Risk-Constrained Thompson Sampling for CVaR Bandits

Joel Q. L. Chang$^1$, Qiuyu Zhu$^2$ and Vincent Y. F. Tan$^{1,2,3}$

$^1$Department of Mathematics, National University of Singapore, Singapore.
$^2$Institute of Operations Research and Analytics, National University of Singapore, Singapore
$^3$Department of Electrical and Computer Engineering, National University of Singapore, Singapore
joel.chang@u.nus.edu, qiuyu_zhu@u.nus.edu, vtan@nus.edu.sg

Abstract

The multi-armed bandit (MAB) problem is a ubiquitous decision-making problem that exemplifies the exploration-exploitation tradeoff. Standard formulations exclude risk in decision making. Risk notably complicates the basic reward-maximising objective, in part because there is no universally agreed definition of it. In this paper, we consider a popular risk measure in quantitative finance known as the Conditional Value at Risk (CVaR). We explore the performance of a Thompson Sampling-based algorithm CVaR-TS under this risk measure. We provide comprehensive comparisons between our regret bounds with state-of-the-art L/UCB-based algorithms in comparable settings and demonstrate their clear improvement in performance. We also include numerical simulations to empirically verify that CVaR-TS outperforms other L/UCB-based algorithms.

1 Introduction

The multi-armed bandit (MAB) problem analyses sequential decision making, where the learner has access to partial feedback. This problem is applicable to a variety of real-world applications, such as clinical trials, online advertisement, network routing, and resource allocation. In the well-known stochastic MAB setting, a player chooses among $K$ arms, each characterised by an independent reward distribution. During each period, the player plays one arm and observes a random reward from that arm, incorporates the information in choosing the next arm to select. The player repeats the process for a horizon containing $n$ periods. In each period, the player faces a dilemma whether to explore the potential value of other arms or to exploit the arm that the player believes offers the highest estimated reward.

In each of the above-mentioned applications, risk is not taken into account. In this regard, the MAB problem should be explored in a more sophisticated setting, where the player wants to maximise one’s reward while minimising one’s risk incurred, subject to a “maximum risk” condition. This paper proposes a learning algorithm which minimises the Conditional Value at Risk (CVaR) risk measure [Rockafellar and Uryasev, 2000], also known as the expected shortfall (ES).

1.1 Related Work

Various analyses of MABs involving risk measures have been undertaken. Sani et al. [2012] considered the mean-variance as their risk measure. Each arm $i$ was distributed according to a Gaussian with mean $\mu_i \in [0, 1]$ and variance $\sigma_i^2 \in [0, 1]$. The authors provided an LCB-based algorithm with accompanying regret analyses. Galichet et al. [2013] proposed the Multi-Armed Risk-Aware Bandit (MARB) algorithm with the objective of minimising the number of pulls of risky arms, using CVaR as their risk measure. Vakili and Zhao [2016] showed that the instance-dependent and instance-independent regrets in terms of the mean-variance of the reward process over a horizon $n$ are lower bounded by $\Omega((\log n))$ and $\Omega(n^{2/3})$ respectively. Sun et al. [2017] considered contextual bandits with risk constraints, and developed a meta algorithm utilizing the online mirror descent algorithm which achieves near-optimal regret in terms of minimizing the total cost. Huo and Fu [2017] studied applications of risk-aware MAB into portfolio selection, achieving a balance between risk and return. Khajonchotpanya et al. [2020] similarly used CVaR as their risk measure and proposed the CVaR-UCB algorithm, which chooses the arm with the highest gap with respect to the arm with the highest CVaR. Note that the authors associated the highest CVaR to the highest reward, while our paper considers the highest CVaR to result in the highest loss.

The paper closest to our work is that by Kagrecha et al. [2020a] who regard a large CVaR as corresponding to a large loss incurred by the user. They proposed and analysed RC-LCB and considered the scenario where users have a predetermined risk tolerance level $\tau$. Any arm whose CVaR exceeds $\tau$ is deemed infeasible. This motivates the authors to define three different regrets, each corresponding to the number of times the non-optimal arms of different classes (e.g., feasible/infeasible) were pulled. This problem formulation also includes that of Khajonchotpanya et al. [2020] in the “infeasible instance” case. Having drawn inspiration from Kagrecha et al. [2020a] in terms of the problem setup, in Section 2, we provide definitions of these classes of arms later which were originally defined in their paper. Zhu and Tan [2020] provided a Thompson sampling-based algorithm for the scenario...
in Sani et al. [2012], where the arm distributions are Gaussian or Bernoulli. Our paper seeks to explore the efficacy of Thompson sampling in the problem setting proposed by Kagrecha et al. [2020a] and to demonstrate, theoretically and empirically, the improvement of Thompson sampling over confidence bound-based approaches.

1.2 Contributions

In this paper, we focus on MABs under the CVaR risk criterion. Our contributions on are as follows:

- **CVaR-TS Algorithm**: We design CVaR-TS, an algorithm that is similar to the structure of the RC-LCB algorithm in Kagrecha et al. [2020a] but using Thompson sampling [Thompson, 1933] as explored for mean-variance bandits in Zhu and Tan [2020].

- **Comprehensive regret analyses**: We provide theoretical analyses of the algorithms and show that in a wide variety of regimes, the regret bounds outperform those of Kagrecha et al. [2020a]. As detailed in Section 5, the novel analyses improve on existing regret bounds for CVaR bandits. They also coincide with existing bounds under other risk measures (such as the mean-variance).

- **Numerical simulations**: We provide an extensive set of simulations to demonstrate that under certain regimes, our algorithms based on Thompson sampling [Thompson, 1933] consistently outperform RC-LCB Kagrecha et al. [2020a]. Furthermore, in Kagrecha et al. [2020a], due to the nature of the constants that were not explicitly defined, numerical simulations could not be implemented. By estimating the constants that they used based on explicit concentration bounds L.A. et al. [2020], we implemented RC-LCB.

This paper is structured as follows. We introduce the formulation of the CVaR MAB problem in Section 2. In Section 3, we present the CVaR-TS algorithm and demonstrate how it is a Thompson sampling generalization of Khajonchetpanya et al. [2020]. This is followed by Sections 4 and 5, where we state upper bounds on the regret and outline their proofs respectively. In Section 6, we provide numerical simulations to validate the regret bounds. Finally, we conclude our discussion in Section 7, suggesting avenues for future research. We defer detailed proofs of the theorems and statements and proofs of the corollaries to the supplementary material.

2 Problem formulation

In this section we define the CVaR MAB problem. Throughout the paper, denote $|m| = \{1, \ldots, m\}$ for any $m \in \mathbb{N}$ and $(x)^+ = \max\{0, x\}$ for $x \in \mathbb{R}$.

**Definition 1.** For any random variable $X$, given a confidence level $\alpha \in (0, 1)$, we define the Value at Risk (VaR) and Conditional Value at Risk (CVaR) metrics of $X$ by

$$v_\alpha(X) = \inf\{v \in \mathbb{R} : \mathbb{P}(X \leq v) \geq \alpha\}, \quad \text{and}$$

$$c_\alpha(X) = v_\alpha(X) + \frac{1}{1-\alpha} \mathbb{E}[(X - v_\alpha(X))^+].$$

As explained by Kagrecha et al. [2020a], $v_\alpha(X)$ is the worst case loss corresponding to a confidence level $\alpha$ (which is usually taken to be in the interval $(0.05, 1)$), where $X$ is the loss associated with a portfolio. Working with a continuous cumulative distribution function (CDF) $F_X(\cdot)$ of $X$ that is strictly increasing over its support, direct computations give $c_\alpha(X) = \mathbb{E}[X \mid X \geq v_\alpha(X)]$, and thus $c_\alpha(X)$ can be interpreted as the expected loss given that the loss exceeds $v_\alpha(X)$. We remark that CVaR is preferred to VaR as a risk measure since it is coherent [Artzner et al., 1999] and satisfies more mathematically useful properties for analysis.

In our paper, we will be working with Gaussian distributions $X \sim \mathcal{N}(\mu, \sigma^2)$, which satisfy the continuity and strict monotonicity assumptions of $F_X(\cdot)$. Letting $\Phi$ denote the cumulative distribution function of $Z \sim \mathcal{N}(0, 1)$, direct computations using (1) and (2) yield

$$c_\alpha(Z) = \frac{1}{(1-\alpha)^{1/2}} \exp \left( -\frac{1}{2} \left( \Phi^{-1}(\alpha) \right)^2 \right),$$

$$c_\alpha(X) = \mu \left( \frac{\alpha}{1-\alpha} \right) + \sigma c_\alpha(Z).$$

In the rest of the paper, we denote $c_\alpha^* = c_\alpha(Z)$. Intuitively, by (3), as $\alpha \to 1^-$, we have $c_\alpha^* \to +\infty$, i.e., the CVaR of the standard Gaussian arm increases without bound, and the arm gets riskier as the user demands higher confidence. However, since $(\sigma c_\alpha^*)/\mu \to 0$ as $\alpha \to 1^-$, we have that $1/\sigma^2 c_\alpha^*$ more quickly than $c_\alpha^*$. Thus, for $\alpha$ sufficiently close to $1^-$, $c_\alpha(X) \approx \mu (1/\sigma^2)$ and the problem reduces to a standard (risk-neutral, cost minimization or reward maximization) $K$-armed MAB problem.

Consider a $K$-armed MAB $\nu = (\nu(i))_{i \in [K]}$ and a player with a risk threshold $\tau > 0$ that represents her risk appetite. The CVaR MAB problem is played over a horizon of length $n$. Roughly speaking, our goal is to choose the arm with the lowest average loss, subject to an upper bound on the risk (measured by the CVaR) associated with the arm. For any arm $i$, the loss associated with arm $i$ has distribution $\nu(i)$ and $X(i)$ denotes a random variable with distribution $\nu(i)$. Furthermore, $\mu(i) = \mu_i$ denotes the mean of $X(i)$, and $c_\alpha(i)$ denotes the CVaR of $X(i)$. Similar to Kagrecha et al. [2020a], we define feasible and infeasible instances as follows.

**Definition 2** (Kagrecha et al. [2020a]). An instance of the risk-constrained MAB problem is defined by $(\nu, \tau)$. We denote the set of feasible arms (whose CVaR $\leq \tau$) as $\mathcal{K}_\tau = \{i \in [K] : c_\alpha(i) \leq \tau\}$. The instance $(\nu, \tau)$ is said to be feasible (resp. infeasible) if $\mathcal{K}_\tau \neq \emptyset$ (resp. $\mathcal{K}_\tau = \emptyset$).

In a feasible instance, an arm $i$ is **optimal** if $c_\alpha(i) \leq \tau$ and $\mu_i = \min_{j \in \mathcal{K}_\tau} \mu_j$. Suppose arm $i$ is optimal. Define the set $\mathcal{M} := \{i \in [K] \setminus \{1\} : \mu_i > \mu_1\}$. Arm $i$ is said to be

- a suboptimal arm if $i \in \mathcal{M} \cap \mathcal{K}_\tau$,
- an infeasible arm if $i \in \mathcal{K}_\tau$,
- a deceiver arm if $i \in \mathcal{M} \cap \mathcal{K}_\tau$.

For a suboptimal arm $i$, we define the suboptimality gap as $\Delta(i) = \mu_i - \mu_1 \geq 0$. For an infeasible arm $k$, we define the infeasibility gap as $\Delta_k(i, \alpha) = c_\alpha(i) - \tau > 0$.

In an infeasible instance, without loss of generality, set arm $1$ as an optimal arm, that is, $c_\alpha(1) = \min_{i \in [K]} c_\alpha(i)$. This is usually taken to be in the interval $(0.05, 1)$, where $X$ is the loss associated with a portfolio.
We define the risk gap for a arm $k$ that is not optimal by
\[ \Delta_{\text{risk}}(i, \alpha) = c_\alpha(i) - c_\alpha(1) > 0. \]
We remark that it is in this infeasible instance that the risk gap is defined naturally by Khajonchotpanya et al. [2020]. With the three gaps defined, we are now in the position to define three regrets.

**Definition 3.** Let $T_i,n$ denote the number of times arm $i$ was pulled in the first $n$ rounds.

1. For a feasible instance, let
   \[ K^* = \left\{ i \in [K] : c_\alpha(i) \leq \tau \text{ and } \mu_i = \min_{j \in K^*} \mu_j \right\} \]
   denote the set of optimal arms (and without loss of generality suppose $1 \in K^*$). The suboptimality regret of policy $\pi$ over $n$ rounds is
   \[ R_{\text{sub}}^n(\pi) = \sum_{i \in K^* \setminus K^*} \mathbb{E}[T_i,n] \Delta(i), \]
   and the infeasibility regret of policy $\pi$ over $n$ rounds is
   \[ R_{\text{inf}}^n(\pi) = \sum_{i \in K \setminus K^*} \mathbb{E}[T_i,n] \Delta_T(i, \alpha). \]

2. For an infeasible instance, let
   \[ K^* = \left\{ i \in [K] : c_\alpha(i) = \min_{j \in [K]} c_\alpha(j) \right\} \]
   denote the set of optimal arms (and without loss of generality suppose $1 \in K^*$). The risk regret of policy $\pi$ over $n$ rounds is
   \[ R_{\text{risk}}^n(\pi) = \sum_{i \in [K] \setminus K^*} \mathbb{E}[T_i,n] \Delta_{\text{risk}}(i, \alpha). \]

In the following, we design and analyse an algorithm that simultaneously minimizes the three regrets in (5), (6), and (7).

## 3 The CVaR-TS Algorithm

In this section, we introduce the CVaR Thompson Sampling (CVaR-TS) algorithm for Gaussian bandits. The algorithm is adapted from RC-LCB in Kagrecha et al. [2020a], but instead of choosing the arm based on the optimism in the face of uncertainty principle, the algorithm samples from the posteriors of each arm, then chooses the arm according to a multi-criterion procedure.

As is well known, a crucial step of Thompson sampling algorithms is the updating of parameters based on Bayes rule. Denote the mean and precision of the Gaussian by $\mu$ and $\phi$ respectively. If $(\mu, \phi) \sim \text{Normal-Gamma}(\mu, T, \alpha, \beta)$, then $\phi \sim \text{Gamma}(\alpha, \beta)$, and $\mu/\phi \sim \mathcal{N}(\mu, 1/(\phi T))$. Since the Normal-Gamma distribution is the conjugate prior for the Gaussian with unknown mean and variance, we use Algorithm 1 to update $(\mu, \phi)$.

We present a Thompson sampling-based algorithm to solve the CVaR Gaussian MAB problem. The player chooses a prior over the set of feasible bandits parameters for both the mean and precision. In each round $t$, for each arm $i$, the player samples a pair of parameters $(\theta_{i,t}, \kappa_{i,t})$ from the posterior distribution of arm $i$, then forms the set
\[ \hat{K}_t := \{ k \in [K] : \hat{c}_\alpha(k, t) = \theta_{k,t} \left( \frac{\alpha}{1 - \alpha} + \frac{1}{\sqrt{\kappa_{k,t}}} \right) \leq \tau \}. \]

If $\hat{K}_t$ is nonempty, i.e., it is plausible that there are some feasible arms available, choose arm $j$ if $\theta_{j,t} = \min_{k \in \hat{K}_t} \theta_{k,t}$. Otherwise, choose arm $j$ if $j = \arg \min_{k \in \hat{K}_t} \hat{c}_\alpha(k, t)$; that is, choose the least infeasible arm available. At the end of the algorithm, we also set a FeasibilityFlag that identifies the instance as feasible or not. Kagrecha et al. [2020a] provided bounds on the probability of incorrect flagging by any consistent algorithm, and we have empirically compared the errors induced by both algorithms in Section 6. CVaR-TS does not perform worse than RC-LCB in this aspect; yet regret-wise, CVaR-TS often performs much better.

**Algorithm 1** CVaR Thompson Sampling (CVaR-TS)

1. **Input:** Prior parameters $(\mu_{i,t-1}, T_{i,t-1}, \alpha_{i,t-1}, \beta_{i,t-1})$ and new sample $X_{i,t}$
2. Update the mean: $\mu_{i,t} = \frac{T_{i,t-1}}{T_{i,t-1} + 1} \mu_{i,t-1} + \frac{1}{T_{i,t-1} + 1} X_{i,t}$
3. Update the number of samples, the shape parameter, and the rate parameter:
   \[ T_{i,t} = T_{i,t-1} + 1, \alpha_{i,t} = \alpha_{i,t-1} + 1/2, \beta_{i,t} = \beta_{i,t-1} + \frac{T_{i,t-1}}{T_{i,t-1} + 1} (X_{i,t} - \mu_{i,t-1})^2. \]

4. **for** $t = 1, 2, \ldots, K$ **do**
5. **Play** arm $t$ and update $\hat{\mu}_{t,t}$
6. **end for**
7. **for** $t = K + 1, K + 2, \ldots$ **do**
8. **Sample** $\kappa_{t,t}$ from Gamma$(\alpha_{t-1}, \beta_{t-1})$
9. **Sample** $\theta_{t,t}$ from $\mathcal{N}(\hat{\mu}_{t,t-1}, 1/T_{t,t-1})$
10. **Set** $\hat{c}_\alpha(t, k) = \theta_{t,t} \left( \frac{\alpha}{1 - \alpha} + \frac{1}{\sqrt{\kappa_{t,t}}} \right)$
11. **Set** $\hat{K}_t = \{ k : \hat{c}_\alpha(k, t) \leq \tau \}$
12. **if** $\hat{K}_t \neq \emptyset$ **then**
13. **Play** arm $i(t) = \arg \min_{k \in \hat{K}_t} \hat{c}_\alpha(k, t)$ and observe loss $X_{i(t),t} \sim \mathcal{N}(\mu_{i(t),t})$
14. **end if**
15. **else**
16. **Play** arm $i(t) = \arg \min_{k \in \hat{K}_t} \hat{c}_\alpha(k, t)$ and observe loss $X_{i(t),t} \sim \mathcal{N}(\mu_{i(t),t})$
17. **end if**
18. **end for**
19. **if** $\hat{K}_t \neq \emptyset$ **then**
20. **Set** FeasibilityFlag = true
21. **else**
22. **Set** FeasibilityFlag = false
23. **end if**
Table 1: Comparison of the expected regret of CVaR-TS to those of existing CVaR MAB algorithms: (1) CVaR-LCB by Bhat and L.A. [2019, Theorem 1], (2) RC-LCB by Kagrecha et al. [2020a, Theorem 2], (3) CVaR-UCB-1 by Khajonchothpanya et al. [2020, Lemma 4], (4) CVaR-UCB-2 by Tamkin et al. [2019]. We abbreviate $1 - \alpha$ by $\beta$. The second column corresponds the expected number of pulls of a non-optimal arm $i$ over horizon $n$, which suffices to provide comparisons on the expected regret bounds. The last column states conditions under which CVaR-TS performs better than the algorithm in comparison. See Remark 3 for details.

4 Regret Bounds of CVaR-TS

We present our regret bounds in the following theorems. Most comparisons of results are with respect to Theorem 1, since a natural definition of regret induced by the CVaR risk measure is considered in Khajonchothpanya et al. [2020], Tamkin et al. [2019], Xi et al. [2020], and Soma and Yoshida [2020].

Theorem 1. Fix $\xi \in (0, 1)$. In an infeasible instance, the asymptotic expected risk regret of CVaR-TS for CVaR Gaussian bandits satisfies

$$\limsup_{n \to \infty} \frac{R_{\text{risk}}^i(\text{CVaR-TS})}{\log n} \leq \sum_{i \in [K] \setminus K_*} C_{\alpha, \xi}^i \Delta_{\text{risk}}(i, \alpha),$$

where $C_{\alpha, \xi}^i = \max \{ A_{\alpha, \xi}^i, B_{\alpha, \xi}^i \}$.

$$A_{\alpha, \xi}^i = \frac{2 \alpha^2}{\xi^2 (1 - \alpha)^2 \Delta_{\text{risk}}^2(i, \alpha)},$$

$$B_{\alpha, \xi}^i = \frac{1}{h \left( \frac{\sigma^2 \epsilon^2}{(\sigma^2 \epsilon^2 - (1 - \xi) \Delta_{\text{risk}}^2(i, \alpha))^2} \right)},$$

and

$$h(x) = \frac{1}{2} (x - 1 - \log x).$$

Remark 1. For $\alpha \to 1^-$, an appropriate choice of $\xi_{n, \alpha}$ yields $\xi_{n, \alpha} \to 1^-$ and $B_{n, \xi_{n, \alpha}} \to 0$ (see Corollary 1). Thus, the upper bound is characterised by $A_{\alpha, \xi_{n, \alpha}}$. By continuity, we obtain the regret bound involving $A_{\alpha, \xi}$ as defined in (8). We remark that this asymptotic regret bound is tighter than several existing results under certain regimes, as summarised in Table 1, which lists the upper bounds on the expected number of pulls $E[T_{i,n}]$ of a non-optimal arm $i$ over a horizon $n$ by various policies $\pi$. See Remark 3 for details.

Remark 2. For any arm $i$, $\lim_{\alpha \to 1^-} (1 - \alpha) \Delta_{\text{risk}}(i, \alpha) = \mu_i - \mu_1$, and the upper bound simplifies to $2 / (\mu_i - \mu_1)^2$. This agrees with our intuition because as $\alpha \to 1^-$, $\epsilon_i(i) = \mu_i(1 - \alpha) / \epsilon_\alpha$ is dominated by $\mu_i(1 - \alpha)$, implying that we are in the risk-neutral setting. Thus, the results are analogous to those derived for mean-variance bandits [Zhu and Tan, 2020] for the risk-neutral setting when $\rho \to +\infty$ (recall the mean-variance of arm $i$ is $\text{MV}_i = \rho \mu_i - \sigma_i^2$).

Remark 3. From Table 1, we see that CVaR-TS outperforms existing state-of-the-art CVaR MAB algorithms on Gaussian bandits under certain regimes.

1. We see that CVaR-TS outperforms CVaR-LCB [Bhat and L.A., 2019] unconditionally since $\alpha \in (0, 1)$ in Definition 1, and hence $\alpha^2 < 1 \leq 8$ trivially.

2. Our regret bound for CVaR-TS is tighter than that of RC-LCB when $\alpha^2 \leq 4/d_\sigma$, where $\sigma$ is the fixed sub-Gaussianity parameter of the bandits. We remark that Kagrecha et al. [2020a, Lemma 1] included the implicit constants $D_\sigma$ and $d_\sigma$ in their algorithm design due to an LCB-style concentration bound derived from Bhat and L.A. [2019, Corollary 1], which in turn was derived from a concentration of measure result involving the Wasserstein distance in Fournier and Guillin [2015]. Note that $d_\sigma$ is non-trivial to compute, so we estimate it to be $1 / (8 \sigma^2)$ based on an explicit concentration bound for $d_\sigma$ in L.A. et al. [2020, Theorem 3.1]. In this case, CVaR-TS outperforms RC-LCB when the sub-Gaussianity parameter of the $K$ bandits $\sigma$ satisfies $\alpha^2 \geq 8 \sigma^2$.

3. Replacing 8 in Remark 3 with $2 \leq \gamma$, where $\gamma$ is the UCB parameter in Khajonchothpanya et al. [2020] (denoted as $\alpha$ therein), yields the conclusion that CVaR-TS outperforms CVaR-UCB-1 unconditionally.

4. Our regret bound for CVaR-TS is tighter than that of CVaR-UCB-2 in Tamkin et al. [2019] when $\alpha^2 \leq 2U^2$, where $U > 0$ is any chosen upper bound on the supports of the distributions of the $K$ bandits. Notice that we do not need to assume that the arm distributions are bounded, which is different from Tamkin et al. [2019].

Theorem 2. Fix $\xi \in (0, 1)$. In a feasible instance, the asymptotic expected value regret of CVaR-TS for CVaR Gaussian bandits satisfies

$$\limsup_{n \to \infty} \frac{R_{\text{inf}}^i(\text{CVaR-TS})}{\log n} \leq \sum_{i \in K_*} D_{\alpha, \xi}^i \Delta_r(i, \alpha),$$

where $D_{\alpha, \xi}^i = \max \{ E_{\alpha, \xi}^i, F_{\alpha, \xi}^i \}$.

$$E_{\alpha, \xi}^i = \frac{2 \alpha^2}{\xi^2 (1 - \alpha)^2 \Delta_r^2(i, \alpha)},$$

and

$$F_{\alpha, \xi}^i = \frac{1}{h \left( \frac{\sigma^2 \epsilon^2}{(\sigma^2 \epsilon^2 - (1 - \xi) \Delta_r^2(i, \alpha))^2} \right)}.$$

Remark 4. For $\alpha \to 1^-$, an appropriate choice of $\xi_{n, \alpha}$ yields $\xi_{n, \alpha} \to 1^-$ and $E_{n, \xi_{n, \alpha}}^i \to 0$ (see Corollary 2). Thus, the upper bound is characterised by $E_{\alpha, \xi_{n, \alpha}}^i$. By continuity, we obtain the regret bound involving $E_{\alpha, \xi}$ as defined in (10). This is expected, since $(\epsilon_i(1), \Delta_{\text{risk}}(i, \alpha))$ can be replaced by $(\tau, \Delta_r(i, \alpha))$ analogously for infeasible arms in the infeasible case, and the computations can be similarly reused. Likewise, this regret bound is tighter than RC-LCB ($E[T_{i,n}] \leq 4 \log(2D_\sigma n^2) / (\beta^2 \Delta^2(1, \alpha) d_\sigma)$) provided that $d_\sigma \leq 4$.
Theorem 3. In a feasible instance, the asymptotic expected suboptimality regret of CVaR-TS for CVaR Gaussian bandits satisfies
\[
\limsup_{n \to \infty} \frac{R_n(CVaR-TS)}{\log n} \leq \sum_{i \in K \setminus \{\star\}} \frac{2}{\Delta(i)}. \tag{11}
\]

Remark 5. This is expected, since the problem reduces to a standard MAB. Our regret bound is tighter than RC-LCB (in which \(E[T_{i,n}] \leq (16\sigma^2 \log n)/\Delta^2(i)\)) under the condition that \(\sigma^2 \geq 1/8\) (where \(\sigma\) is the largest sub-Gaussianity parameter of the arms).

Remark 6. For an infeasible and suboptimal arm \(i\) in a feasible instance, the expected number of pulls us upper bounded by \(2/\Delta^2(i)\) and \(D^i_{\alpha,\xi}\), and thus can be compactly written as
\[
\limsup_{n \to \infty} \frac{E[T_{i,n}]}{\log n} \leq \min \left\{ \frac{2}{\Delta^2(i)}, D^i_{\alpha,\xi} \right\}.
\]
This agrees with the regret analysis of RC-LCB if we take \(\alpha \to 1^-\) and choose \(\xi_{\alpha} \to 1^+\) (according to Corollary 2).

Although the bounds in Theorems 1–3 are asymptotic, finite-time bounds are available in the supp. material.

5 Proof Outline for Theorem 1

We outline the proof of Theorem 1, since it is the most involved. Theorem 2 follows by a straightforward substitution and Theorem 3 is analogous to the proof of Thompson sampling for standard MAB. Denote the sample CVaR as \(\hat{c}_\alpha(i, t) = \theta_{i,t}(\frac{\alpha}{1-\alpha}) + \frac{1}{\sqrt{\kappa_{i,t}}} c_{\alpha}^*\). Fix \(\varepsilon > 0\) and define
\[E_i(t) := \{\hat{c}_{\alpha}(i, t) > c_{\alpha}(1) + \varepsilon\},\]
the event that the Thompson sample mean of arm \(i\) is \(\varepsilon\)-riskier than a certain threshold or, more precisely, \(\varepsilon\)-higher than the optimal arm (which has the lowest mean, quantifying the expected loss). Intuitively, event \(E_i(t)\) is highly likely to occur when the algorithm has explored sufficiently. However, the algorithm does not choose arm \(i\) when \(E_i(t)\), an event with small probability under Thompson sampling, occurs. We can divide \(E[T_{i,n}]\) into two parts by Lemma 1 as follows.

Lemma 1 (Lattimore and Szepesvári [2020]). Let \(P_t(\cdot) = P(\cdot | A_1, X_1, \ldots, A_{t-1}, X_{t-1})\) be the probability measure conditioned on the history up to time \(t - 1\) and \(G_{is} = P_t(E^i_t(t) | T_{i,t} = s)\), where \(E^i_t(t)\) is any specified event for arm \(i\) at time \(t\). Then
\[
E[T_{i,n}] \leq \sum_{s=0}^{n-1} \left( \frac{1}{G_{is}} - 1 \right) + \sum_{s=0}^{n-1} \mathbb{I}\{G_{is} > \frac{1}{n}\} + 1.
\]

It remains to upper bound \(\Lambda_1 = \mathbb{E}\left[\sum_{s=0}^{n-1} \mathbb{I}\{G_{is} > \frac{1}{n}\}\right]\) and \(\Lambda_2 = \sum_{s=0}^{n-1} P(G_{is} > \frac{1}{n})\). The techniques used to upper-bound \(\Lambda_1\) resemble those in Zhu and Tan [2020], and are relegated to the supplementary material. To upper bound \(\Lambda_2\), we split the event \(E^i(t)\) into \(\Psi_1(\xi) = \left\{\theta_{i,t} - \mu_i \leq \left(\frac{\alpha}{1-\alpha}\right) \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right\}\) and \(\Psi_2(\xi) = \left\{\frac{1}{\sqrt{\kappa_{i,t}}} - \sigma_i \right\} c_{\alpha}^* \leq (1 + \xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right\} + (1 + \xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right\} + \mathbb{I}(\xi > 0).\]

Figure 1: Risk regrets averaged over 100 runs of instances whose arms follow Gaussian distributions with means and variances according to \((\mu, \sigma_1^2)\) and \((\mu, \sigma_2^2)\) respectively. The error bars indicate \(\pm 1\) standard deviations over the 100 runs. The proportions of wrong flags are \(r/N = 0.01\%, 0.02\%\) respectively.

That is, \(E^i_t(t) \subseteq \Psi_1(\xi) \cup \Psi_2(\xi)\). We use the union bound to control \(P(E^i_t(t)) \leq P(\Psi_1(\xi)) + P(\Psi_2(\xi))\). A good choice of the free parameter \(\xi\) allows us to allocate “weights” to \(\Psi_1(\xi)\) and \(\Psi_2(\xi)\) which then yields \(A^i_{\alpha,\xi}\) and \(B^i_{\alpha,\xi}\) in (8) and (9) without incurring further residual terms. See Lemma 6.

The sample Gaussian CVaR is of the form \(af(\theta) + bg(1/\kappa)\), where \(\theta\) follows a Gaussian distribution and \(\kappa\) follows a Gamma distribution, \(f(x) = x, g(x) = \sqrt{x}\) are bijective on their natural domains and have inverses \(f^{-1}, g^{-1}\) respectively, and \(a = \frac{\alpha}{1-\alpha}, b = c_{\alpha}^*\) depend on \(\alpha\). By splitting the Thompson sample as described in the previous paragraph, and using \(f^{-1}\) and \(g^{-1}\), we reduced the problem to upper bounding probabilities of the form \(P(\theta \leq \cdot)\) and \(P(\kappa \geq \cdot)\) with arguments as functions of \(\xi\) and concentration bounds can then be readily applied. This suggests that for general risk measures on Gaussian bandits with Thompson samples of the form \(af(\theta) + bg(1/\kappa)\), a similar strategy can be used to establish tight regret bounds. If, however, \(f\) and \(g\) take very different forms, establishing regret bounds might be more challenging.

6 Numerical Simulations

We verify the theory developed using numerical simulations. Set parameters \((N, \alpha, \tau, K, n) = (1000, 0.95, 4.6, 15, 1000)\). We present regret that is averaged over \(N\) statistically independent runs of horizon \(n\). For the \(K = 15\) arms, we
set their means as per the experiments in Sani et al. [2012] and Zhu and Tan [2020], namely, $\mu = (0.1, 0.2, 0.23, 0.27, 0.32, 0.32, 0.34, 0.41, 0.43, 0.54, 0.55, 0.56, 0.67, 0.71, 0.79)$.

By L.A. et al. [2020, Theorem 3.1] and Kagrecha et al. [2020a, Lemma 1], we have $(D_\sigma, d_\sigma) = (3.1/(8\sigma^2))$ where we set $\varepsilon < 2\delta$ therein for simplicity. Under this regime, the regret bounds for CVaR-TS are tighter than those of RC-LCB when the sub-Gaussianity parameter $\sigma$ satisfies $\sigma^2 \geq 1/32 = 0.03125$. Thus, when investigating Gaussian bandits with variances significantly larger than (resp. close to) $1/32$, we set the variances of the arms according to $\sigma_d^2$ (resp. $\sigma_{\underline{d}}^2$), where $\sigma_d^2 = (0.045, 0.144, 0.248, 0.339, 0.243, 0.172, 0.039, 0.144, 0.244, 0.353, 0.244, 0.146, 0.056, 0.149, 0.285)$ (resp. $\sigma_{\underline{d}}^2 = (0.0321, 0.0332, 0.0355, 0.0464, 0.0375, 0.0486, 0.0397, 0.0398, 0.0387, 0.0378, 0.0567, 0.0456, 0.0345, 0.0334, 0.0323)$).

In this setup, the first 2 arms are feasible, and arm 1 is optimal. We additionally set $\tau = 2$ to investigate an infeasible instance. We run both RC-LCB [Kagrecha et al., 2020a] and CVaR-TS.

Each pair of figures compares the regrets under the regime with variances $\sigma^2_d$ versus $\sigma^2_{\underline{d}}$. In the captions, we also state the ratio $r/N$ where $r$ is the total number of runs that CVaR-TS declares that the flag is wrong, i.e., it flags a feasible instance as being infeasible or vice versa. This helps us numerically verify that CVaR-TS does not perform worse than RC-LCB in flagging instances incorrectly. We see a much stronger performance by CVaR-TS under instances when $\sigma^2 \gg 1/32$, agreeing with the theoretical conclusion of the regret analyses. When $\sigma^2 \approx 1/32$, CVaR-TS still outperforms RC-LCB but not by much. The Python code for the simulations is included in the supplementary material.

7 Conclusion

This paper applies Thompson sampling [Thompson, 1933] to provide a solution for CVaR MAB problems [Galichet et al., 2013; Kagrecha et al., 2020b; Khajonchoptpanya et al., 2020] which were approached from the L/UCB perspectives previously. The regret bounds are notable improvements of those attained by the state-of-the-art L/UCB techniques, when the bandit environments are Gaussians satisfying certain assumptions. We corroborated the theoretical results through simulations and verified that under the conditions predicted by the theorems, the gains over previous approaches are significant.

Noting the similarity of the mean-variance of a Gaussian arm when $\rho \to +\infty$ and the CVaR of the same arm when $\alpha \to 1^-$, we believe there is a unifying theory of risk measures for Gaussian bandits; see Cassel et al. [2018], Lee et al. [2020] and Xi et al. [2020]. Furthermore, most papers consider sub-Gaussian bandits, while our work focuses on Gaussians. This is perhaps why we obtain better regret bounds in general. Further work includes analysing Thompson sampling of Gaussian MABs under general risk measures and exploring the performance of Thompson sampling for CVaR sub-Gaussian bandits.
References

Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions with formulas, graphs, and mathematical tables, volume 55. US Government printing office, 1970.

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

Sanjay P Bhat and Prashanth L.A. Concentration of risk measures: A Wasserstein distance approach. In Advances in Neural Information Processing Systems, pages 11762–11771, 2019.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In Proceedings of the 31st Conference On Learning Theory, pages 1295–1306, 2018.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. Probability Theory and Related Fields, 162(3-4):707–738, 2015.

Peter Harremoës. Bounds on tail probabilities for negative binomial distributions. Kybernetika, 52(6):943–966, 2016.

Xiao Guang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. Royal Society Open Science, 4(11):171377, 2017.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna Jaganathan. Constrained regret minimization for multi-criterion multi-armed bandits. arXiv preprint arXiv:2006.09649, 2020.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna Jaganathan. Statistically robust, risk-averse best arm identification in multi-armed bandits. arXiv preprint arXiv:2008.13629, 2020.

Najakorn Khajonchotpanya, Xue Yilin, and Napat Rujeerapaiboon. A revised approach for risk-averse multi-armed bandits under cvar criterion. ScholarBank NUS, 2020.

Prashanth L.A., Krishna Jaganathan, and Ravi Kumar Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In International Conference on Machine Learning, pages 3657–3666, 2020.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. Annals of Statistics, pages 1302–1338, 2000.

Jaeho Lee, Sejun Park, and Jinwoo Shin. Learning bounds for risk-sensitive learning. arXiv preprint arXiv:2006.08138, 2020.

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

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In Advances in Neural Information Processing Systems, pages 3275–3283, 2012.

Tasuku Soma and Yuichi Yoshida. Statistical learning with conditional value at risk. arXiv preprint arXiv:2002.05826, 2020.

Wen Sun, Debadeepta Dey, and Ashish Kapoor. Risk-aversion in multi-armed bandits. In International Conference on Machine Learning, pages 3280–3288, 2017.

Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for CVaR bandits. In Neural Information Processing Systems 2019 Workshop on Safety and Robustness in Decision Making, 2019.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4):285–294, 1933.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. IEEE Journal of Selected Topics in Signal Processing, 10(6):1093–1111, 2016.

Guangyu Xi, Chao Tao, and Yuan Zhou. Near-optimal MNL bandits under risk criteria. arXiv preprint arXiv:2009.12511, 2020.

Qiuyu Zhu and Vincent YF Tan. Thompson sampling algorithms for mean-variance bandits. In International Conference on Machine Learning, pages 2645–2654, 2020.
Supplementary Material

A Proof of Theorem 1

Proof of Theorem 1. Denote the sample CVaR at 0 by $\hat{c}_\alpha(i, t) = \theta_{i,t} \left( \frac{\alpha}{1-\alpha} \right) + \frac{1}{\sqrt{n_i,t}} c_\alpha^*$ (see (4)). Fix $\varepsilon > 0$, and define

$$E_{1}(t) := \{c_{\alpha}(i, t) > c_{\alpha}(1) + \varepsilon \},$$

the event that the Thompson sample mean of arm $i$ is $\varepsilon$-riskier than the threshold or $\varepsilon$-higher than the optimal arm (which has the lowest mean, quantifying expected loss). Intuitively, event $E_{1}(t)$ is highly likely to occur when the algorithm has explored sufficiently. However, the algorithm does not choose arm $i$ when $E_{1}(t)$, an event with small probability under Thompson sampling, occurs. By Lemma 1, and the linearity of expectation, we can divide $E[T_{i,n}]$ into two parts as

$$E[T_{i,n}] \leq \sum_{s=0}^{n-1} E \left[ \frac{1}{G_{1s}} - 1 \right] + \sum_{s=0}^{n-1} \mathbb{P} \left( G_{is} > \frac{1}{n} \right) + 1.$$  \hspace{1cm} (12)

By Lemmas 3 and 8 in the following, we have

$$\sum_{s=1}^{n} E \left[ \frac{1}{G_{1s}} - 1 \right] \leq \frac{C_{1}}{\varepsilon^3} + \frac{C_{2}}{\varepsilon^2} + \frac{C_{3}}{\varepsilon} + C_{4},$$

and

$$\sum_{s=1}^{n} \mathbb{P} \left( G_{is} > \frac{1}{n} \right) \leq 1 + \max \left\{ \frac{2\alpha^2 \log(2n)}{\xi^2(1-\alpha)^2 (\Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2}, \frac{\log(2n)}{h \left( \frac{\sigma_{i}^2 c_{\alpha}^*}{(\sigma_{i}^2 c_{\alpha})^2} \right)^2} \right\} + \frac{C_{5}}{\varepsilon^3} + \frac{C_{6}}{\varepsilon^2}.$$  \hspace{1cm} (13)

Plugging the two displays into (12), we have

$$E[T_{i,n}] \leq 1 + \max \left\{ \frac{2\alpha^2 \log(2n)}{\xi^2(1-\alpha)^2 (\Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2}, \frac{\log(2n)}{h \left( \frac{\sigma_{i}^2 c_{\alpha}^*}{(\sigma_{i}^2 c_{\alpha})^2} \right)^2} \right\} + \frac{C_{4}'}{\varepsilon^3} + \frac{C_{2}'}{\varepsilon^2} + \frac{C_{4}'}{\varepsilon} + C_{5}' \hspace{1cm} (14)$$

where $C_{1}', C_{2}', C_{3}', C_{4}', C_{5}'$ are constants. Setting $\varepsilon = (\log n)^{-\frac{1}{2}}$ into (13), we get

$$\limsup_{n \to \infty} \frac{\mathcal{R}_{n}^{\text{risk}} (\text{CVaR-TS})}{\log n} \leq \sum_{i \in [K]\setminus \mathcal{K}^*} \max \left\{ \frac{2\alpha^2 \log(2n)}{\xi^2(1-\alpha)^2 (\Delta_{\text{risk}}(\alpha)^2)}, \frac{\log(2n)}{h \left( \frac{\sigma_{i}^2 c_{\alpha}^*}{(\sigma_{i}^2 c_{\alpha})^2} \right)^2} \right\} \Delta_{\text{risk}}(i, \alpha)$$

$$= \sum_{i \in [K]\setminus \mathcal{K}^*} \max \left\{ A_{i,\alpha} \right\} \Delta_{\text{risk}}(i, \alpha) = \sum_{i \in [K]\setminus \mathcal{K}^*} C_{i,\alpha} \Delta_{\text{risk}}(i, \alpha).$$

Lemma 2. We can lower bound

$$\mathbb{P}_{t} (E_{1}^{*}(t) \mid T_{1,t} = s, \mu_{1,s} = \mu, \hat{\sigma}_{1,s} = \sigma) = \mathbb{P}_{t} (\hat{c}_{\alpha}(1, t) \leq c_{\alpha}(1) + \varepsilon \mid T_{1,t} = s, \mu_{1,s} = \mu, \hat{\sigma}_{1,s} = \sigma)$$

by

$$\mathbb{P}_{t} (\hat{c}_{\alpha}(1, t) \leq c_{\alpha}(1) + \varepsilon \mid T_{1,t} = s, \mu_{1,s} = \mu, \hat{\sigma}_{1,s} = \sigma) \geq \begin{cases} \mathbb{P}_{t} \left( \theta_{1,t} - \mu_1 \leq \frac{(1-\alpha)\varepsilon}{2\alpha} \right) \cdot \mathbb{P}_{t} \left( \frac{1}{\sqrt{\mathcal{K}_1,t}} - \sigma_1 \leq \frac{\varepsilon}{2c_{\alpha}^*} \right) & \text{if } \mu \geq \mu_1, \sigma \geq \sigma_1, \\ \frac{1}{2} \mathbb{P}_{t} \left( \frac{1}{\sqrt{\mathcal{K}_1,t}} - \sigma_1 \leq \frac{\varepsilon}{2c_{\alpha}^*} \right) & \text{if } \mu < \mu_1, \sigma \geq \sigma_1, \\ \frac{1}{2} \mathbb{P}_{t} \left( \frac{1}{\sqrt{\mathcal{K}_1,t}} - \sigma_1 \leq \frac{(1-\alpha)\varepsilon}{2\alpha} \right) & \text{if } \mu \geq \mu_1, \sigma < \sigma_1, \\ \frac{1}{2} \mathbb{P}_{t} \left( \frac{1}{\sqrt{\mathcal{K}_1,t}} - \sigma_1 \leq \frac{(1-\alpha)\varepsilon}{2\alpha} \right) & \text{if } \mu < \mu_1, \sigma < \sigma_1. \end{cases} \hspace{1cm} (14)$$

Proof of Lemma 2. Given $T_{1,t} = s, \mu_{1,s} = \mu, \hat{\sigma}_{1,s} = \sigma \leq 1$, a direct calculation gives us,

$$\mathbb{P}_{t} (\hat{c}_{\alpha}(1, t) \leq c_{\alpha}(1) + \varepsilon \mid T_{1,t} = s, \mu_{1,s} = \mu, \hat{\sigma}_{1,s} = \sigma)$$

$$= \mathbb{P}_{t} \left( \theta_{1,t} - \mu_1 \left( \frac{\alpha}{1-\alpha} \right) + \left( \frac{1}{\sqrt{\mathcal{K}_1,t}} - \sigma_1 \right) c_{\alpha}^* \leq \varepsilon \right) \geq \mathbb{P}_{t} \left( \theta_{1,t} - \mu_1 \left( \frac{1-\alpha)\varepsilon}{2\alpha} \right) + \frac{1}{\sqrt{\mathcal{K}_1,t}} - \sigma_1 \leq \frac{\varepsilon}{2c_{\alpha}^*} \right).$$
If $\mu < \mu_1$, then
\[
P_t\left(\theta_{1,t} - \mu_1 \leq \frac{(1-\alpha)\varepsilon}{2\alpha}\right) = P_t\left(\theta_{1,t} - \mu \leq \mu_1 - \mu + \frac{(1-\alpha)\varepsilon}{2\alpha}\right) \geq \frac{1}{2}
\]
by the properties of the median of the Gaussian distribution.

If $\sigma < \sigma_1$, then
\[
P_t\left(\frac{1}{\sqrt{k_{1,t}}} - \sigma_1 \leq \frac{\varepsilon}{2\sigma_1^2} \leq \sigma_1^2 \leq \left(\frac{\varepsilon}{2\sigma_1^2} + 2\sigma_1\right)\frac{\varepsilon}{2\sigma_1^2} \right) \geq \frac{1}{2}
\]
by the properties of the median of the Gamma distribution.

\[\square\]

**Lemma 3** (Upper bounding the first term of (12)). We have
\[
\sum_{s=1}^{n} E\left[\frac{1}{G_{1s}} - 1\right] \leq \frac{C_1}{\varepsilon^3} + \frac{C_2}{\varepsilon^2} + \frac{C_3}{\varepsilon} + C_4,
\]
where $C_1, C_2, C_3, C_4$ are constants.

**Proof of Lemma 3.** We now attempt to bound $E\left[\frac{1}{G_{1s}} - 1\right]$ by conditioning on the various values of $\hat{\mu}_{i,s}$ and $\hat{\sigma}_{i,s}$. Firstly, we define the conditional version of $G_{1s}$ by
\[
\tilde{G}_{1s} = G_{1s}|_{\hat{\mu}_{i,s} = \mu, \hat{\sigma}_{i,s} = \beta} = P_t(\hat{\epsilon} \alpha(1, t) \leq c_\alpha(1) + \varepsilon | T, \hat{\mu}_{i,s} = \mu, \hat{\sigma}_{i,s} = \beta),
\]
which is the LHS of (14). Define $c_1 = \frac{1}{\sqrt{2\pi\sigma_1^2}}$, $c_2 = \frac{1}{2^{\frac{\alpha}{2}} \Gamma(\frac{\alpha}{2})\sigma_1^{\frac{\alpha}{2}}}$, and $\omega = s\sqrt{2}\sigma_1$. For clarity, we partition the parameter space $(\beta, \mu) \in [0, \infty) \times (-\infty, \infty) = A \cup B \cup C \cup D$, where
\[
A = [0, \omega) \times \left[\mu_1 + \frac{(1-\alpha)\varepsilon}{2\alpha}, \infty\right), \quad B = [0, \omega) \times (-\infty, \frac{(1-\alpha)\varepsilon}{2\alpha}), \quad C = [\omega, \infty) \times \left[\frac{(1-\alpha)\varepsilon}{2\alpha}, \infty\right), \quad D = [\omega, \infty) \times (-\infty, \frac{(1-\alpha)\varepsilon}{2\alpha}].
\]
We can then partition $E\left[\frac{1}{G_{1s}} - 1\right]$ into four parts:
\[
E\left[\frac{1}{G_{1s}} - 1\right] = c_1 c_2 \int_0^\infty \int_{-\infty}^\infty \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp\left(-\frac{s(\mu - \mu_1)^2}{2\sigma_1^2}\right) \beta^{\frac{\alpha}{2}} e^{-\frac{s^2}{2\sigma_1^2}} d\mu d\beta
\]
\[
= c_1 c_2 \left(\int_A + \int_B + \int_C + \int_D\right) \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp\left(-\frac{s(\mu - \mu_1)^2}{2\sigma_1^2}\right) \beta^{\frac{\alpha}{2}} e^{-\frac{s^2}{2\sigma_1^2}} d\mu d\beta.
\]
For Part $B$, using the fourth case in Lemma 2,
\[
\frac{1 - \tilde{G}_{1s}}{G_{1s}} \leq 4(1 - \tilde{G}_{1s}).
\]
It follows that
\[
c_1 c_2 \int_B \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp\left(-\frac{s(\mu - \mu_1)^2}{2\sigma_1^2}\right) \beta^{\frac{\alpha}{2}} e^{-\frac{s^2}{2\sigma_1^2}} d\mu d\beta
\]
\[
\leq 4 c_1 c_2 \int_B (1 - \tilde{G}_{1s}) \exp\left(-\frac{s(\mu - \mu_1)^2}{2\sigma_1^2}\right) \beta^{\frac{\alpha}{2}} e^{-\frac{s^2}{2\sigma_1^2}} d\mu d\beta
\]
\[
\leq 4 c_1 c_2 \int_B \left(P_t\left(\theta_{1,t} - \mu_1 \geq \frac{(1-\alpha)\varepsilon}{2\alpha}, \hat{\mu}_{i,s} = \mu\right) + P_t\left(\frac{1}{\sqrt{k_{1,t}}} - \sigma_1 \geq \frac{\varepsilon}{2\sigma_1^2}, \hat{\sigma}_{i,s} = \beta\right)\right) \exp\left(-\frac{s(\mu - \mu_1)^2}{2\sigma_1^2}\right) \beta^{\frac{\alpha}{2}} e^{-\frac{s^2}{2\sigma_1^2}} d\mu d\beta
\]
\[
\leq 4 c_1 \int_{-\infty}^{\mu_1 + \frac{(1-\alpha)\varepsilon}{2\alpha}} \left[P_t\left(\theta_{1,t} - \mu_1 \geq \frac{(1-\alpha)\varepsilon}{2\alpha}, \hat{\mu}_{i,s} = \mu\right) \exp\left(-\frac{s(\mu - \mu_1)^2}{2\sigma_1^2}\right) d\mu
\]
\[
+ 4 c_2 \int_0^\omega P_t\left(\frac{1}{\sqrt{k_{1,t}}} - \sigma_1 \geq \frac{\varepsilon}{2\sigma_1^2}, \hat{\sigma}_{i,s} = \beta\right) \beta^{\frac{\alpha}{2}} e^{-\frac{s^2}{2\sigma_1^2}} d\beta
\]
Then where (15) and (16) respectively follow from applying standard tail upper bounds on the Gaussian and Gamma distributions.

**Lemma 4** (Abramowitz and Stegun [1970])

For Part \( D \), the third case in the above lemma,

\[
\frac{1 - \tilde{G}_{1s}}{G_{1s}} \leq \frac{2}{P_t \left( \theta_{1,t} - \mu_1 \leq \frac{(1-\alpha)\varepsilon}{2\alpha} \mid \hat{\sigma}_{1,s} = \mu \right)}.
\]

Then

\[
c_1c_2 \int_A \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s/2} e^{-\frac{\beta^2}{2\sigma_1^2}} d\beta
\leq \frac{2c_1c_2}{\sqrt{\pi}} \int_{\mu_1 + \frac{(1-\alpha)\varepsilon}{2\alpha}}^{\infty} \frac{1}{P_t \left( \theta_{1,t} - \mu_1 \leq \frac{(1-\alpha)\varepsilon}{2\alpha} \mid \hat{\sigma}_{1,s} = \mu \right)} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s/2} e^{-\frac{\beta^2}{2\sigma_1^2}} d\beta
\leq \frac{2c_1\sqrt{2}}{\sqrt{\pi}} \int_{\mu_1 + \frac{(1-\alpha)\varepsilon}{2\alpha}}^{\infty} \left( \sqrt{s \left( \mu - \mu_1 \right)^2 - \frac{(1-\alpha)\varepsilon}{2\alpha} \right) + \sqrt{s \left( \mu - \mu_1 \right)^2 - (1-\alpha)\varepsilon \right)^2 + 4)
\exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) d\mu
\leq \frac{2c_1\sqrt{2}}{\sqrt{\pi}} \int_0^{\infty} \left( \sqrt{s z} + \sqrt{s \varepsilon^2 + 4} \right) \exp \left( -\frac{s\varepsilon^2}{2} \right) d\beta
\leq \frac{2c_1\sqrt{2}}{\sqrt{\pi}} \exp \left( -\frac{s\varepsilon^2}{2} \right) \int_0^{\infty} 3 \left( \sqrt{s} z \right) \exp (sz \varepsilon) dz \leq \frac{6}{\pi \varepsilon^2 s^\frac{1}{2}} \exp \left( -\frac{s(1-\alpha)^2\varepsilon^2}{8\alpha^2} \right),
\]

where (17) follows from Lemma 4, which we state below.

**Lemma 4** (Abramowitz and Stegun [1970]). For \( X \sim \mathcal{N}(\mu, 1/s) \),

\[
P(X \leq \mu - x) = P(X \geq \mu + x) \geq \sqrt{\frac{2}{\pi}} \frac{\exp \left( -\frac{s\varepsilon^2}{2} \right)}{\sqrt{s \varepsilon^2 + \sqrt{s \varepsilon^2 + 4}}}.
\]

For Part \( D \), using the second case in Lemma 2,

\[
\frac{1 - \tilde{G}_{1s}}{G_{1s}} \leq \frac{2}{P_t \left( \frac{1}{\sqrt{\sigma_{1,t}}} - \sigma_1 \leq \frac{\varepsilon}{2\alpha} \mid \hat{\sigma}_{1,s} = \beta \right)}.
\]
Then
\[
c_{1}c_{2} \int \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp \left( - \frac{s(\mu - \mu_{1})^2}{2\sigma_{1}^2} \right) \beta^{s-2} e^{-\frac{\beta^2}{2\beta_1^2}} d\mu d\beta
\]
\[
\leq 2c_{1}c_{2} \int_{D} \mathbb{P}_t \left( \frac{1}{\sqrt{\kappa_{1,t}}} - \sigma_{1} \leq \frac{\varepsilon}{2\epsilon_{c}} | \bar{\sigma}_{1,s} = \beta \right) \exp \left( - \frac{s(\mu - \mu_{1})^2}{2\sigma_{1}^2} \right) \beta^{s-2} e^{-\frac{\beta^2}{2\beta_1^2}} d\mu d\beta
\]
\[
\leq c_{1} \int_{-\infty}^{\mu_{1}} \exp \left( - \frac{s(\mu - \mu_{1})^2}{2\sigma_{1}^2} \right) d\mu \cdot 2c_{2} \int_{\omega}^{\infty} \mathbb{P}_t \left( \frac{1}{\sqrt{\kappa_{1,t}}} - \sigma_{1} \leq \frac{\varepsilon}{2\epsilon_{c}} | \bar{\sigma}_{1,s} = \beta \right) \beta^{s-2} e^{-\frac{\beta^2}{2\beta_1^2}} d\beta
\]
\[
\leq 2c_{2} \int_{\omega}^{\infty} \mathbb{P}_t \left( \frac{1}{\sqrt{\kappa_{1,t}}} - \sigma_{1} \leq \frac{\varepsilon}{2\epsilon_{c}} | \bar{\sigma}_{1,s} = \beta \right) \beta^{s-2} e^{-\frac{\beta^2}{2\beta_1^2}} d\beta
\]
\[
\leq 2c_{2} \Gamma \left( \frac{s}{2} \right) \int_{\omega}^{\infty} \exp \left( \frac{\beta^2}{(\frac{\varepsilon}{2\epsilon_{c}} + \sigma_{1})^2} - \frac{\beta^2}{2\sigma_{1}^2} \right) \beta^{s-2} \left( 1 + \frac{\beta^2}{(\frac{\varepsilon}{2\epsilon_{c}} + \sigma_{1})^2} \right)^{-(\frac{s-1}{2})} d\beta
\]
\[
\leq 2c_{2} \Gamma \left( \frac{s}{2} \right) \cdot \left( \frac{\varepsilon}{2\epsilon_{c}} + \sigma_{1} \right)^{s-1} \int_{\omega}^{\infty} \exp \left( \frac{\beta^2 - \beta^2 (\frac{\varepsilon}{2\epsilon_{c}} + \sigma_{1})^2}{2\sigma_{1}^2} \right) y^{s-2} (1 + y^2)^{-(\frac{s-1}{2})} dy
\]
\[
\leq \frac{c_{2}^{2} \Gamma \left( \frac{s}{2} \right) \cdot \left( \frac{\varepsilon}{2\epsilon_{c}} + \sigma_{1} \right)^{s-1}}{2\sigma_{1}^2 - (\frac{\varepsilon}{2\epsilon_{c}} + \sigma_{1})^2} \exp \left( \frac{\beta^2 - \beta^2 (\frac{\varepsilon}{2\epsilon_{c}} + \sigma_{1})^2}{2\sigma_{1}^2} \right)
\]
where (18) follows from Lemma 5, which we state below.

**Lemma 5 (Zhu and Tan [2020, Lemma S.1]).** If $X \sim \text{Gamma}(\alpha, \beta)$ with $\alpha \geq 1$ and rate $\beta > 0$, we have the lower bound on the complementary cumulative distribution function

\[
\mathbb{P}(X \geq x) \geq \frac{1}{\Gamma(\alpha)} \exp(-\beta x)(1 + \beta x)^{\alpha-1}.
\]

For Part C, using the first case in Lemma 2,

\[
\frac{1 - \tilde{G}_{1s}}{G_{1s}} \leq \mathbb{P}(\theta_{1,t} - \mu_{1} \leq \varepsilon | \bar{\mu}_{1,s} = \mu) \cdot \mathbb{P}(\frac{1}{\sqrt{\kappa_{1,t}}} - \sigma_{1} \leq \varepsilon | \bar{\sigma}_{1,s} = \beta).
\]

Reusing the integrations in Parts A and D,

\[
c_{1}c_{2} \int \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp \left( - \frac{s(\mu - \mu_{1})^2}{2\sigma_{1}^2} \right) \beta^{s-2} e^{-\frac{\beta^2}{2\beta_1^2}} d\mu d\beta
\]
\[
\leq c_{1}c_{2} \int D \mathbb{P}_t (\theta_{1,t} - \mu_{1} \leq \frac{(1-\alpha)\varepsilon}{2\alpha} | \bar{\mu}_{1,s} = \mu) \cdot \mathbb{P}_t \left( \frac{1}{\sqrt{\kappa_{1,t}}} - \sigma_{1} \leq \frac{\varepsilon}{2\epsilon_{c}} | \bar{\sigma}_{1,s} = \beta \right)
\]
\[
\cdot \exp \left( - \frac{s(\mu - \mu_{1})^2}{2\sigma_{1}^2} \right) \beta^{s-2} e^{-\frac{\beta^2}{2\beta_1^2}} d\mu d\beta
\]
\[
\leq \frac{1}{4} \cdot 2c_{2} \int_{\mu_{1} + \frac{(1-\alpha)\varepsilon}{2\alpha}}^{\infty} \mathbb{P}_t (\theta_{1,t} - \mu_{1} \leq \frac{(1-\alpha)\varepsilon}{2\alpha} | \bar{\mu}_{1,s} = \mu) \exp \left( - \frac{s(\mu - \mu_{1})^2}{2\sigma_{1}^2} \right) d\mu
\]
\[
\cdot 2c_{2} \int_{\omega}^{\infty} \mathbb{P}_t \left( \frac{1}{\sqrt{\kappa_{1,t}}} - \sigma_{1} \leq \frac{\varepsilon}{2\epsilon_{c}} | \bar{\sigma}_{1,s} = \beta \right) \beta^{s-2} e^{-\frac{\beta^2}{2\beta_1^2}} d\beta.
\]
\[
\leq \frac{3}{\pi \varepsilon^2 s \sqrt{s}} \exp \left( -\frac{s(1 - \alpha)^2 \varepsilon^2}{8\alpha^2} \right) \cdot \frac{\sigma_1^2 c_2 \Gamma \left( \frac{\varepsilon}{\alpha} \right)}{2\sigma_1^2 - \left( \frac{\varepsilon}{\alpha} + \sigma_1 \right)^2} \exp \left( -\frac{2\sigma_1^2 - \left( \frac{\varepsilon}{\alpha} + \sigma_1 \right)^2}{\sigma_1^2} s^2 \right) .
\]

Combining these four parts, we can upper bound \( \mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right] \) by
\[
\mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right] = 4 \exp \left( -\frac{s(1 - \alpha)^2 \varepsilon^2}{16\alpha^2} \right) + 4 \exp \left( -\frac{s \varepsilon^2}{16(c_\alpha^*)^2} \right) + \frac{6}{\pi \varepsilon^2 s \sqrt{s}} \exp \left( -\frac{s(1 - \alpha)^2 \varepsilon^2}{8\alpha^2} \right) + \frac{2\sigma_1 \left( \frac{\varepsilon}{\alpha} + \sigma_1 \right)^2}{2\sigma_1^2 - \left( \frac{\varepsilon}{\alpha} + \sigma_1 \right)^2} \exp \left( -\frac{2\sigma_1^2 - \left( \frac{\varepsilon}{\alpha} + \sigma_1 \right)^2}{\sigma_1^2} s^2 \right) + \frac{3}{\pi \varepsilon^2 s \sqrt{s}} \exp \left( -\frac{s(1 - \alpha)^2 \varepsilon^2}{8\alpha^2} \right) .
\]

Summing over \( s \), we have
\[
\sum_{s=1}^{n} \mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right] \leq \frac{C_1}{\varepsilon^2} + \frac{C_2}{\varepsilon^2} + \frac{C_3}{\varepsilon} + C_4 .
\]

**Lemma 6.** For \( \xi \in (0, 1) \), we have
\[
\mathbb{P} \left( \hat{c}_i(t) \leq c_\alpha(1) + \varepsilon \mid T_{i,s} = s, \hat{\mu}_{i,s} = \mu, \hat{\sigma}_{i,s} = \sigma \right) \\
\leq \exp \left( -\frac{s}{2} \left( \mu - \mu_i + \frac{\xi(\Delta_{\text{risk}}(i, \alpha) - \varepsilon)(1 - \alpha)}{\alpha} \right)^2 \right) + \exp \left( -\frac{2\sigma_i^2 - \left( \frac{\varepsilon}{\alpha} + \sigma_i \right)^2}{\sigma_i^2} s^2 \right) .
\]

where \( h(x) = \frac{1}{2}(x - 1 - \log x) \).

*Proof of Lemma 6.* Given \( T_{i,t} = s, \hat{\mu}_{i,s} = \mu > \mu_i, \hat{\sigma}_{i,s} = \sigma > \sigma_i \), a direct calculation gives us,
\[
\mathbb{P} \left( \hat{c}_i(t) \leq c_\alpha(1) + \varepsilon \mid T_{i,t} = s, \hat{\mu}_{i,s} = \mu, \hat{\sigma}_{i,s} = \sigma \right) \\
= \mathbb{P} \left( \frac{\alpha}{1 - \alpha} \right) + \frac{1}{\sqrt{\kappa_{i,t}}} c_\alpha^* \leq c_\alpha(1) + \varepsilon \right) \\
= \mathbb{P} \left( \frac{\alpha}{1 - \alpha} \right) + \frac{1}{\sqrt{\kappa_{i,t}}} c_\alpha^* \leq -\Delta_{\text{risk}}(i, \alpha) + \varepsilon \right) \\
= \mathbb{P} \left( \frac{\alpha}{1 - \alpha} \right) + \frac{1}{\sqrt{\kappa_{i,t}}} c_\alpha^* \leq -\xi + (-1 + \xi)(\Delta_{\text{risk}}(i, \alpha) - \varepsilon) \right) \\
\leq \mathbb{P} \left( \frac{\alpha}{1 - \alpha} \right) - \xi + (-1 + \xi)(\Delta_{\text{risk}}(i, \alpha) - \varepsilon) \right) \\
\leq \exp \left( -\frac{s}{2} \left( \mu - \mu_i + \frac{\xi(\Delta_{\text{risk}}(i, \alpha) - \varepsilon)(1 - \alpha)}{\alpha} \right)^2 \right) + \mathbb{P} \left( \frac{\alpha}{1 - \alpha} \right) - \xi + (-1 + \xi)(\Delta_{\text{risk}}(i, \alpha) - \varepsilon) \right) \\
\leq \exp \left( -\frac{s}{2} \left( \mu - \mu_i + \frac{\xi(\Delta_{\text{risk}}(i, \alpha) - \varepsilon)(1 - \alpha)}{\alpha} \right)^2 \right) + \exp \left( -\frac{2\sigma_i^2 - \left( \frac{\varepsilon}{\alpha} + \sigma_i \right)^2}{\sigma_i^2} s^2 \right) .
\]

where (19) follows from Lemma 7, which we state below.

**Lemma 7** (Harremoës [2016]). For a Gamma r.v. \( X \sim \text{Gamma}(\alpha, \beta) \) with shape \( \alpha \geq 2 \) and rate \( \beta > 0 \), we have
\[
\mathbb{P}(X \geq x) \leq \exp \left( -2\alpha h \left( \frac{\beta x}{\alpha} \right) \right) , \quad x > \frac{\alpha}{\beta},
\]
where \( h(x) = \frac{1}{2}(x - 1 - \log x) \).
Lemma 8 (Upper bounding the second term of (12)). We have

\[
\sum_{s=1}^{n} P_t \left( G_{is} > \frac{1}{n} \right) \leq 1 + \max \left\{ \frac{2a^2 \log(2n)}{\xi^2(1-\alpha)^2 \left( \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right)^2}, \frac{\log(2n)}{h \left( \frac{\sigma^2(c_\alpha^*)^2}{(\sigma_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2} \right)} \right\} + \frac{C_5}{\varepsilon^4} + \frac{C_6}{\varepsilon^2},
\]

where \( C_5, C_6 \) are constants.

Proof of Lemma 8. From Lemma 6 we have the inclusions

\[
\left\{ \hat{\mu}_{i,s} - \sqrt{\frac{2 \log(2n)}{s}} \geq \mu_i - \frac{\xi \left( \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right) (1-\alpha)}{\alpha} \right\} \subseteq \left\{ \exp \left( -\frac{s}{2} \left( \hat{\mu}_{i,s} - \mu_i - \frac{\xi \left( \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right) (1-\alpha)}{\alpha} \right)^2 \right) \leq \frac{1}{2n} \right\}
\]

and

\[
\left\{ \frac{\hat{\sigma}^2_{i,s} (c_\alpha^*)^2}{(\sigma_i c_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2} \geq \frac{1}{h_+^{-1} \left( \frac{\log(2n)}{s} \right)} \right\} \cup \left\{ \frac{\hat{\sigma}^2_{i,s} (c_\alpha^*)^2}{(\sigma_i c_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2} \leq \frac{1}{h_+^{-1} \left( \frac{\log(2n)}{s} \right)} \right\} \subseteq \left\{ \exp \left( -sh \left( \frac{\sigma^2(c_\alpha^*)^2}{(\sigma_i c_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2} \right) \right) \leq \frac{1}{2n} \right\}
\]

where \( h_+^{-1}(y) = \max \{ x : h(x) = y \} \) and \( h_+^{-1}(y) = \min \{ x : h(x) = y \} \). Hence, for

\[
s \geq u = \max \left\{ \frac{2a^2 \log(2n)}{\xi^2(1-\alpha)^2 \left( \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right)^2}, \frac{\log(2n)}{h \left( \frac{\sigma^2(c_\alpha^*)^2}{(\sigma_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2} \right)} \right\},
\]

we have

\[
P_t \left( G_{is} > \frac{1}{n} \right) \leq P_t \left( \hat{\mu}_{i,s} - \sqrt{\frac{2 \log(2n)}{s}} \leq \mu_i - \frac{\xi \left( \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right) (1-\alpha)}{\alpha} \right)
\]

\[
+ P_t \left( \frac{\sqrt{2 \log(2n)}}{s} \leq \frac{\sigma^2_{i,s} (c_\alpha^*)^2}{(\sigma_i c_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2} \leq \frac{1}{h_+^{-1} \left( \frac{\log(2n)}{s} \right)} \right)
\]

\[
\leq P_t \left( \hat{\mu}_{i,s} - \mu_i \leq \sqrt{\frac{2 \log(2n)}{s}} - \frac{\xi \left( \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right) (1-\alpha)}{\alpha} \right)
\]

\[
+ P_t \left( \sigma^2_{i,s} \leq \frac{(\sigma_i c_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2}{(c_\alpha^*)^2} \leq \frac{1}{h_+^{-1} \left( \frac{\log(2n)}{s} \right)} \right)
\]

\[
\leq \exp \left( -s \left( \frac{\xi \Delta_{\text{risk}}(i, \alpha) - \varepsilon)(1-\alpha)}{\alpha} - \sqrt{\frac{2 \log(2n)}{s}} \right)^2 \right)
\]

\[
+ \exp \left( -s \left( \frac{\sigma^2_i (\sigma_i c_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2}{(c_\alpha^*)^2} \right)^2 \frac{1}{h_+^{-1} \left( \frac{\log(2n)}{s} \right)} \right)
\]

\[
\leq \exp \left( -\frac{s \varepsilon^2}{\sigma^2_i} \right) + \exp \left( -(s-1) \frac{\varepsilon^4}{4 \sigma^4_i} \right),
\]

where (21) follows from Lemma 9, which we state below. Summing over \( s \),

\[
\sum_{s=1}^{n} P_t \left( G_{is} > \frac{1}{n} \right) \leq u + \sum_{s=[u]}^{n} \left[ \exp \left( -\frac{s \varepsilon^2}{\sigma^2_i} \right) + \exp \left( -(s-1) \frac{\varepsilon^4}{4 \sigma^4_i} \right) \right]
\]

\[
\leq 1 + \max \left\{ \frac{2a^2 \log(2n)}{\xi^2(1-\alpha)^2 \left( \Delta_{\text{risk}}(i, \alpha) - \varepsilon \right)^2}, \frac{\log(2n)}{h \left( \frac{\sigma^2(c_\alpha^*)^2}{(\sigma_\alpha^* - (1-\xi) \Delta_{\text{risk}}(i, \alpha) - \varepsilon)^2} \right)} \right\} + \frac{C_5}{\varepsilon^4} + \frac{C_6}{\varepsilon^2}.
\]
Lemma 9 (Laurent and Massart [2000]). For any $X \sim \chi^2_{s-1}$,

$$P(X \leq x) \leq \exp \left(-\frac{(s-1-x)^2}{4(s-1)}\right).$$

\[ \square \]

Corollary 1. Let $B^i_{\alpha,\xi}$ be defined as in Theorem 1. Then choosing

$$\xi_\alpha = 1 - \frac{\sigma_i c^*_\alpha}{\Delta_{\text{risk}}(i, \alpha)} \left(1 - \frac{1}{c^*_\alpha(2-\alpha)}\right)$$

yields $\xi_\alpha \to 1^-$ and $B^i_{\alpha,\xi_\alpha} \to 0$ as $\alpha \to 1^-$.\[ \square \]

Proof of Corollary 1. Recall that $c^*_\alpha \to +\infty$ and $c^*(1-\alpha) \to 0$. Then

$$c^*_\alpha(2-\alpha) = c^*_\alpha(1-\alpha) + c^*_\alpha \to 0 + \infty = +\infty \Rightarrow \frac{1}{c^*_\alpha(2-\alpha)} \to 0.$$ 

Furthermore, since $(1-\alpha)c_\alpha(i) = \mu_i + \sigma_i c^*_\alpha(1-\alpha) \to (1-\alpha)\Delta_{\text{risk}}(i, \alpha) = \mu_i + \sigma_i c^*_\alpha(1-\alpha) - c_\alpha(1-\alpha) \to \mu_i - \mu_1,$ we have

$$\frac{\sigma_i c^*_\alpha}{\Delta_{\text{risk}}(i)} = \frac{\sigma_i c^*_\alpha(1-\alpha)}{\mu_i + \sigma_i c^*_\alpha(1-\alpha) - c_\alpha(1-\alpha)} \to \frac{\sigma_i}{\mu_i - \mu_1} = 0.$$ 

Hence, $\xi_\alpha = 1 - \frac{\sigma_i c^*_\alpha}{\Delta_{\text{risk}}(i, \alpha)} \left(1 - \frac{1}{c^*_\alpha(2-\alpha)}\right) \to 1 - 0(1-0) = 1.$ Thus,

$$\frac{\sigma_i c^*_\alpha}{\sigma_i c^*_\alpha - (1 - \xi_\alpha)\Delta_{\text{risk}}(i, \alpha)} = \frac{\sigma_i c^*_\alpha}{\sigma_i c^*_\alpha - (1 - \frac{1}{c^*_\alpha(2-\alpha)})} = \frac{1}{1 - \frac{1}{c^*_\alpha(2-\alpha)}} = \frac{1}{c^*_\alpha(2-\alpha)} \to +\infty,$$

and $h(c^*_\alpha(2-\alpha)) \to +\infty$. Finally,

$$B^i_{\alpha,\xi_\alpha} = \frac{1}{h\left(\frac{\sigma^2(c^*_\alpha)^2}{(\sigma_i c^*_\alpha - (1 - \xi_\alpha)\Delta_{\text{risk}}(i, \alpha))^2}\right)} = \frac{1}{h(c^*_\alpha(2-\alpha))} \to 0,$$

as required.\[ \square \]

B Proof of Theorem 2

The proof of Theorem 2 is similar to that of Theorem 1.

Proof of Theorem 2. Fix $\varepsilon > 0$, and define

$$E_i(t) := \{\hat{\ell}_\alpha(i, t) > \tau + \varepsilon\},$$

By Lemma 1, and the linearity of expectation, we can split $E[T_{i,n}]$ into two parts as follows

$$E[T_{i,n}] \leq \sum_{s=0}^{n-1} \mathbb{E}\left[\frac{1}{G_{is}} - 1\right] + \sum_{s=0}^{n-1} \mathbb{P}\left(G_{is} > \frac{1}{n}\right) + 1. \tag{22}$$

By Lemmas 10 and 11, we have

$$\sum_{s=1}^{n} \mathbb{E}\left[\frac{1}{G_{is}} - 1\right] \leq \frac{C^7}{\varepsilon^2} + C_8,$$

and

$$\sum_{s=1}^{n} \mathbb{P}_t\left(G_{is} > \frac{1}{n}\right) \leq 1 + \max\left\{\frac{2\alpha^2 \log(2n)}{\xi^2(1 - \alpha)^2 (\Delta_r(i, \alpha) - \varepsilon)^2}, \frac{\log(2n)}{h\left(\frac{\sigma^2(c^*_\alpha)^2}{(\sigma_i c^*_\alpha - (1 - \xi_\alpha)\Delta_{\text{risk}}(i, \alpha))^2}\right)}\right\} + \frac{C_9}{\varepsilon^2} + \frac{C_{10}}{\varepsilon^2}.$$
Proof of Lemma 10. We note that

\[ \text{Plugging the two displays into (22), we have} \]

\[
E[T_{1,n}] \leq 1 + \max \left\{ \frac{2\alpha^2 \log(2n)}{\xi^2(1-\alpha)^2(D(i,\alpha) - \epsilon)^2}, \frac{\log(2n)}{h} \left( \frac{\sigma^2(c_{\alpha})^2}{(\sigma_i c_{\alpha}^*)^2} \right) \right\} + C_6' \frac{e^2}{\epsilon^2} + C_8', \tag{23}
\]

where \( C_6', C_7', C_8' \) are constants. Setting \( \epsilon = (\log n)^{-\frac{1}{2}} \) into (23), we get

\[
\limsup_{n \to \infty} \frac{R_{n \to \infty}^{(\text{CVaR-TS})}}{\log n} \leq \sum_{i \in \mathcal{K}_1} \left( \max \left\{ \frac{2\alpha^2}{\xi^2(1-\alpha)^2 \Delta^2(i)}, \frac{1}{h} \left( \frac{\sigma^2(c_{\alpha})^2}{(\sigma_i c_{\alpha}^*)^2} \right) \right\} \right) \Delta(i, \alpha) = \sum_{i \in \mathcal{K}^*} \max \{ E_{\alpha,\xi}^i, E_{\alpha,\xi}^i \} \Delta(i, \alpha) = \sum_{i \in |\mathcal{K}| \setminus \mathcal{K}^*} D_{\alpha,\xi}^i \Delta(i, \alpha).
\]

\[ \square \]

Lemma 10 (Upper bounding the first term of (22)). We have

\[
\sum_{s=1}^n E \left[ \frac{1}{G_{1s}} - 1 \right] \leq \frac{C_7'}{\epsilon^2} + C_8,
\]

where \( C_7', C_8 \) are constants.

Proof of Lemma 10. We note that \( i \) is either a deceiver arm or a non-deceiver arm.

(a) Suppose \( i \) is a deceiver arm. Then

\[
E_i(t) = \{ \hat{c}_i(t) > \tau + \epsilon \} \cap \{ \hat{\mu}_i,t \leq \mu_1 - \epsilon \} \Rightarrow E_i^\prime(t) \supseteq \{ \hat{\mu}_i,t > \mu_1 - \epsilon \}.
\]

This allows us to establish lower bound

\[
\mathbb{P}_t(\mathcal{E}_i^\prime(t)|T_{1,t} = s, \hat{\mu}_{1,s} = \mu, \hat{\sigma}_{1,s} = \sigma) \geq \begin{cases} 
\frac{1}{2} \mathbb{P}_t(\theta_{1,t} > \mu_1 - \epsilon) & \text{if } \mu \geq \mu_1 \\
\mathbb{P}_t(\theta_{1,t} > \mu_1 - \epsilon) & \text{if } \mu < \mu_1.
\end{cases} \tag{24}
\]

Define the conditional version of \( G_{1s} \) by

\[
\tilde{G}_{1s} = G_{1s}|\hat{\mu}_{i,s}, \hat{\sigma}_{1,s} = \beta = \mathbb{P}_t(\hat{\mu}_{1,t} > \mu_1 - \epsilon \text{ or } \hat{c}_i(1,t) \leq \tau + \epsilon | \hat{\mu}_{1,s}, \hat{\sigma}_{1,s} = \beta).
\]

Define \( c_1 = \frac{1}{\sqrt{2\pi\sigma_i^2}} \) and \( c_2 = \frac{1}{2^2 \Gamma(s/2)\sigma_i^2} \). We partition the parameter space

\[
(\beta, \mu) \in [0, \infty) \times (-\infty, \infty) = A' \cup B'
\]

where \( A' = [0, \infty) \times [\mu_1, \infty) \) and \( B' = [0, \infty) \times (-\infty, \mu_1) \). We can then partition \( E \left[ \frac{1}{G_{1s}} - 1 \right] \) into two parts:

\[
E \left[ \frac{1}{G_{1s}} - 1 \right] = c_1 c_2 \int_0^\infty \int_{-\infty}^\infty \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp \left( -\frac{(\mu - \mu_1)^2}{2\sigma_i^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_i^2}} d\mu d\beta
\]

\[
= c_1 c_2 \left( \int_{A'} + \int_{B'} \right) \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp \left( -\frac{(\mu - \mu_1)^2}{2\sigma_i^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_i^2}} d\mu d\beta.
\]

For Part \( A' \), using the first case in (24),

\[
\frac{1 - \tilde{G}_{1s}}{G_{1s}} \leq 2(1 - \tilde{G}_{1s}).
\]
Thus,
\[
c_1c_2 \int_{A'} \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma^2}} \, d\mu \, d\beta \\
\leq 2c_1c_2 \int_{A'} (1 - \tilde{G}_{1s}) \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma^2}} \, d\mu \, d\beta \\
\leq 2c_1c_2 \int_{A'} \mathbb{P}_t(\theta_{1,t} \leq \mu_1 - \varepsilon) \mathbb{P}_t(\hat{\epsilon}_n(1, t) > \tau + \varepsilon) \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma^2}} \, d\mu \, d\beta \\
\leq 2c_1 \int_{\mu_1}^{\infty} \exp \left( -\frac{s}{2} (\mu - \mu_1 + \varepsilon)^2 \right) \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \, d\mu \cdot c_2 \int_{0}^{\infty} \beta^{s-1} e^{-\frac{\beta^2}{2\sigma^2}} \, d\beta \\
\leq 2 \exp \left( -\frac{s\varepsilon^2}{4} \right)
\]  
(25)

where (25) follows from using tail bounds on the Gaussian distribution.

For Part $B''$, using the second case (24),

\[
\frac{1 - \tilde{G}_{1s}}{G_{1s}} \leq \frac{1}{\mathbb{P}_t(\theta_{1,t} > \mu_1 - \varepsilon| \hat{\mu}_{1,s} = \mu)}.
\]

Then, reusing a calculation in Zhu and Tan [2020],

\[
c_1c_2 \int_{B''} \frac{1 - \tilde{G}_{1s}}{G_{1s}} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma^2}} \, d\mu \, d\beta \\
\leq c_1c_2 \int_{B''} \frac{1}{\mathbb{P}_t(\theta_{1,t} > \mu_1 - \varepsilon| \hat{\mu}_{1,s} = \mu)} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma^2}} \, d\mu \, d\beta \\
\leq c_1 \int_{-\infty}^{\mu_1 - \varepsilon} \frac{1}{\mathbb{P}_t(\theta_{1,t} > \mu_1 - \varepsilon| \hat{\mu}_{1,s} = \mu)} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \, d\mu \cdot c_2 \int_{0}^{\infty} \beta^{s-1} e^{-\frac{\beta^2}{2\sigma^2}} \, d\beta \\
\leq \frac{\sqrt{2}}{\sqrt{\pi}} \cdot \frac{3}{\sqrt{2\pi\varepsilon^2 s}} \cdot \exp \left( -\frac{s\varepsilon^2}{2} \right) = \frac{3}{\pi\varepsilon^2 s} \sqrt{8} \exp \left( -\frac{s\varepsilon^2}{2} \right) 
\]

Combining both parts, we can upper bound $\mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right]$ by

\[
\mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right] \leq 2 \exp \left( -\frac{s\varepsilon^2}{4} \right) + \frac{3}{\pi\varepsilon^2 s} \sqrt{8} \exp \left( -\frac{s\varepsilon^2}{2} \right). 
\]

Summing over $s$, we have

\[
\sum_{s=1}^{n} \mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right] \leq \frac{C_7^{(a)}}{\varepsilon^2} + C_8^{(a)}. 
\]

(b) Suppose $i$ is a non-deceiver arm. Then

\[
E_i(t) = \{ \hat{\epsilon}_n(i, t) > \tau + \varepsilon \} \cap \{ \hat{\mu}_{i,t} \geq \mu_1 + \varepsilon \} \Rightarrow E_i^c(t) \supseteq \{ \hat{\mu}_{i,t} < \mu_1 + \varepsilon \}. 
\]

This allows us to establish lower bound

\[
\mathbb{P}_t(E_i^c(t)|T_{1,t} = s, \hat{\mu}_{1,s} = \mu, \hat{\sigma}_{1,s} = \sigma) \geq \begin{cases} 
\frac{1}{2} & \text{if } \mu < \mu_1 \\
\mathbb{P}_t(\theta_{1,t} < \mu_1 + \varepsilon) & \text{if } \mu \geq \mu_1
\end{cases}
\]  
(26)

Define the conditional version of $G_{1s}$ by

\[
\tilde{G}_{1s} = G_{1s}|_{\hat{\mu}_{1,s}, \hat{\sigma}_{1,s} = \beta} = \mathbb{P}_t(\hat{\mu}_{1,t} \leq \mu_1 + \varepsilon \text{ or } \hat{\epsilon}_n(1, t) > \tau - \varepsilon | \hat{\mu}_{1,s}, \hat{\sigma}_{1,s} = \beta).
\]

Define $c_1 = \frac{1}{\sqrt{2\pi\sigma_1^4}}$ and $c_2 = \frac{1}{2\pi \Gamma(s/2)\sigma_1^4}$, we partition the parameter space

\[(\beta, \mu) \in [0, \infty) \times (-\infty, \infty) = A'' \cup B''\]
where $A'' = [0, \infty) \times [\mu_1, \infty)$ and $B'' = [0, \infty) \times (-\infty, \mu_1]$. We can then partition $\mathbb{E}\left[ \frac{1}{G_{1s}} - 1 \right]$ into two parts:

\[
\mathbb{E}\left[ \frac{1}{G_{1s}} - 1 \right] = c_1 c_2 \int_0^\infty \int_{-\infty}^\infty \frac{1 - G_{1s}}{G_{1s}} \exp \left( -\frac{(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \, d\beta
\]

\[
= c_1 c_2 \left( \int_{A''} + \int_{B''} \right) \frac{1 - G_{1s}}{G_{1s}} \exp \left( -\frac{(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \, d\beta.
\]

For Part $B''$, using the first case in (26),

\[
\frac{1 - G_{1s}}{G_{1s}} \leq 2(1 - G_{1s}).
\]

Thus,

\[
c_1 c_2 \int_{B''} \frac{1 - G_{1s}}{G_{1s}} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \, d\beta
\]

\[
\leq 2c_1 c_2 \int_{B''} (1 - G_{1s}) \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \, d\beta
\]

\[
\leq 2c_1 c_2 \int_{B''} \mathbb{P}_t(\theta_{1,t} \geq \mu_1 + \varepsilon) \mathbb{P}_t(\hat{\epsilon}_a(1,t) \leq \tau - \varepsilon) \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \, d\beta
\]

\[
\leq 2c_1 \int_{-\infty}^{\mu_1} \mathbb{P}_t(\theta_{1,t} \geq \mu_1 + \varepsilon) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\beta \]

\[
\leq 2 \exp \left( -\frac{s\varepsilon^2}{4} \right).
\]

For Part $A''$, using the second case in (26),

\[
\frac{1 - G_{1s}}{G_{1s}} \leq \frac{1}{\mathbb{P}_t(\theta_{1,t} < \mu_1 + \varepsilon | \hat{\mu}_{1,s} = \mu)}.
\]

By reusing a calculation from Part A of the proof of Lemma 3,

\[
c_1 c_2 \int_{A''} \frac{1 - G_{1s}}{G_{1s}} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \, d\beta
\]

\[
\leq c_1 c_2 \int_{A''} \frac{1}{\mathbb{P}_t(\theta_{1,t} - \mu_1 \leq \varepsilon | \hat{\mu}_{1,s} = \mu)} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \, d\beta
\]

\[
\leq c_1 \int_{\mu_1 + \varepsilon}^{\infty} \frac{1}{\mathbb{P}_t(\theta_{1,t} - \mu_1 \leq \varepsilon | \hat{\mu}_{1,s} = \mu)} \exp \left( -\frac{s(\mu - \mu_1)^2}{2\sigma_1^2} \right) \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\mu \cdot c_2 \int_0^\infty \beta^{s-1} e^{-\frac{\beta^2}{2\sigma_1^2}} \, d\beta
\]

\[
\leq \frac{3}{\pi^{\varepsilon^2 s\sqrt{s}}} \exp \left( -\frac{s\varepsilon^2}{2} \right).
\]

Combining both parts, we can upper bound $\mathbb{E}\left[ \frac{1}{G_{1s}} - 1 \right]$ by

\[
\mathbb{E}\left[ \frac{1}{G_{1s}} - 1 \right] \leq 2 \exp \left( -\frac{s\varepsilon^2}{4} \right) + \frac{3}{\pi^{\varepsilon^2 s\sqrt{s}}} \exp \left( -\frac{s\varepsilon^2}{2} \right)
\]

Summing over $s$, we have

\[
\sum_{s=1}^n \mathbb{E}\left[ \frac{1}{G_{1s}} - 1 \right] \leq \frac{C_{7}^{(b)}}{\varepsilon^2} + C_{8}^{(b)}.
\]

Setting $C_7 = \max\left\{ C_{7}^{(a)}, C_{7}^{(b)} \right\}$ and $C_8 = \max\left\{ C_{8}^{(a)}, C_{8}^{(b)} \right\}$, we get

\[
\sum_{s=1}^n \mathbb{E}\left[ \frac{1}{G_{1s}} - 1 \right] \leq \frac{C_7}{\varepsilon^2} + C_8.
\]
Lemma 11 (Upper bounding the second term of (22)). For $\xi \in (0, 1)$,

$$
\mathbb{P}(E_i^c(t)|T_{i,t} = s, \hat{\mu}_{i,t} = \mu, \hat{\sigma}_{i,t} = \sigma) \\
\leq \exp \left(-\frac{s}{2} \left((\mu - \mu_i) + \frac{\xi(1-\alpha)(\Delta_r(i,\alpha) - \varepsilon)}{\alpha}\right)^2\right) + \exp \left(-sh \left(\frac{\sigma^2(c^*_\alpha)^2}{(\sigma, c^*_\alpha - (1-\xi)(\Delta_r(i,\alpha) - \varepsilon))^2}\right)\right).
$$

Furthermore,

$$
\sum_{s=1}^{n} \mathbb{P}_t \left(G_{is} > \frac{1}{n}\right) \leq 1 + \max \left\{ \frac{2\alpha^2 \log(2n)}{\xi^2(1-\alpha)^2 (\Delta_r(i,\alpha) - \varepsilon)^2}, \frac{\log(2n)}{h(\frac{\sigma^2(c^*_\alpha)^2}{(\sigma, c^*_\alpha - (1-\xi)(\Delta_r(i,\alpha) - \varepsilon))^2})}\right\} + \frac{C_9}{\varepsilon^2} + \frac{C_{10}}{\varepsilon^2},
$$

where $C_9, C_{10}$ are constants.

Proof of Lemma 11. The result follows immediately from Lemma 8 by replacing $(c_\alpha(1), \Delta_{\text{risk}}(i,\alpha))$ with $(\tau, \Delta_r(i,\alpha))$.

Corollary 2. Let $F_{i,\xi}^\alpha$ be defined as in Theorem 2. Then choosing

$$
\xi_\alpha = 1 - \frac{\sigma_\alpha}{\Delta_r(i,\alpha)} \left(1 - \frac{1}{c_\alpha(2-\alpha)}\right)
$$

yields $\xi_\alpha \to 1^-$ and $F_{i,\xi_\alpha}^\alpha \to 0$ as $\alpha \to 1^-$. 

Proof of Corollary 2. The result follows immediately from Corollary 1 by replacing $(c_\alpha(1), \Delta_{\text{risk}}(i,\alpha))$ with $(\tau, \Delta_r(i,\alpha))$.

C Proof of Theorem 3

The proof of Theorem 3 follows the same strategy and is even more straightforward. Nevertheless, we include it here for completeness.

Proof of Theorem 3: For any arm $i \in \mathcal{K}_r \setminus \mathcal{K}^*$, define the good event by

$$
E_i(t) = \{\hat{\mu}_{i,t} > \mu_1 + \varepsilon\}.
$$

We first observe that

$$
E_i^c(t) = \{\hat{\mu}_{i,t} \leq \mu_1 + \varepsilon\}.
$$

By Lemma 1, and the linearity of expectation, we can divide $\mathbb{E}[T_{i,n}]$ into two parts as

$$
\mathbb{E}[T_{i,n}] \leq \sum_{s=0}^{n-1} \mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right] + \sum_{s=0}^{n-1} \mathbb{P}(G_{is} > \frac{1}{n}) + 1.
$$

(27)

By the computations in part (b) of the proof of Lemma 10, we have

$$
\sum_{s=1}^{n} \mathbb{E} \left[ \frac{1}{G_{1s}} - 1 \right] \leq \frac{C_{11}}{\varepsilon^2} + C_{12}.
$$

(28)

By a Gaussian concentration bound,

$$
\mathbb{P}_t(E_i^c(t)) = \mathbb{P}_t(\theta_{i,t} - \mu \leq -(\mu - \mu_1 - \varepsilon)) \leq \exp \left(-\frac{s}{2} (\mu - \mu_1 - \varepsilon)^2\right).
$$

By similar computations in Zhu and Tan [2020],

$$
\sum_{s=1}^{n} \mathbb{P}_t \left(G_{is} > \frac{1}{n}\right) \leq 1 + \frac{2 \log(n)}{(\mu_1 - \mu_1 - \varepsilon)^2} + \frac{C_{13}}{\varepsilon^4} + \frac{C_{14}}{\varepsilon^2}.
$$

(29)

Plugging (28) and (29) into (27), we have the expected number of pulls on arm $i$ up to round $n$ given by

$$
\mathbb{E}[T_{i,n}] \leq \frac{2 \log(n)}{(\mu_1 - \mu_1 - \varepsilon)^2} + \frac{C_9}{\varepsilon^2} + \frac{C_{10}}{\varepsilon^2} + C'_{11}.
$$

Setting $\varepsilon = (\log n)^{-\frac{1}{5}}$, we get the following result for the suboptimality regret:

$$
\limsup_{n \to \infty} \frac{\mathcal{R}_{\text{sub}}^{\text{1/2}}(\text{CVaR-TS})}{\log n} \leq \sum_{i \in \mathcal{K}_r \setminus \mathcal{K}^*} \frac{2}{(\mu_i - \mu_1)^2} \Delta(i).
$$

□