Tail-Risk-Aware Learning in Cooperative Stackelberg Bandits

author names withheld

Under Review for NExT-Game 2026

Abstract

We study cooperative Stackelberg bandits where a leader acts first, a follower observes only that action, and both agents seek to maximize the lower-tail CVaR of the realized reward. Mean reward can hide rare failures, and the Stackelberg information pattern rules out centralized joint-action control. We prove an information-theoretic lower bound for spectral risk objectives, which gives $\Omega(\sqrt{ABK/\tau})$ for CVaR. We also propose a Stackelberg-factorized CVaR-UCB algorithm based on public pairwise shortfall indices, and prove a matching $\widetilde{O}(\sqrt{ABK/\tau})$ regret bound without proxy-realizability or pointwise tail-variance assumptions.

1. Introduction

Many learning systems are neither fully centralized nor symmetric. A platform sets allocation rules, an AI service chooses a routing or safety policy, or an operator sets a control level, while downstream agents respond locally. These settings share a Stackelberg information pattern: the first action is also the only coordination signal available to the second decision maker.

In safety-critical settings, average performance is often the wrong target. A policy may look good on average while still producing rare routing collapses, unsafe interventions, severe service degradation, or coordinated failures. These events live in the lower tail. Conditional value-at-risk (CVaR), and more generally spectral risk measures, are standard tools for evaluating such outcomes (Artzner et al., 1999; Acerbi, 2002; Rockafellar and Uryasev, 2000). The question is how to learn these risk-sensitive objectives when the agents are cooperative but cannot freely communicate.

This problem lies between Stackelberg learning and risk-sensitive online learning. Work on online Stackelberg games largely studies expected utilities or equilibrium identification under bandit feedback (Blum et al., 2014; Balcan et al., 2015; Bai et al., 2021; Fiez et al., 2020). Risk-sensitive bandit and reinforcement-learning results quantify the cost of estimating CVaR, but mainly in single-agent settings (Wang et al., 2023). We consider the case where tail-risk estimation and leader-follower exploration occur at the same time, while the follower observes only the leader's current action.

We study a minimal model for this interaction. At each round, a leader chooses $a_k \in \mathcal{A}$, a follower observes a_k and chooses $b_k \in \mathcal{B}$, and both observe a reward from $\nu(a_k, b_k)$. The shared objective is lower-tail CVaR. There is no message passing, counterfactual feedback, or centralized joint-action planner, so the leader action alone must coordinate tail-risk estimation and conditional follower exploration.

© .

The algorithm certifies the realizable Stackelberg value directly. Both agents maintain the same public confidence table over pairs (a, b), with each entry given by an optimistic CVaR index from the shortfall representation in (1). For a leader action a, the certified follower value is the largest pair index over $b \in \mathcal{B}$; the leader chooses the action with the largest certified value. This keeps the Stackelberg protocol intact and avoids artificial diversification from portfolio-style leader proxies.

We formulate cooperative Stackelberg bandits for lower-tail CVaR and bounded spectral risks. For CVaR, we prove the lower bound

$$\Omega\left(\sqrt{\frac{K(AB-1)}{\tau}}\right).$$

We also give a decentralized CVaR-UCB algorithm with high-probability regret

$$\mathrm{Reg}_\tau(K) \le \widetilde{O}\left(\sqrt{\frac{ABK}{\tau}}\right).$$

Thus both the joint-action cost AB and the tail-risk cost $1/\tau$ are unavoidable up to logarithmic factors.

2. Model

Let $\mathcal{A}$ and $\mathcal{B}$ be finite action sets with $|\mathcal{A}| = A$ and $|\mathcal{B}| = B$. At round $k = 1, \ldots, K$, the leader first selects $a_k \in \mathcal{A}$. The follower observes a_k and then selects $b_k \in \mathcal{B}$. Both agents observe (a_k, b_k, r_k), where $r_k \sim \nu(a_k, b_k)$ and each distribution $\nu(a, b)$ is supported on $[0, 1]$. No other communication is allowed.

For a reward random variable X, we use the lower-tail CVaR utility

$$\mathrm{CVaR}_\tau(X) = \max_{c\in[0,1]} \left\{ c - \frac{1}{\tau}\mathbb{E}[(c - X)^+] \right\}, \qquad \tau \in (0, 1]. \tag{1}$$

The cooperative Stackelberg optimum is

$$(a^\star, b^\star) \in \arg\max_{a\in\mathcal{A}, b\in\mathcal{B}} \mathrm{CVaR}_\tau(\nu(a, b)).$$

The regret after K rounds is

$$\mathrm{Reg}_\tau(K) = \sum_{k=1}^{K} \left(\mathrm{CVaR}_\tau(\nu(a^\star, b^\star)) - \mathrm{CVaR}_\tau(\nu(a_k, b_k))\right). \tag{2}$$

We also use spectral risk utilities. For an integrable reward X with quantile function F_X^{-1},

$$M_\phi(X) = \int_0^1 \phi(u) F_X^{-1}(u)\, du,$$

where $\phi \ge 0$, $\int_0^1 \phi(u)du = 1$, and $C_\phi = \|\phi\|_\infty < \infty$. For lower-tail risk aversion in a reward-maximization problem, larger weight is placed on smaller quantiles; CVaR corresponds to $\phi(u) = \tau^{-1}\mathbb{I}\{u \le \tau\}$.

3. Lower Bound

The lower bound follows from the sensitivity of spectral risk near the quantile where the spectrum places its largest mass.

Lemma 1 (local spectral separation) *For any bounded spectrum ϕ and any $\epsilon > 0$, there are Bernoulli distributions P and Q whose means differ by ϵ and whose spectral utilities satisfy*

$$|M_\phi(Q) - M_\phi(P)| \geq \epsilon(C_\phi - o(1)).$$

Theorem 2 (spectral lower bound) *For any online algorithm and any bounded spectral risk measure, there exists a cooperative Stackelberg bandit instance with Bernoulli rewards such that, whenever $K \geq (AB-1)/(6u^\star)$,*

$$\mathbb{E}[\mathrm{Reg}_\phi(K)] \geq \frac{C_\phi}{40e}\sqrt{K(AB-1)u^\star},$$

where $u^\star$ is a quantile at which the spectrum is arbitrarily close to its essential supremum.

For CVaR, $C_\phi = 1/\tau$ and $u^\star = \tau$. Hence Theorem 2 gives the following minimax lower bound.

Corollary 3 (CVaR lower bound) *For every algorithm, there is a cooperative Stackelberg bandit instance with A leader actions and B follower actions such that*

$$\mathbb{E}[\mathrm{Reg}_\tau(K)] \geq \frac{1}{40e}\sqrt{\frac{K(AB-1)}{\tau}}.$$

4. Decentralized CVaR-UCB

The algorithm uses the shortfall form (1) and a public confidence table over executable pairs. For each pair define

$$\mu(c,a,b) = \mathbb{E}_{R\sim\nu(a,b)}[(c-R)^+].$$

Let $\mathcal{C}_\varepsilon$ be an ε-net of $[0,1]$. The algorithm initializes every pair once, so after initialization

$$N_k(a,b) = \sum_{i<k} \mathbb{I}\{(a_i,b_i) = (a,b)\} \geq 1.$$

For $c \in \mathcal{C}_\varepsilon$, define

$$\widehat{\mu}_k(c,a,b) = \frac{1}{N_k(a,b)}\sum_{i<k}(c-r_i)^+\mathbb{I}\{(a_i,b_i) = (a,b)\}.$$

Set

$$L = \log\frac{2ABK|\mathcal{C}_\varepsilon|}{\delta}, \qquad \beta_k(a,b) = \sqrt{\frac{2\tau L}{N_k(a,b)}} + \frac{2L}{3N_k(a,b)}.$$

Algorithm 1: Stackelberg-Factorized CVaR-UCB

Input: risk tolerance τ, horizon K, confidence δ, grid $\mathcal{C}_\varepsilon$

Sample each pair (a, b) once and initialize N, $\widehat{\mu}$, β, and U by (3);

For each $a \in \mathcal{A}$, set $G(a) = \max_{b\in\mathcal{B}} U(a, b)$ and $b^{\text{cert}}(a) \in \arg\max_b U(a, b)$;

Set $a^{\text{cert}} \in \arg\max_a G(a)$;

for *each subsequent round* k **do**

Leader plays

$$a_k = a^{\text{cert}}.$$

Follower observes a_k and plays

$$b_k = b^{\text{cert}}(a_k).$$

Observe $r_k \sim \nu(a_k, b_k)$;

For every $c \in \mathcal{C}_\varepsilon$, update the shortfall sum for (a_k, b_k) by adding $(c - r_k)^+$;

Update $N(a_k, b_k)$, recompute $\beta(a_k, b_k)$ and $U(a_k, b_k)$;

Refresh only

$$G(a_k) = \max_{b\in\mathcal{B}} U(a_k, b), \qquad b^{\text{cert}}(a_k) \in \arg\max_{b\in\mathcal{B}} U(a_k, b),$$

and then refresh $a^{\text{cert}} \in \arg\max_{a\in\mathcal{A}} G(a)$;

end

The pairwise optimistic shortfall index is

$$U_k(a, b) = \min\left\{1, \max_{c\in\mathcal{C}_\varepsilon}\left[c - \frac{1}{\tau}\big(\widehat{\mu}_k(c, a, b) - \beta_k(a, b)\big)\right] + \frac{2\varepsilon}{\tau}\right\}. \tag{3}$$

For a leader action a, define the certified follower-response value

$$G_k(a) = \max_{b\in\mathcal{B}} U_k(a, b).$$

The leader maximizes $G_k(a)$. After observing a_k, the follower maximizes the same table over b.

Write

$$u(a, b) = \mathrm{CVaR}_\tau(\nu(a, b)), \qquad g(a) = \max_{b\in\mathcal{B}} u(a, b).$$

Lemma 4 (grid approximation) *For $\psi(c) = c - \tau^{-1}\mathbb{E}[(c - R)^+]$ and any ε-net $\mathcal{C}_\varepsilon$ of $[0, 1]$,*

$$0 \le \max_{c\in[0,1]} \psi(c) - \max_{c\in\mathcal{C}_\varepsilon} \psi(c) \le \left(1 + \frac{1}{\tau}\right)\varepsilon \le \frac{2\varepsilon}{\tau}.$$

Lemma 5 (tail-budget Bernstein pessimism) *For each pair (a, b), choose a lower τ-quantile*

$$c^\star(a, b) \in \arg\max_{c\in[0,1]}\left\{c - \frac{1}{\tau}\mu(c, a, b)\right\}, \qquad \Pr_{R\sim\nu(a,b)}(R < c^\star(a, b)) \le \tau.$$

Then, with probability at least $1-\delta/2$*, for all* $k \le K$ *and all* (a,b)*,*

$$\widehat{\mu}_k(c^\star(a,b),a,b) - \beta_k(a,b) \le \mu(c^\star(a,b),a,b).$$

Consequently, $u(a,b) \le U_k(a,b)$ *for all* k,a,b*.*

Lemma 6 (empirical CVaR concentration) *Let* $\widehat{u}_n(a,b)$ *be the empirical reward-side CVaR computed from the first* n *samples of pair* (a,b)*. With probability at least* $1-\delta/2$*, for all pairs and all sample sizes* $n \le K$ *with* $n \ge 25L$*,*

$$\widehat{u}_n(a,b) - u(a,b) \le \sqrt{\frac{3L}{n\tau}} + \frac{15L}{n\tau}.$$

Lemma 7 (Stackelberg optimism) *On the event of Lemma 5, for every round* k*,*

$$g(a^\star) \le G_k(a_k) = U_k(a_k,b_k).$$

Lemma 8 (one-step overestimation) *On the event of Lemma 5, for every round* k*,*

$$g(a^\star) - u(a_k,b_k) \le U_k(a_k,b_k) - u(a_k,b_k).$$

Lemma 9 (cumulative Bernstein-CVaR overestimation) *Let* $M = AB$*. On the events of Lemmas 5 and 6, for any adaptive sequence of played pairs* $x_k = (a_k,b_k)$*,*

$$\sum_{k=1}^{K} [U_k(x_k) - u(x_k)] \le \frac{2K\varepsilon}{\tau} + 2\sqrt{\frac{10MKL}{\tau}} + 16\frac{ML(1+\log K)}{\tau} + 25ML.$$

Theorem 10 (upper bound) *Run Algorithm 1 with a public* ε*-net* $\mathcal{C}_\varepsilon$*. With probability at least* $1-\delta$*,*

$$\mathrm{Reg}_\tau(K) \le \frac{2K\varepsilon}{\tau} + 2\sqrt{\frac{10ABKL}{\tau}} + 16\frac{ABL(1+\log K)}{\tau} + 25ABL.$$

Choosing $\varepsilon = 1/K$ *gives*

$$\mathrm{Reg}_\tau(K) = \widetilde{O}\left(\sqrt{\frac{ABK}{\tau}}\right).$$

The proof in Appendix B does not use a uniform Bernstein band over all budgets c. Tail-budget Bernstein gives optimism, and empirical-CVaR concentration together with the summed shortfall bonus controls overestimation.

Relation to prior work. Online Stackelberg learning mostly studies expected utilities (Blum et al., 2014; Balcan et al., 2015; Bai et al., 2021; Fiez et al., 2020), while CVaR bandit and RL results are mostly single-agent (Wang et al., 2023). Our setting keeps the objectives aligned but imposes strict Stackelberg information and a lower-tail risk objective.

Table 1: Two-point Stackelberg bandit. Lower regret/unsafe rate and higher tail CVaR are better.

Algorithm	CVaR regret	Tail CVaR	Unsafe rate
Scheme B shortfall-UCB	265.9 ± 16.5	0.607 ± 0.005	0.003 ± 0.001
Joint shortfall-UCB	265.9 ± 16.5	0.607 ± 0.005	0.003 ± 0.001
Mean-UCB	2224.9 ± 94.5	0.215 ± 0.008	0.240 ± 0.036

5. Numerical Illustration

We use a two-point instance with explicit means and CVaRs. Here $A = B = 5$, $\tau = 0.1$, $K = 6000$, and 40 trials. The mean-optimal pair $(0, 1)$ has mean 0.880 but CVaR 0.000, whereas the CVaR-optimal pair $(3, 2)$ has mean 0.743 and CVaR 0.680.

Table 1 reports final averages. Mean-UCB is attracted to the unsafe high-mean pair, while Scheme B almost never selects it and obtains much higher realized lower-tail reward. In this public-table bandit implementation, Scheme B and the centralized shortfall-UCB baseline make the same pair selections; the distinction is that Scheme B executes the same certificate through the leader-follower protocol.

6. Discussion

Decentralized optimism recovers the minimax joint-action and tail-risk scaling up to logarithmic factors despite the Stackelberg information constraint. The model is small, but it captures a persistent difficulty: tail events require targeted exploration, while hierarchical information constraints make exploration nonstationary across agents. Natural extensions include stateful Stackelberg MDPs, multiple followers, delayed or private observations, and AI service-routing benchmarks.

References

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

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

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. In *Advances in Neural Information Processing Systems*, volume 34, pages 25799–25811, 2021.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 61–78, 2015.

Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Learning optimal commitment to overcome insecurity. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Tanner Fiez, Benjamin Chasnov, and Lillian J. Ratliff. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3133–3144, 2020.

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

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near minimax optimal risk-sensitive reinforcement learning with CVaR. In *Proceedings of the 40th International Conference on Machine Learning*, pages 35864–35907, 2023.

Appendix A. Proofs

A.1. Proof of Lemma 1

Let $u^\star$ be a point at which ϕ is arbitrarily close to its essential supremum C_ϕ. Set $P = \mathrm{Ber}(1-u^\star)$ and $Q = \mathrm{Ber}(1-u^\star+\epsilon)$, assuming ϵ is small enough that the parameter remains in $[0,1]$. The quantile functions differ only on an interval of length ϵ adjacent to $u^\star$. Hence

$$|M_\phi(Q) - M_\phi(P)| = \int_{u^\star-\epsilon}^{u^\star} \phi(u)\, du.$$

By the definition of essential supremum, this integral is at least $\epsilon(C_\phi - o(1))$ along a sequence of intervals shrinking to a point at which the spectrum is essentially maximal. The Bernoulli means differ by exactly ϵ.

A.2. Proof of Theorem 2

We first prove the case $A = B = 2$. Consider two environments. In M_1, the optimal pair is (a_1, b_2); in M_2, the optimal pair is (a_1, b_1). The rewards are Bernoulli:

M_1	b_1	b_2
a_1	$\mathrm{Ber}(1-u^\star-\epsilon)$	$\mathrm{Ber}(1-u^\star+\epsilon)$
a_2	$\mathrm{Ber}(1-u^\star)$	$\mathrm{Ber}(1-u^\star)$

M_2	b_1	b_2
a_1	$\mathrm{Ber}(1-u^\star+2\epsilon)$	$\mathrm{Ber}(1-u^\star+\epsilon)$
a_2	$\mathrm{Ber}(1-u^\star)$	$\mathrm{Ber}(1-u^\star)$.

Here $\epsilon > 0$ will be chosen later and is small enough that all Bernoulli parameters lie in $[0,1]$. By Lemma 1, each adjacent change in the Bernoulli parameter around $u^\star$ creates a spectral-utility gap at least of order ϵC_ϕ. We absorb the arbitrarily small approximation term in the universal constants below.

Let

$$T_{a,b}(K) = \sum_{k=1}^{K} \mathbb{I}\{(a_k, b_k) = (a, b)\}.$$

In M_1, pair (a_1, b_2) is optimal. Hence

$$\begin{aligned}\mathbb{E}_{M_1}[\mathrm{Reg}_\phi(K)] &\geq \epsilon C_\phi\, \mathbb{E}_{M_1}[K - T_{a_1,b_2}(K)] \\ &\geq \frac{K\epsilon C_\phi}{2} \Pr_{M_1}\left(T_{a_1,b_2}(K) \leq \frac{K}{2}\right).\end{aligned}$$

In M_2, pair (a_1, b_1) is optimal, so playing (a_1, b_2) incurs at least ϵC_ϕ regret. Therefore

$$\mathbb{E}_{M_2}[\mathrm{Reg}_\phi(K)] \geq \frac{K\epsilon C_\phi}{2} \Pr_{M_2}\left(T_{a_1,b_2}(K) \geq \frac{K}{2}\right).$$

Applying the Bretagnolle-Huber inequality to the event $E = \{T_{a_1,b_2}(K) \leq K/2\}$ yields

$$\begin{aligned}&\mathbb{E}_{M_1}[\mathrm{Reg}_\phi(K)] + \mathbb{E}_{M_2}[\mathrm{Reg}_\phi(K)] \\ &\quad\geq \frac{K\epsilon C_\phi}{4} \exp\{-\mathrm{KL}(\mathbb{P}_{M_1}, \mathbb{P}_{M_2})\}.\end{aligned}$$

Only observations from (a_1, b_1) distinguish M_1 and M_2, so

$$\mathrm{KL}(\mathbb{P}_{M_1}, \mathbb{P}_{M_2}) = \mathbb{E}_{M_1}[T_{a_1,b_1}(K)]\,\mathrm{KL}(\mathrm{Ber}(1-u^\star-\epsilon), \mathrm{Ber}(1-u^\star+2\epsilon))\,.$$

We choose the perturbed nonoptimal pair to be the least-sampled nonoptimal pair in M_1, so in the two-action case $\mathbb{E}_{M_1}[T_{a_1,b_1}(K)] \le K/3$. For Bernoulli distributions,

$$\begin{aligned}&\mathrm{KL}(\mathrm{Ber}(1-u^\star-\epsilon), \mathrm{Ber}(1-u^\star+2\epsilon))\\ &= (1-u^\star-\epsilon)\log\frac{1-u^\star-\epsilon}{1-u^\star+2\epsilon}(u^\star+\epsilon)\log\frac{u^\star+\epsilon}{u^\star-2\epsilon}.\end{aligned}$$

When $u^\star \le 1/2$, using $-\log(1-x) \le 2x$ for $x \in [0, 1/2]$ gives

$$\mathrm{KL}(\mathrm{Ber}(1-u^\star-\epsilon), \mathrm{Ber}(1-u^\star+2\epsilon)) \le \frac{18\epsilon^2}{u^\star}.$$

The case $u^\star > 1/2$ is handled symmetrically by applying the same argument to the complementary Bernoulli mass, after decreasing constants if necessary. Taking $\epsilon^2 = u^\star/(6K)$ gives

$$\mathbb{E}_{M_1}[\mathrm{Reg}_\phi(K)] + \mathbb{E}_{M_2}[\mathrm{Reg}_\phi(K)] \ge \frac{C_\phi}{12e}\sqrt{Ku^\star}.$$

Thus one of the two environments has expected regret at least $(C_\phi/(24e))\sqrt{Ku^\star}$.

For general A and B, keep one pair as the optimal pair and choose the changed pair among the $AB-1$ nonoptimal pairs with smallest expected count under the base environment. Then

$$\mathbb{E}_{M_1}[T_{\text{changed}}(K)] \le \frac{K}{AB-1}.$$

Repeating the same argument with $\epsilon^2 = (AB-1)u^\star/(6K)$ gives

$$\mathbb{E}[\mathrm{Reg}_\phi(K)] \ge \frac{C_\phi}{40e}\sqrt{K(AB-1)u^\star},$$

after absorbing constants. This proves the theorem.

A.3. Proof of Corollary 3

For the lower-tail CVaR utility, the spectrum is $\phi(u) = \tau^{-1}\mathbb{I}\{u \le \tau\}$. Thus $C_\phi = 1/\tau$ and the maximally weighted quantile satisfies $u^\star = \tau$. Substituting these values into Theorem 2 gives the stated lower bound.

Appendix B. Upper-Bound Proofs

B.1. Proof of Lemma 4

For any $c, c' \in [0, 1]$, the map $c \mapsto (c-R)^+$ is one-Lipschitz, so

$$|\psi(c) - \psi(c')| \le |c-c'| + \tau^{-1}|c-c'| \le \left(1+\frac{1}{\tau}\right)|c-c'|.$$

Choosing $c' \in \mathcal{C}_\varepsilon$ within ε of a maximizer over $[0, 1]$ gives the first bound, and $\tau \le 1$ gives the second.

B.2. Proof of Lemma 5

Fix (a, b) and write $c^\star = c^\star(a, b)$. For $Y = (c^\star - R)^+$, we have $0 \le Y \le 1$ and $Y > 0$ only when $R < c^\star$. Hence $\mathbb{E}Y^2 \le \Pr(R < c^\star) \le \tau$, so $\mathrm{Var}(Y) \le \tau$. Bernstein's inequality and a union bound over all pairs and all sample sizes $n \le K$ give the displayed pessimism event.

Choose $\bar{c} \in \mathcal{C}_\varepsilon$ with $|\bar{c} - c^\star| \le \varepsilon$. Since $(c - r)^+$ is one-Lipschitz in c,

$$\widehat{\mu}_k(\bar{c}, a, b) \le \widehat{\mu}_k(c^\star, a, b) + \varepsilon.$$

On the pessimism event,

$$\begin{aligned}
\max_{c \in \mathcal{C}_\varepsilon} \left[c - \frac{1}{\tau}(\widehat{\mu}_k(c, a, b) - \beta_k(a, b)) \right] &\ge \bar{c} - \frac{1}{\tau}(\widehat{\mu}_k(\bar{c}, a, b) - \beta_k(a, b)) \\
&\ge c^\star - \frac{1}{\tau}\mu(c^\star, a, b) - \left(1 + \frac{1}{\tau}\right)\varepsilon \\
&\ge u(a, b) - \frac{2\varepsilon}{\tau}.
\end{aligned}$$

The extra $2\varepsilon/\tau$ in (3) gives $u(a, b) \le U_k(a, b)$.

B.3. Proof of Lemma 6

This is the standard empirical CVaR concentration inequality for bounded rewards, applied to the reward samples of each pair. A tape construction makes adaptive sampling equivalent to reading from independent sample tapes for each pair, and a union bound over the AB pairs and all $n \le K$ gives the simultaneous statement.

B.4. Proof of Lemma 7

By Lemma 5, $u(a, b) \le U_k(a, b)$ for every pair. Therefore

$$g(a^\star) = \max_b u(a^\star, b) \le \max_b U_k(a^\star, b) = G_k(a^\star).$$

The leader chooses $a_k \in \arg\max_a G_k(a)$, so $G_k(a^\star) \le G_k(a_k)$. The follower chooses $b_k \in \arg\max_b U_k(a_k, b)$, hence $G_k(a_k) = U_k(a_k, b_k)$.

B.5. Proof of Lemma 8

This follows immediately from Lemma 7.

B.6. Proof of Lemma 9

For $x = (a, b)$, let $\widehat{u}_k(x)$ be the empirical CVaR from the $N_k(x)$ samples of x. From the definition of the index and $\mathcal{C}_\varepsilon \subset [0, 1]$,

$$U_k(x_k) \le \widehat{u}_k(x_k) + \frac{\beta_k(x_k)}{\tau} + \frac{2\varepsilon}{\tau}.$$

Thus

$$U_k(x_k) - u(x_k) \le \frac{\beta_k(x_k)}{\tau} + [\widehat{u}_k(x_k) - u(x_k)] + \frac{2\varepsilon}{\tau}.$$

For rounds with $N_k(x_k) < 25L$, use $U_k(x_k) - u(x_k) \le 1$; each pair contributes fewer than $25L$ such rounds, for total at most $25ML$. On the remaining rounds, Lemma 6 gives

$$\frac{\beta_k(x_k)}{\tau} + \widehat{u}_k(x_k) - u(x_k) \le \sqrt{\frac{10L}{\tau N_k(x_k)}} + \frac{16L}{\tau N_k(x_k)}.$$

Finally,

$$\sum_{k=1}^{K} \frac{1}{\sqrt{N_k(x_k)}} \le 2\sqrt{MK}, \qquad \sum_{k=1}^{K} \frac{1}{N_k(x_k)} \le M(1 + \log K),$$

which proves the claim.

B.7. Proof of Theorem 10

By Lemma 8,

$$\mathrm{Reg}_\tau(K) \le \sum_{k=1}^{K} \left[U_k(a_k, b_k) - u(a_k, b_k) \right].$$

Apply Lemma 9 with $M = AB$.