Concentration Inequalities for Conditional Value at Risk

Philip S. Thomas [1] **Erik Learned-Miller** [1]

Abstract

In this paper we derive new concentration inequalities for the *conditional value at risk* (CVaR) of a random variable, and compare them to the previous state of the art (Brown, 2007). We show analytically that our lower bound is strictly tighter than Brown's, and empirically that this difference is significant. While our upper bound may be looser than Brown's in some cases, we show empirically that in most cases our bound is significantly tighter. After discussing when each upper bound is superior, we conclude with empirical results which suggest that both of our bounds will often be significantly tighter than Brown's.

1. Introduction

Many standard machine learning algorithms optimize the expected value (mean) of a random variable that quantifies performance or loss, for example, the *mean* squared error (for regression) or the *expected* discounted return (for reinforcement learning). However, for some applications the mean performance of an algorithm does not capture the desired objective. For example, for some medical applications the mean outcome should not be optimized, since doing so could increase the number of disastrous outcomes if enough mediocre outcomes are slightly improved so that the average outcome is still improved. To address this problem, *risk sensitive* methods use alternatives to mean performance that emphasize optimizing for the worst outcomes.

Two popular statistics that are used in risk sensitive machine learning, in place of the mean, are the *value at risk* (VaR) and *conditional value at risk* (CVaR). CVaR is also called the *expected shortfall* (ES), *average value at risk* (AVaR), and *expected tail loss* (ETL). Although CVaR has become popular within the operations research and machine learning communities (Kashima, 2007; Chen et al., 2009; Prashanth & Ghavamzadeh, 2013; Chow & Ghavamzadeh, 2014; Tamar et al., 2015; Pinto et al., 2017; Morimura et al., 2010), it was originally introduced in economics research as a tool for quantifying the risk associated with a portfolio (Rockafellar & Uryasev, 2000). For an introduction to CVaR and VaR see, for example, the works of Pflug (2000) and Acerbi & Tasche (2002).

In machine learning research and applications, it is often not enough to estimate the performance of a method. Such estimates usually include some amount of error, and without quantification of how much error there is in an estimate, it is not clear how much the estimate can be trusted. It is therefore important that we not only estimate the performance of an algorithm, but we provide a confidence interval along with our estimate. That is, if the goal is to optimize for the expected value of a random variable and we estimate this expected value from random samples, we should provide a confidence interval that quantifies how much our estimate can be trusted. The same applies when the goal is to optimize for CVaR: when estimating the CVaR of a random variable from samples, we should also provide a confidence interval.

Confidence intervals on estimates are often given by *concentration inequalities*: inequalities that describe how far sample statistics (e.g., estimates of performance) deviate from the statistics that they approximate (e.g., true performance). There are many concentration inequalities for the mean of a random variable (Massart, 2007; Student, 1908), which allow for tight confidence intervals around estimates of performance that are based on the mean. However, there are relatively few concentration inequalities for CVaR—to the best of our knowledge, the current state of the art for concentration inequalities for CVaR were derived by Brown (2007).

In this paper we improve upon the concentration inequalities derived by Brown (2007). Specifically, we derive two new concentration inequalities for the CVaR of a random variable—one provides a high-probability upper bound on CVaR, and the other provides a high-probability lower bound on CVaR. Not only are our concentration inequalities applicable in common settings where Brown's are not (e.g., when the random variable is only bounded above or below, but not both, or when the random variable is discrete), but

[*]Equal contribution [1]College of Information and Computer Sciences, University of Massachusetts Amherst. Correspondence to: Philip S. Thomas <pthomas@cs.umass.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

we prove that our high-probability upper bound is a strict improvement over Brown's (given practical assumptions), and show analytically that our high-probability lower-bound is often significantly tighter than Brown's.

Our inequalities have a few notable drawbacks relative to Brown's. First, they have $O(n\ln(n))$ time complexity as opposed to the $O(n)$ time complexity of Brown's. Second, they are not expressed in a form that clearly shows a confidence interval being added or subtracted from the sample CVaR (although we show later in the proof of Theorem 5 that a looser form of our bounds can be written in this form). Finally, our inequalities do not provide improved asymptotic behavior. This means that, although our inequalities should be used in place of Brown's for any actual implementations, for purely theoretical asymptotic analysis, they do not provide improvement beyond their reliance on weaker assumptions.

In the remainder of this paper we first formally define CVaR and VaR before formalizing the problem of constructing new concentration inequalities for CVaR (Section 2). We then review Brown's concentration inequalities in Section 3 before presenting our new concentration inequalities in Section 4. Finally, we provide analytic and empirical comparisons of our and Brown's inequalities in Sections 5 and 6 respectively before concluding in Section 7.

2. Problem

Let (Ω, Σ, P) be a probability space. The *conditional value at risk* (CVaR) at level $\alpha \in (0,1)$ of a random variable $X : \Omega \to \mathbb{R}$ is

$$\mathrm{CVaR}_\alpha(X) := \inf_x \left\{ x + \frac{1}{\alpha}\mathbf{E}\left[(X-x)^+\right] \right\},$$

where $x^+ := \max\{x, 0\}$ (Brown, 2007).[1] Acerbi & Tasche (2002) showed that, if X is a continuous random variable, then

$$\mathrm{CVaR}_\alpha(X) = \mathbf{E}\left[X | X \geq \mathrm{VaR}_\alpha(X)\right],$$

where $\mathrm{VaR}_\alpha(X)$ denotes the *value at risk* (VaR) of X, and is defined as

$$\mathrm{VaR}_\alpha(X) := \sup\left\{x \in \mathbb{R} | \Pr(X \geq x) \geq \alpha\right\}.$$

Figure 1 illustrates the concepts of CVaR and VaR. Due to space restrictions, hereafter we use the shorthand:

$$C(X) := \mathrm{CVaR}_\alpha(X).$$

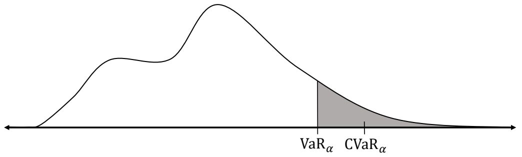

Figure 1. Illustrative example of VaR and CVaR, where the curve denotes the probability density function of a continuous random variable, with larger values denoting undesirable outcomes. VaR_α is the largest value such that at least $100\alpha\%$ of samples will be larger than it. That is, in the figure above the area of the shaded region has area α. CVaR_α is the expected value if we only consider the samples that are at least VaR_α—it is the expected value if we were to view the shaded region as a probability density function (which would require it to be normalized to integrate to one).

Let $X_1, \dots, X_n$ be n independent and identically distributed (*i.i.d.*) random variables with the same distribution as X, i.e., X_i and X are *i.i.d.* for all $i \in \{1, \dots, n\}$. In this paper we consider the problem of finding functions f and g such that the following hold (in some cases given further assumptions on X):

$$\Pr\left(\mathrm{C}(X) \geq f(\alpha, \delta, X_1(\omega), \dots, X_n(\omega))\right) \geq 1 - \delta,$$

and

$$\Pr\left(\mathrm{C}(X) \leq g(\alpha, \delta, X_1(\omega), \dots, X_n(\omega))\right) \geq 1 - \delta.$$

That is, f and g should produce high-confidence lower and upper bounds on CVaR, respectively.

3. Brown's Inequalities

Let

$$\widehat{\mathrm{C}} := \inf_{x \in \mathbb{R}} \left\{ x + \frac{1}{n\alpha}\sum_{i=1}^{n}(X_i(\omega) - x)^+ \right\}$$

denote a sample-based estimate of $\mathrm{C}(X)$ (Brown, 2007). Brown (2007) proved the following two inequalities that bound the deviation of the sample CVaR from the true CVaR with high probability.

Theorem 1. *If* $\mathrm{supp}(X) \subseteq [a, b]$ *and* X *has a continuous distribution function, then for any* $\delta \in (0, 1]$*,*

$$\Pr\left(\mathrm{C}(X) \leq \widehat{\mathrm{C}} + (b-a)\sqrt{\frac{5\ln(3/\delta)}{\alpha n}}\right) \geq 1 - \delta.$$

Theorem 2. *If* $\mathrm{supp}(X) \subseteq [a, b]$*, then for any* $\delta \in (0, 1]$*,*

$$\Pr\left(\mathrm{C}(X) \geq \widehat{\mathrm{C}} - \frac{b-a}{\alpha}\sqrt{\frac{\ln(1/\delta)}{2n}}\right) \geq 1 - \delta.$$

4. New Concentration Inequalities for CVaR

We present two new concentration inequalities for CVaR in Theorems 3 and 4 (for clarity they span both columns and are therefore presented in Figure 2).

[1] Conventions differ about whether larger X are more desirable (Pflug, 2000; Acerbi & Tasche, 2002) or less desirable (Brown, 2007), and thus whether CVaR should focus on the larger or smaller values that X can take. We adopt the notation of the most closely related prior work—that of Brown (2007)—wherein larger values of X are less desirable.

Theorem 3. *If $X_1, \dots, X_n$ are independent and identically distributed random variables and $\Pr(X_1 \le b) = 1$ for some finite b, then for any $\delta \in (0, 0.5]$,*

$$\Pr\left(\mathrm{CVaR}_\alpha(X_1) \le Z_{n+1} - \frac{1}{\alpha}\sum_{i=1}^{n}(Z_{i+1} - Z_i)\left(\frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)^+\right) \ge 1 - \delta,$$

where $Z_1, \dots, Z_n$ are the order statistics (i.e., $X_1, \dots, X_n$ sorted in ascending order), $Z_{n+1} = b$, and $x^+ := \max\{0, x\}$ for all $x \in \mathbb{R}$.

Theorem 4. *If $X_1, \dots, X_n$ are independent and identically distributed random variables and $\Pr(X_1 \ge a) = 1$ for some finite a, then for any $\delta \in (0, 0.5]$,*

$$\Pr\left(\mathrm{CVaR}_\alpha(X_1) \ge Z_n - \frac{1}{\alpha}\sum_{i=0}^{n-1}(Z_{i+1} - Z_i)\left(\min\left\{1, \frac{i}{n} + \sqrt{\frac{\ln(1/\delta)}{2n}}\right\} - (1-\alpha)\right)^+\right) \ge 1 - \delta,$$

where $Z_1, \dots, Z_n$ are the order statistics (i.e., $X_1, \dots, X_n$ sorted in ascending order), $Z_0 = a$, and where $x^+ := \max\{0, x\}$ for all $x \in \mathbb{R}$.

Figure 2. Main results, presented in a standalone fashion, and where $x^+ := \max\{0, x\}$ for all $x \in \mathbb{R}$.

Before providing proofs for these two theorems in the following subsections, we present a lemma that is used in the proofs of both theorems. Let $H : \mathbb{R} \to [0, 1]$ be the *cumulative distribution function* (CDF) for a random variable, Y. That is,

$$H(y) = \Pr(Y \le y).$$

Our first lemma provides an expression for the CVaR of Y in terms of its CDF, H:

Lemma 1. *If H is the CDF for a random variable, Y, and $H(b) = 1$ for some finite b, then*

$$C(Y) = b - \frac{1}{\alpha}\int_{-\infty}^{b}(H(y) - (1-\alpha))^+ \, dy. \tag{1}$$

Proof. Acerbi & Tasche (2002) showed that expected shortfall and CVaR are equivalent. They also gave the following expression for expected shortfall, and thus CVaR (Acerbi & Tasche, 2002, Proposition 3.2):[2]

$$C(Y) = \frac{1}{\alpha}\int_{1-\alpha}^{1} \mathrm{VaR}_\gamma(Y) \, \mathrm{d}\gamma. \tag{2}$$

This form for CVaR is depicted in Figure 3. Following the reasoning presented in Figure 3, (2) can be written as:

$$C(Y) = \frac{1}{\alpha}\left(\alpha b - \int_{-\infty}^{b} \max\{0, H(y) - (1-\alpha)\} \, \mathrm{d}y\right),$$

from which Lemma 1 follows by algebraic manipulations. □

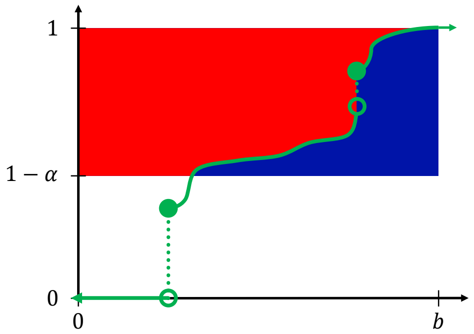

Figure 3. The green curve depicts a possible CDF, H, for a hybrid (both continuous and discrete) distribution, where $H(b) = 1$. $\mathrm{CVaR}_\alpha(Y)$, as given by (2), is $1/\alpha$ times the area of the red region (which spans horizontally from zero to the CDF, and vertically from $1-\alpha$ to 1). We will express this area as the area of the rectangle formed by the red and blue regions, which is αb, minus the area of the blue region, which is $\int_{-\infty}^{b} \max\{0, H(y) - (1-\alpha)\} \, \mathrm{d}y$.

4.1. Proof of Theorem 3

We begin by reviewing the *Dvoretzky-Kiefer-Wolfowitz* (DKW) inequality (Dvoretzky et al., 1956) using the tight constants found by Massart (1990). Let $F : \mathbb{R} \to [0, 1]$ be the CDF of X and $F_\omega : \mathbb{R} \to [0, 1]$ be the empirical CDF:

$$F_\omega(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i(\omega) \le x\}}.$$

The DKW inequality bounds the probability that the empirical CDF, F_ω, will be far from the true CDF, F, where the distance between CDFs is measured using a variant of the Kolmogorov-Smirnov statistic, $\sup_{x \in \mathbb{R}}(F_\omega(x) - F(x))$.

[2] Notice that Acerbi & Tasche use the alternate convention where larger values of X are more desirable. The expression we present can be derived by applying their expression to $-X$.

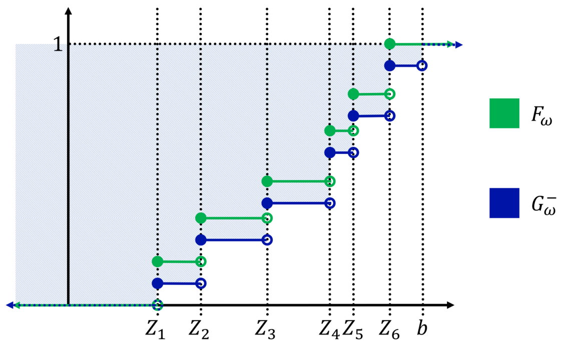

Figure 4. Illustration (not to scale) of a possible empirical CDF, F_ω, where $n = 6$. Also shown is G_ω^-, the $1-\delta$ confidence lower bound on F_ω produced by the DKW inequality. Thus, by (5) we have that with probability at least $1-\delta$, the true CDF will lie entirely within the shaded region (if $Z_1, \ldots, Z_6$ are viewed as random variables).

Property 1 (DKW Inequality)**.** *For any* $\lambda \geq \sqrt{\frac{\ln(2)}{2n}}$,

$$\Pr\left(\sup_{x\in\mathbb{R}}\left(F_\omega(x) - F(x)\right) > \lambda\right) \leq \exp\left(-2n\lambda^2\right). \quad (3)$$

Notice that the DKW inequality holds for any distribution function, F, not just continuous distribution functions (Massart, 1990, Page 1271, Comment (iii)).

Next, we will perform basic manipulations to convert the DKW inequality into the form that we will require later. Let $\delta := \exp(-2n\lambda^2)$, so that $\lambda = \sqrt{\ln(1/\delta)/(2n)}$. Thus, the requirement that $\lambda \geq \sqrt{\ln(2)/(2n)}$ becomes that $\sqrt{\ln(1/\delta)/(2n)} \geq \sqrt{\ln(2)/(2n)}$, which is the same as requiring that $\ln(1/\delta) \geq \ln(2)$, or equivalently, that $\delta \in (0, 0.5]$. Furthermore, using this definition of δ, we can rewrite (3) as:

$$\Pr\left(\sup_{x\in\mathbb{R}}\left(F_\omega(x) - F(x)\right) > \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq \delta,$$

or, by the complement rule,

$$\Pr\left(\sup_{x\in\mathbb{R}}\left(F_\omega(x) - F(x)\right) \leq \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \geq 1-\delta,$$

which implies that

$$\Pr\left(\forall x \in \mathbb{R},\, F_\omega(x) - \sqrt{\frac{\ln(1/\delta)}{2n}} \leq F(x)\right) \geq 1-\delta.$$

Furthermore, because F is a CDF, $F(x) \geq 0$ for all x, and so

$$\Pr\left(\forall x \in \mathbb{R},\, \left(F_\omega(x) - \sqrt{\frac{\ln(1/\delta)}{2n}}\right)^+ \leq F(x)\right) \geq 1-\delta. \quad (4)$$

By the assumption that $\Pr(X_1 \leq b) = 1$, $F(x) = 1$ for all $x \geq b$, and so we can tighten (4):

$$\Pr\left(\forall x \in \mathbb{R},\, G_\omega^-(x) \leq F(x)\right) \geq 1-\delta, \quad (5)$$

where

$$G_\omega^-(x) := \begin{cases} 1 & \text{if } x \geq b \\ \left(F_\omega(x) - \sqrt{\frac{\ln(1/\delta)}{2n}}\right)^+ & \text{otherwise.} \end{cases}$$

We now have, in (5), the form of the DKW inequality that we require later. A visualization of (5) is provided in Figure 4.

Intuitively, the next step of this proof is to argue that, since the true CDF of X is within the shaded blue region of Figure 4 with high probability, $\text{CVaR}_\alpha(X)$ is, with high probability, at most the maximum CVaR_α possible for random variables with CDFs that do not leave the shaded blue region. To show this formally, let $\mathcal{H}_\omega^-$ be the set of CDFs that are greater than or equal to G_ω^- at all points (the CDFs that are contained within the shaded blue region of Figure 4):

$$\mathcal{H}_\omega^- = \left\{H : \mathbb{R} \to [0,1] \middle| \forall x \in \mathbb{R},\, G_\omega^-(x) \leq H(x)\right\}.$$

Furthermore, we abuse notation and redefine C to be a function that takes as input a distribution function rather than a random variable. That is, $\text{C}(H)$ is equivalent to $\text{C}(Y)$ if Y is a random variable with CDF H.

Let u_ω denote the largest possible CVaR_α associated with a CDF that does not leave the shaded blue region of Figure 4. Formally,

$$u_\omega := \sup_{H\in\mathcal{H}_\omega^-} \text{C}(H).$$

Notice that, if $\forall x$, $G_\omega^-(x) \leq F(x)$, then $F \in \mathcal{H}_\omega^-$ and so $u_\omega \geq \text{C}(F) = \text{C}(X)$. Thus, from (5) we have that

$$\Pr\left(\text{C}(X) \leq u_\omega\right) \geq 1-\delta.$$

In the remainder of the proof of Theorem 3 we derive an expression for u_ω by showing that it is the CVaR of the distribution characterized by G_ω^-, and then leverage the fact that G_ω^- is a step function to obtain a simple expression for u_ω. By Lemma 1, we have that, for any CDF, H, where $H(b) = 1$ (which is the case for all $H \in \mathcal{H}_\omega^-$):

$$\text{C}(H) = b - \frac{1}{\alpha}\int_{-\infty}^{b} \left(H(y) - (1-\alpha)\right)^+ \text{d}y. \quad (6)$$

This expression is maximized when H is minimized (since the integral is subtracted from b). By the definition of $\mathcal{H}_\omega^-$ we have that for all $H \in \mathcal{H}_\omega^-$ and all $x \in \mathbb{R}$, $G_\omega^-(x) \leq H(x)$. So, G_ω^- is the CDF in $\mathcal{H}_\omega^-$ that maximizes C, and thus $u_w = \text{C}(G_\omega^-)$.

Since G^-_ω is a step function, its CVaR is straightforward to compute from (6) as:

$$\mathrm{C}(G^-_\omega) = \underbrace{Z_{n+1}}_{=b} - \frac{1}{\alpha}\sum_{i=1}^{n} \underbrace{(Z_{i+1} - Z_i)}_{\text{step width}} \times \underbrace{\left(\frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)^+}_{\text{step height}},$$

where × denotes scalar multiplication split across two lines.

4.2. Proof of Theorem 4

This proof follows the same steps as the proof of Theorem 3, but begins with an alternate form of the DKW inequality (Dvoretzky et al., 1956; Massart, 1990):

Property 2. *For any $\lambda \geq \sqrt{\frac{\ln(2)}{2n}}$,*

$$\Pr\left(\sup_{x\in\mathbb{R}} \left(F(x) - F_\omega(x)\right) > \lambda\right) \leq \exp\left(-2n\lambda^2\right).$$

If follows from Property 2 that

$$\Pr\left(\forall x \in \mathbb{R},\, G^+_\omega(x) \geq F(x)\right) \geq 1-\delta, \tag{7}$$

where

$$G^+_\omega(x) := \begin{cases} 0 & \text{if } x \leq a \\ \min\left\{1, F_\omega(x) + \sqrt{\frac{\ln(1/\delta)}{2n}}\right\} & \text{otherwise.}\end{cases}$$

Let

$$\mathcal{H}^+_\omega = \left\{H : \mathbb{R} \to [0,1] \middle| \forall x \in \mathbb{R},\, G^+_\omega(x) \geq H(x)\right\},$$

and $l_\omega := \inf_{H\in\mathcal{H}^+_\omega} \mathrm{C}(H)$. If $\forall x, G^+_\omega(X) \geq F(x)$, then $F \in \mathcal{H}^+_\omega$ and so $l_\omega \leq \mathrm{C}(X)$. Thus, from (7) we have that $\Pr\left(\mathrm{C}(X) \geq l_\omega\right) \geq 1-\delta$.

As a consequence of Lemma 1, for all $H \in \mathcal{H}^+_\omega$:

$$\mathrm{C}(H) \geq Z_n - \frac{1}{\alpha}\int_a^\infty \left(H(x) - (1-\alpha)\right)^+ \mathrm{d}x,$$

where the inequality holds with equality if $H = G^+_\omega$. This expression is minimized when H is maximized, and so G^+_ω is the $H \in \mathcal{H}^+_\omega$ that minimizes C. Thus, $l_\omega = \mathrm{C}(G^+_\omega)$. Since G^+_ω is a step function, like G^-_ω, it is also straightforward to compute as:

$$\mathrm{C}(G^+_\omega) = Z_n - \frac{1}{\alpha}\sum_{i=0}^{n-1}(Z_{i+1} - Z_i) \times \left(\min\left\{1, \frac{i}{n} + \sqrt{\frac{\ln(1/\delta)}{2n}}\right\} - (1-\alpha)\right)^+.$$

5. Discussion

By considering the derivations of Theorems 3 and 4, we can obtain some intuition for the complicated forms of our inequalities. At a high level, the right sides of our inequalities are the CVaR associated with a CDF (G^-_ω or G^+_ω) that is the sum of the sample CDF and a term that comes from the DKW inequality. The term from the DKW inequality scales with $1/\sqrt{n}$, and so as $n \to \infty$, this term converges to zero. Hence, as $n \to \infty$, the right sides of our inequalities become the CVaR of the sample CDF, which converges almost surely to the true CVaR.

In the remainder of this section we show that our inequalities are a significant improvement over those of Brown (2007). We begin by considering the relationship between our lower bound and Brown's, i.e., Theorem 4 and Theorem 2. First, notice the difference in requirements. Brown's inequality requires the random variable, X, to be bounded both above and below, while our inequality only requires it to be bounded below. Thus, our bound will be applicable is cases where Brown's will not—when X has no upper bound, such as if X has log-normal distribution. However, Brown's inequality allows for $\delta \in (0, 1]$, while ours only allows for $\delta \in (0, 0.5]$. Thus, Brown's inequality can be applied in cases where ours cannot. However, these cases are rare: typically we want guarantees that hold with probability greater than 0.5 (i.e., bounds that hold at least 50% of the time).

Next, consider the setting where both Brown's and our inequalities can be applied. In Theorem 5 we prove that our lower bound is a strict improvement over Brown's if $n > 2$ (for nearly all practical applications, n will be much larger than 2):

Theorem 5. *Under the conditions required by Theorems 2 4, and if $n > 3$, then Theorem 4 is strictly tighter than Theorem 2.*

Proof. For clarity, we present the proof at the end of this paper in Section 8. □

The proof of Theorem 5 provides some insight into when our inequality will be much tighter than Brown's, and when our inequality may only be slightly tighter than Brown's. Specifically, an intermediate step shows that

$$Z_{n+1} - \frac{1}{\alpha}\sum_{i=1}^{n}(Z_{i+1} - Z_i)\left(\frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)^+ \tag{8}$$

$$> \widehat{C} - \frac{1}{\alpha}\sqrt{\frac{\ln(1/\delta)}{2n}}(Z_n - Z_0),$$

which provides a strict loosening of Theorem 4 that puts it into the same form as Brown's inequality (the sample CVaR minus a term). Since $Z_n - Z_0 \leq (b - a)$, the right side of (8) is always at least as tight as the corresponding term in Theorem 2. So, whereas Brown's inequality depends on b, the upper bound on X, ours is strictly tighter than a bound that is similar to Brown's, but which depends only on the largest observed sample, Z_n. This will yield significant improvements for random variables with heavy upper tails—cases where the upper bound, b, is likely far larger than the largest observed sample, Z_n.

We now turn to comparing our upper bound to Brown's, i.e., Theorem 3 to Theorem 1. In this case, Brown's inequality requires X to be bounded both above and below, while our inequality only requires X to be bounded above with probability one. Another difference between our and Brown's upper bounds, which was not present when considering lower bounds, is that Brown's requires X to be a continuous random variable, while ours does not. This means that, again, our inequality will be applicable when Brown's is not. However, just as with the lower bounds, our bound only holds for $\delta \in (0, 0.5]$, while Brown's holds for $\delta \in (0, 1]$.

Unlike with the lower bounds, further comparison is more challenging. Neither bound is a strict improvement on the other in the settings where both are applicable. Our inequality has no dependence on the upper bound, and so for random variables with large upper bounds that are rarely realized, our inequality tends to perform better. However, our confidence interval scales with $1/\alpha$, while Brown's scales with $1/\sqrt{\alpha}$. Since $\alpha < 1$, this means that Brown's inequality has a better dependence on α.[3]

Theorems 3 and 4 can be used to create high-confidence upper and lower bounds on CVaR, and are presented in the form that would be used to do so. Although concentration inequalities are sometimes expressed in this form (Maurer & Pontil, 2009, Theorem 11), they are also sometimes expressed in an alternative form (Hoeffding, 1963). In this alternative form, a predetermined tolerance, t, is specified, and the probability of the sample statistic (sample CVaR) deviating from the target statistic (CVaR) by more than t is bounded. A limitation of Theorems 3 and 4 is that it is not clear how they could be converted into this form, which is the form originally used by Brown (2007), since the standard approach of setting t equal to the high-probability upper or lower bounds and then solving for δ in terms of t would result in the probability of the bound containing a random variable.

6. Empirical Comparisons

In order to better visualize the benefits of our new inequalities relative to those of Brown (2007), we conducted a series of empirical comparisons. The results of these comparisons are presented in Figure 8. These experiments do not test coverage rates because both our approach and Brown's provide guaranteed coverage. Instead, in these experiments we compare the upper and lower bounds produced by our approach and Brown's, for a variety of different distributions and values of n, δ, and α. Specifically, we considered seven different possible distributions for the random variable X: one log-normal, five beta, and one where $-X$ is log-normal. The *probability density functions* (PDFs) of these distributions are depicted in the top row of images in Figure 8.

The remainder of the rows show the upper and lower bounds produced by the different methods using various settings. In all cases, the left-most column denotes whether a row shows upper or lower bounds along with what label should be on the horizontal axis of all plots in that row (they were omitted to save vertical space). Also, unless otherwise specified, we always used $n = 10{,}000$ samples, $\alpha = 0.05$, and $\delta = 0.05$. The second and right-most columns correspond to distributions that are not bounded above and below. Brown's inequalities therefore are not valid for these two columns, and so their corresponding curves are omitted. Also, in all plots the blue curves correspond to bounds produced by Brown's inequalities, red curves correspond to the bounds produced by our inequalities, and the dotted black curve is the sample CVaR.

Since the quantity of primary interest is the width of the confidence intervals, rather than report the true CVaR along with our bounds, we report the *sample* CVaR. So, the gap between the dotted black line and the solid lines shows exactly the amount added or subtracted from the sample CVaR by the different bounds. Due to the large sample sizes, in most cases the sample CVaR is a good estimate of the true CVaR—except for when n is small and when using a log-normal distribution (the long tail makes it difficult to accurately estimate CVaR from samples).

Rows 2 and 3 of Figure 8 show how the upper and lower bounds change as the confidence level, δ, of the bounds is varied from 0 to 0.5. As δ becomes larger, the desired confidence level decreases and the bounds become tighter. In all cases, our inequalities were tighter.

Rows 4 and 5 of Figure 8 show how the upper and lower bounds change as the percentile, α, is varied. Notice that changing α changes the true value of $\text{CVaR}_\alpha(X)$, and so the dotted black line is not a horizontal line. This setting where α is varied is of particular interest because Brown's upper bound has a better dependence on α than our bound.

[3] Just as we loosened our lower bound to produce an inequality that is in a form similar to Hoeffding's inequality, we can loosen our upper bound to: $\Pr(C(X) \leq \widehat{C} + (b-a)\alpha^{-1}\sqrt{\ln(1/\delta)/2n} \geq 1 - \delta$. This inequality is not emphasized because it is never tighter than our bound and is often looser than Brown's.

However, in almost all cases the better dependence on α is dwarfed by our bound's better constants, dependence on δ, and lack of dependence on the true range of the random variable. One would expect Brown's upper bound to perform well when samples near the lower bound, a, are frequent so that the benefit of our inequality not depending on a is mitigated. This occurs in the distribution depicted in the second column. Furthermore, one would expect Brown's inequality to perform better when α is small. Notice that in the plot in the second column and fourth row, we show the upper bounds in exactly this case, and when α is sufficiently small, we find the single instance in these plots where Brown's inequality outperforms ours.

The sixth and seventh rows of Figure 8 show how the upper and lower bounds change as the amount of data, n, is varied. As seen elsewhere, we find that out bound remains tighter than Brown's in all cases.

7. Conclusion and Future Work

We have derived new concentration inequalities for CVaR, and shown that they compare favorably, analytically and empirically, to those of Brown (2007). The *Dvoretzky-Kiefer-Wolfowitz* (DKW) inequality (Dvoretzky et al., 1956) served as the foundation for the derivations of our new concentration inequalities. This high-level approach has also been used to produce concentration inequalities for the mean (Anderson, 1969), entropy (Learned-Miller & DeStefano, 2008), and variance (Romano & Wolf, 2002). One limitation of this approach is that the DKW inequality bounds the Kolmogorov-Smirnov statistic, which results in a uniform confidence interval around the empirical CDF. There may exist non-uniform confidence intervals that result in tighter high-probability bounds for each of these statistics (including CVaR).

8. Proof of Theorem 5

First, we provide a useful property:

Property 3. *If $x_1, \dots, x_n$ are n real numbers, $z_1, \dots, z_n$ are these numbers sorted into ascending order, z_0 is any real number, and $\alpha \in (0,1)$, then:*

$$\begin{aligned}&\inf_{x \in \mathbb{R}} \left\{ x + \frac{1}{n\alpha} \sum_{i=1}^{n} (x_i - x)^+ \right\} \\ =& z_n - \frac{1}{\alpha} \sum_{i=0}^{n-1} (z_{i+1} - z_i) \left(\frac{i}{n} - (1-\alpha) \right)^+ .\end{aligned}$$

Proof. It is known that the sample CVaR, $\hat{C}$, is equivalent to the CVaR of the distribution characterized by the sample CDF (Gao & Wang, 2011, Equation 1.4). That is, if H_n is the sample CDF built from samples $x_1, \dots, x_n$, then $\inf_{x \in \mathbb{R}} \left\{ x + \frac{1}{n\alpha} \sum_{i=1}^{n} (x_i - x)^+ \right\} = C(H_n)$, where $C(H_n)$ uses the abuse of notation described in the proof of Lemma 1. Because H_n is a step function, its CVaR, as expressed in (1), can be written as:

$$C(H_n) = \underbrace{z_n}_{b} - \frac{1}{\alpha} \sum_{i=0}^{n-1} \underbrace{(z_{i+1} - z_i)}_{\text{step width}} \underbrace{\left(\frac{i}{n} - (1-\alpha) \right)^+}_{\text{step height}} .$$

□

We now show that

$$Z_n - \frac{1}{\alpha} \sum_{i=0}^{n-1} (Z_{i+1} - Z_i) \left(\min \left\{ 1, \frac{i}{n} + s \right\} - (1-\alpha) \right)^+$$

is strictly greater than $\widehat{\mathrm{C}} - \frac{b-a}{\alpha} s$, where s is shorthand for $\sqrt{\frac{\ln(1/\delta)}{2n}}$. That is:

$$\begin{aligned}&Z_n - \frac{1}{\alpha} \sum_{i=0}^{n-1} (Z_{i+1} - Z_i) \left(\min \left\{ 1, \frac{i}{n} + s \right\} - (1-\alpha) \right)^+ \\ \overset{\text{(a)}}{>}& Z_n - \frac{1}{\alpha} \sum_{i=0}^{n-1} (Z_{i+1} - Z_i) \left(\frac{i}{n} + s - (1-\alpha) \right)^+ \\ \overset{\text{(b)}}{\geq}& Z_n - \frac{1}{\alpha} \sum_{i=0}^{n-1} (Z_{i+1} - Z_i) \left(\left(\frac{i}{n} - (1-\alpha) \right)^+ + s \right) \\ \overset{\text{(c)}}{=}& \widehat{C} - \frac{1}{\alpha} \sum_{i=0}^{n-1} (Z_{i+1} - Z_i) s \\ =& \widehat{C} - \frac{1}{\alpha} s (Z_n - Z_0) \\ \overset{\text{(d)}}{=}& \widehat{C} - \frac{Z_n - a}{\alpha} s \\ \geq& \widehat{\mathrm{C}} - \frac{b-a}{\alpha} s,\end{aligned}$$

where **(a)** holds because the $\min$ with 1 can only decrease the amount that is subtracted from Z_n, **(b)** holds because $(\ln(1/\delta)/2n)^{1/2}$ is positive, **(c)** holds by Property 3, and **(d)** holds because $Z_0 = a$. Notice that **(a)** is a strict inequality because when $i = n-1$:

$$\begin{aligned}\frac{n-1}{n} + \sqrt{\frac{\ln(1/\delta)}{2n}} \geq & \frac{n-1}{n} + \sqrt{\frac{\ln(1/0.5)}{2n}} \\ =& 1 - \frac{1}{n} + \sqrt{\frac{\ln(2)}{2}} \frac{1}{\sqrt{n}} \\ \overset{\text{(a)}}{>}& 1,\end{aligned}$$

where **(a)** holds by the assumption that $n > 3$.

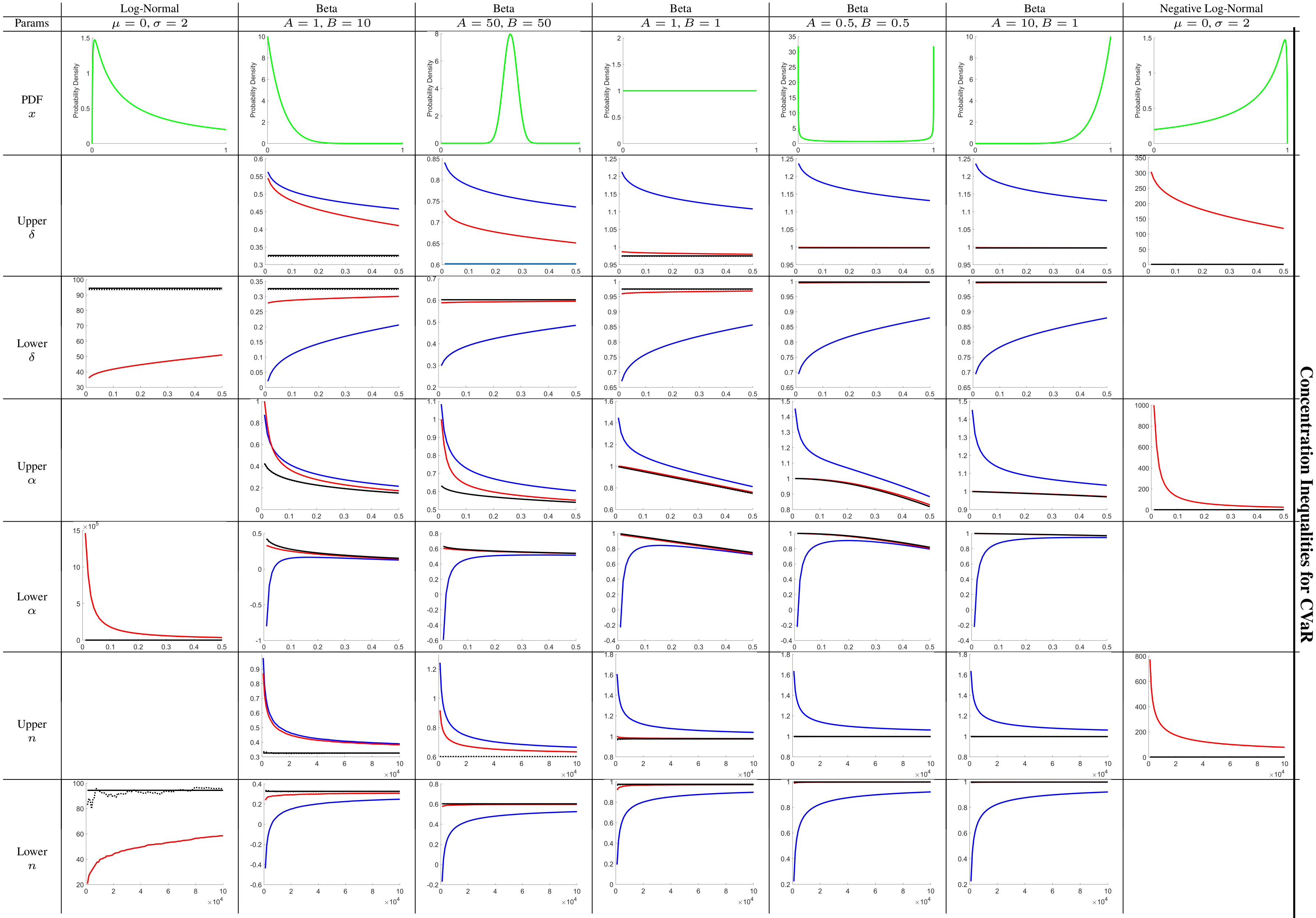

Figure 5. In all cases, red = Theorems 3 and 4, blue = Brown, dotted-black = sample CVaR, and solid-black = actual CVaR. On the left, the second line is the horizontal axis label. In some cases the actual CVaR (solid-black) obscures the sample CVaR (dotted-black) and our theorems (red).

Acknowledgements

We thank the many reviewers who provided numerous corrections, feedback, and insights that significantly improved this paper.

References

Acerbi, C. and Tasche, D. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Anderson, T. W. Confidence limits for the value of an arbitrary bounded random variable with a continuous distribution function. *Bulletin of The International and Statistical Institute*, 43:249–251, 1969.

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

Chen, Y., Xu, M., and Zhang, Z. G. A risk-averse newsvendor model under the cvar criterion. *Operations research*, 57(4):1040–1044, 2009.

Chow, Y. and Ghavamzadeh, M. Algorithms for CVaR optimization in MDPs. In *Advances in neural information processing systems*, pp. 3509–3517, 2014.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27:642–669, 1956.

Gao, F. and Wang, S. Asymptotic behavior of the empirical conditional value-at-risk. *Insurance: Mathematics and Economics*, 49(3):345–352, 2011.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Kashima, H. Risk-sensitive learning via minimization of empirical conditional value-at-risk. *IEICE TRANSACTIONS on Information and Systems*, 90(12):2043–2052, 2007.

Learned-Miller, E. and DeStefano, J. A probabilistic upper bound on differential entropy. *IEEE Transactions on Information Theory*, 54(11):5223–5230, 2008.

Massart, P. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3): 1269–1283, 1990.

Massart, P. *Concentration Inequalities and Model Selection*. Springer, 2007.

Maurer, A. and Pontil, M. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pp. 115–124, 2009.

Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., and Tanaka, T. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 799–806. Citeseer, 2010.

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

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. *arXiv preprint arXiv:1703.02702*, 2017.

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

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

Romano, J. P. and Wolf, M. Explicit nonparametric confidence intervals for the variance with guaranteed coverage. *Communications in Statistics-Theory and Methods*, 31(8): 1231–1250, 2002.

Student. The probable error of a mean. *Biometrika*, pp. 1–25, 1908.

Tamar, A., Glassner, Y., and Mannor, S. Optimizing the CVaR via sampling. In *AAAI*, pp. 2993–2999, 2015.