PAC-Bayesian Bound for the Conditional Value at Risk

Zakaria Mhammedi Benjamin Guedj Robert C. Williamson

Australian National University UCL DATA 61 Inria inventors for the digital world

The Question

How to measure **the generalization performance** of a learning algorithms when **the risk measure** is **not** the standard expected risk?

Motivation

The **mean performance** of an algorithm in a given setting may **not** be the best objective! This includes applications where **mistakes** mean **disastrous outcomes**; this may be the case, for example, when dealing with medical, environmental, or sensitive engineering tasks.

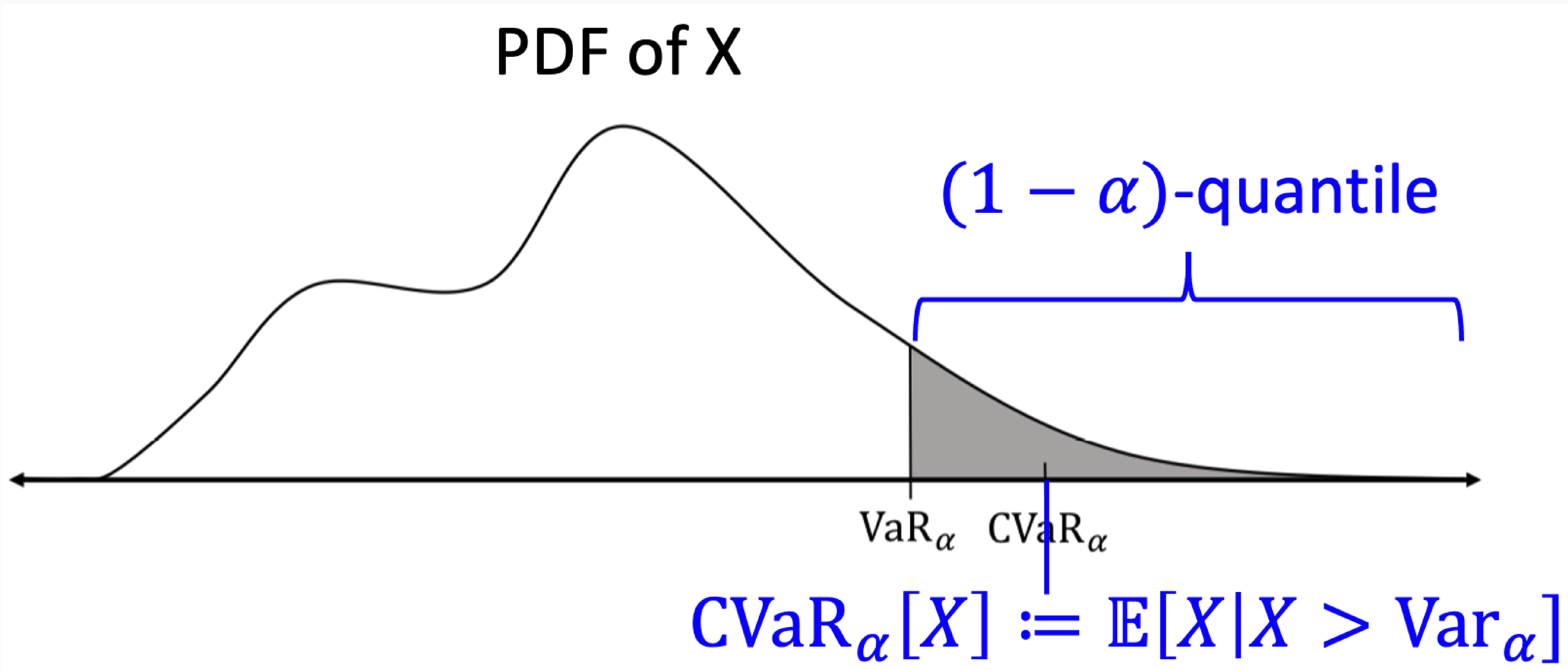
Overview of the Contributions

Motivated by the idea of protecting against the "worst" events in a learning setting, we consider the statistical learning setting, where **the objective is the CVaR of a loss** instead the expectation.

We derive a **tight PAC-Bayesian generalization bound for CVaR**.

We also derive **state-of-the-art concentration inequalities for CVaR** for bounded as well as unbounded random variables with sub-Gaussian or sub-Exponential distributions.

The Conditional Value at Risk (CVaR)



The Setting

We consider the statistical learning setting where we have

- **A bounded loss** function $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, where $\mathcal{H}$ is an hypothesis space, and $\mathcal{Z}$ is a data space. For example, the square loss: $\ell(h,(x,y)) = (y - h(x))^2, z = (x,y)$.
- **A data set** $\mathcal{D}_n := \{Z_i = (X_i, Y_i) \in \mathcal{Z} : i \in [n]\}$, where $Z_1, \dots, Z_n$ are sampled i.i.d. from an **unknown** distribution P.
- **A learning algorithm** which takes in $\mathcal{D}_n$ and outputs a distribution $\hat{\rho}$ on $\mathcal{H}$.
- **A risk measure** $\mathbb{R}\left[\mathbb{E}_{h\sim\hat{\rho}}[\ell(h,Z)]\right]$; this is typically the expected risk $\mathbb{E}_{Z\sim P}\left[\mathbb{E}_{h\sim\hat{\rho}}[\ell(h,Z)]\right]$, but we are interested in $\mathrm{CVaR}_\alpha\left[\mathbb{E}_{h\sim\hat{\rho}}[\ell(h,Z)]\right]$.

Main Contribution

Our main contribution is a PAC-Bayesian bound for algorithms which optimize the CVaR of a loss.

Theorem 1 (Informal). *Let $\alpha \in (0,1)$. Given an algorithm which outputs a distribution $\hat{\rho}$ on $\mathcal{H}$ based on i.i.d. samples $Z_{1:n}$, we have, with high probability,*

$$\mathrm{CVaR}_\alpha[\ell(\hat{\rho},Z)] \le \widehat{\mathrm{CVaR}}_\alpha + c\sqrt{\frac{\widehat{\mathrm{CVaR}}_\alpha \cdot KL}{\alpha n}} + \tilde{O}\left(\frac{KL}{\alpha n}\right), \tag{1}$$

where c is a universal constant; $KL := KL(\hat{\rho}\|\rho_0)$; ρ_0 is a prior distribution on $\mathcal{H}$ (before seeing the data); and $\widehat{\mathrm{CVaR}}_\alpha$ is a consistent estimator of $\mathrm{CVaR}_\alpha[\ell(\hat{\rho},Z)]$.

- This bound is on par with **state-of-the-art bounds** for the standard expected risk, where the square-root error term **vanishes** when the empirical risk is small.
- We also achieve the optimal dependence in the quantile level α as it appears inside the square-root error term; applying uniform convergence arguments result in α appearing **outside** this term.

New Tight Concentration Inequalities for CVaR

As a by-product of our analysis, we derive new concentration inequalities for both bound and unbounded random variables.

- For a bound random variable $Z \in [0,1]$, we have, for all $\alpha, \delta \in (0,1)$, with probability at least $1-2\delta$,

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

  This bound has the **optimal** dependence in α as it appears inside the dominant square-root terms. It also replaces the range of Z (in this case 1) inside these terms by $\mathrm{CVaR}_\alpha[Z] \le 1$.
- We also derive new concentration inequalities for the CVaR of random variables with sub-Gaussian or sub-exponential distributions (see pre-print for more details).

Key Idea: A Reduction to the Expected Risk

The key idea behind our results involves **reducing the problem of estimating** CVaR **to that of estimating the standard expectation**. In particular, we show that for a real random variable Z 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

1. $\mathbb{E}[Y] = \mathbb{E}[g(Z)] = \mathrm{CVaR}_\alpha[Z]$.
2. For i.i.d. copies $Z_{1:n}$ of Z, the i.i.d. random variables $Y_1 := g(Z_1), \dots, Y_n := g(Z_n)$ satisfy

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \le \widehat{\mathrm{CVaR}}_\alpha[Z](1+\epsilon_n), \quad \text{where } \epsilon_n = O(\alpha^{-1/2}n^{-1/2}), \tag{3}$$

   with high probability.

Thus, due to these two points, bounding the difference $\mathbb{E}[Y] - \frac{1}{n}\sum_{i=1}^{n} Y_i$, is **sufficient** to obtaining a concentration bound for CVaR. To this end, since $Y_1, \dots, Y_n$ are i.i.d., one can apply **standard concentration inequalities**, which are available whenever Y is sub-Gaussian or sub-exponential. Furthermore, the way we construct g ensures that Y **inherits** these properties from Z. (See pre-print for more details.)