (Translation Equivariance)* $R[Z + c] = R[Z] + c$*; (Sub-additivity)* $R[Z + Z'] \leq R[Z] + R[Z']$*.*

It is known that the conditional value at risk is a member of a class of CRMs called φ-entropic risk measures Ahmadi-Javid [2012]. These CRMs are often used in the context of robust optimization Namkoong and Duchi [2017], and are perfect candidates to consider next in the context of this paper:

Definition 15. *Let* $\varphi: [0, +\infty[\to \mathbb{R} \cup \{+\infty\}$ *be a closed convex function such that* $\varphi(1) = 0$*. The* φ*-entropic risk measure with divergence level* c *is defined as*

$$\mathrm{ER}_{\varphi}^c[Z] \coloneqq \sup_{q \in \mathcal{Q}_{\varphi}^c} \mathrm{E}_P[Zq], \quad \text{where}$$

$$\mathcal{Q}_{\varphi}^c \coloneqq \left\{ q \in \mathcal{L}^1(\Omega) \,\middle|\, \begin{array}{c} \exists Q \in \mathcal{M}_P(\Omega), q = \mathrm{d}Q/\mathrm{d}P, \\ \mathrm{D}_{\varphi}(Q\|P) \leq c \end{array} \right\},$$

and $\mathrm{D}_{\varphi}(Q\|P) \coloneqq \mathrm{E}_P[\varphi(q)]$ *is the* φ*-divergence between two distributions* Q *and* P*, where* $Q \ll P$ *and* $q = \frac{\mathrm{d}Q}{\mathrm{d}P}$*.*

As mentioned above, $\mathrm{CVAR}_{\alpha}[Z]$ is a φ-entropic risk measure; in fact, it is the φ-entropic risk measure at level $c = 0$ with $\varphi(\cdot) \coloneqq \iota_{[0,1/\alpha]}(\cdot)$, where for a set $\mathcal{C} \subseteq \mathbb{R}$, $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$; and $+\infty$ otherwise Ahmadi-Javid [2012].

The natural estimator $\widehat{\mathrm{ER}}_{\varphi}^c[Z]$ of $\mathrm{ER}_{\varphi}^c[Z]$ is defined by Ahmadi-Javid [2012]

$$\widehat{\mathrm{ER}}_{\varphi}^c[Z] = \inf_{\nu > 0, \mu \in \mathbb{R}} \left\{ \mu + \nu \, \mathrm{E}_{\widehat{P}_n} \left[\varphi^{\star} \left(\frac{Z - \mu}{\nu} - c \right) \right] \right\}.$$

Extending the results of Lemmas 2 and 3 comes down to finding an auxiliary estimator $\widetilde{\mathrm{ER}}_{\varphi}^c[Z]$ of $\mathrm{ER}_{\varphi}^c[Z]$ which satisfies (as in Lemma 3) $\widetilde{\mathrm{ER}}_{\varphi}^c[Z] \leq \widehat{\mathrm{ER}}_{\varphi}^c[Z] \cdot (1 + \epsilon_n)$, for some "small" ϵ_n, and

$$\frac{1}{n} \sum_{i=1}^{n} Z_i \cdot \mathrm{E}[q_{\star} \mid Z_i] \leq \widetilde{\mathrm{ER}}_{\varphi}^c[Z],$$

with high probability, where $q_{\star} \in \mathrm{argmin}_{q \in \mathcal{Q}_{\varphi}^c} \mathrm{E}[Zq]$. The similarities between the expressions of $\widehat{\mathrm{ER}}_{\varphi}^c[Z]$ and $\widehat{\mathrm{C}}_{\alpha}[Z]$ hint that it might be possible to find such an estimator by carefully constructing a set $\widetilde{\mathcal{Q}}_{\varphi}^c$ to play the role of the $\widetilde{\mathcal{Q}}_{\alpha}$ in Section 4. We leave such investigations for future work.