
DEEP REINFORCEMENT LEARNING

HOMework 6 | RL THEORY, EXPLORATION

Professor Mohammad Hossein Rohban

Sharif University of Technology
Spring 2026



Contents

1	Thompson Sampling	1
2	Action Elimination for Multi-Armed Bandits	4
3	RND vs Bootstrapped DQN on DeepSea	6
4	Variance-Aware UCB	8
4.1	Empirical Bernstein Confidence Bounds	9
4.2	The Variance-Aware UCB Index	9
4.3	Counting Pulls of a Suboptimal Arm	10
4.4	Expected Regret with a General Exploration Function	11
4.5	Logarithmic Exploration	11

Important Notes

- **Grading Policy:** This homework is considered complete once you reach a total of **100 points**. Up to *5 additional bonus points* may be earned; however, any score above 100 will still be recorded as 100. Bonus points do *not* transfer or overflow to other assignments.
- **Homework Deadline:** The assignment is due on the posted deadline. Please make sure to submit your work on time.
- **Delay Usage Policy:** You may use *at most 5 days* of your allowed late-day budget for this homework. Late days will be automatically applied if you submit after the deadline.
- **Cheating and Collaboration Policy:** You must write all solutions in your own words and produce all code independently. Discussing general ideas is allowed, but sharing written solutions, code, or using someone else's work (including from online sources, previous semesters, or AI tools without proper authorization) is strictly prohibited. Violations will result in academic misconduct procedures.

Advice

Please avoid asking LLMs for the answer. Sure, doing so might give you an immediate reward of 5 points, but in the long run, it won't help you become the next Rockstar of DRL. Remember the key lesson of this course: chasing immediate rewards is rarely the best strategy. Instead, struggle with the problems, read the papers, and truly understand the algorithms. The satisfaction of solving a challenging problem yourself is far greater than any grade. Good luck! [1]

Grading Breakdown

Task / Criteria	Points
Thompson Sampling	28 pts
Posterior, KL Concentration, and Quantile Envelope	12 pts
Pull-Count Decomposition and Logarithmic Bound	12 pts
Thompson Sampling versus UCB	4 pts
Action Elimination for Multi-Armed Bandits	18 pts
Elimination Event and Gap-Dependent Regret	10 pts
Gap-Free Bound and Comparison	8 pts
RND vs Bootstrapped DQN on DeepSea	24 pts
RND vs Bootstrapped DQN on DeepSea	16 pts
Empirical Analysis Questions	8 pts
Variance-Aware UCB	30 pts
Regret Decomposition	4 pts
Empirical Bernstein Inequality	5 pts
Variance-Aware UCB Index	4 pts
Bounding Pulls of a Suboptimal Arm	5 pts
Expected Regret for General Exploration	6 pts
Logarithmic Regret	4 pts
Interpretation	2 pt
Bonus	+5 pts
Clarity and Quality of Report	+5 pts
Total	100+5 pts

1 Thompson Sampling

Consider a stochastic K -armed Bernoulli bandit with $K \geq 2$ arms. Arm $a \in \{1, \dots, K\}$ produces i.i.d. rewards

$$X_{a,s} \sim \text{Bernoulli}(\mu_a), \quad \mu_a \in (0, 1),$$

and reward sequences are independent across arms. Assume arm 1 is uniquely optimal:

$$\mu_1 > \mu_2 \geq \dots \geq \mu_K, \quad \Delta_a = \mu_1 - \mu_a \quad (a \neq 1).$$

The learner uses independent Beta(1, 1) priors. Each arm is pulled once initially; these forced pulls are the only fixed initialization terms in the regret expressions below. For $t > K$, let $N_{a,t}$ and $S_{a,t}$ be the number of pulls and successes of arm a before round t , including the initial pull of arm a , and let

$$\hat{\mu}_{a,t} = \frac{S_{a,t}}{N_{a,t}}.$$

Thompson Sampling draws

$$\theta_{a,t} \sim \text{Beta}(S_{a,t} + 1, N_{a,t} - S_{a,t} + 1), \quad a = 1, \dots, K,$$

independently conditional on the history $\mathcal{F}_{t-1}$, and chooses an arm with largest sample. Ties are broken by a fixed rule; after the initial pulls, ties among Beta samples have probability zero. Let A_t denote the arm selected at round t . The pseudo-regret is

$$R_T = \sum_{t=1}^T (\mu_1 - \mu_{A_t}) = \sum_{a \neq 1} \Delta_a N_{a,T+1} + O(1),$$

where the $O(1)$ term only accounts for the mandatory initial pulls. For $p, q \in (0, 1)$ define

$$K(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

For horizon $T \geq 3$, define the history-measurable posterior upper quantile and its KL envelope by

$$q_{a,t} = \inf \left\{ x \in [0, 1] : \mathbb{P}(\theta_{a,t} > x \mid \mathcal{F}_{t-1}) \leq \frac{1}{t \log T} \right\},$$

$$U_{a,t} = \sup \left\{ x \in [\hat{\mu}_{a,t}, 1] : K(\hat{\mu}_{a,t}, x) \leq \frac{\log t + \log \log T}{N_{a,t}} \right\},$$

with $U_{a,t} = 1$ when $\hat{\mu}_{a,t} = 1$. Endpoint cases in posterior-tail and Bernoulli-KL expressions are interpreted by continuity.

Question 1: Posterior, KL Concentration, and Quantile Envelope

[12 Points]

- For a fixed arm with n pulls and s successes, derive the Beta($s + 1, n - s + 1$) posterior from Bayes' rule. Compute its posterior mean and variance, and show that the variance is $\Theta(1/n)$ in the interior of $(0, 1)$.
- Prove the Bernoulli Chernoff bounds in the exact KL orientation

$$\mathbb{P}(\hat{p}_n \geq q) \leq e^{-nK(q,p)} \quad (q > p), \quad \mathbb{P}(\hat{p}_n \leq q) \leq e^{-nK(q,p)} \quad (q < p),$$

where the first argument is the empirical deviation level and the second is the true Bernoulli

mean.

(c) Prove the Beta–Binomial identity

$$\mathbb{P}(Z \geq x) = \mathbb{P}(B \leq s), \quad Z \sim \text{Beta}(s+1, n-s+1), \quad B \sim \text{Binomial}(n+1, x).$$

Then prove the explicit posterior tail bound

$$\mathbb{P}(Z \geq x) \leq \exp\left(-(n+1)K\left(\frac{s}{n+1}, x\right)\right), \quad x > \frac{s}{n+1}.$$

(d) Prove the following one-way quantile-envelope statement. First show that $q_{a,t}$ is well-defined and that

$$\mathbb{P}(\theta_{a,t} > q_{a,t} \mid \mathcal{F}_{t-1}) \leq \frac{1}{t \log T}.$$

Next show that if $x \geq \hat{\mu}_{a,t}$ and

$$(N_{a,t} + 1)K\left(\frac{S_{a,t}}{N_{a,t} + 1}, x\right) \geq \log t + \log \log T,$$

then $q_{a,t} \leq x$. Finally compare this exact finite-sample statement with the displayed definition of $U_{a,t}$: explain why replacing $(N_{a,t} + 1, S_{a,t}/(N_{a,t} + 1))$ by $(N_{a,t}, \hat{\mu}_{a,t})$ changes only lower-order $O(\log N_{a,t})$ terms in the numerator of the KL radius. An exact equality between $q_{a,t}$ and $U_{a,t}$ is not required.

(e) Use the local quadratic expansion of $K(p, q)$ around $q = p$ to compare the KL envelope with the usual Hoeffding-style radius $|q - p| \lesssim \sqrt{(\log t)/n}$. Explain why posterior sampling explores without an explicit optimism bonus.

Question 2: Pull-Count Decomposition and Logarithmic Bound

[12 Points]

Fix a suboptimal arm $a \neq 1$ and choose numbers $\mu_a < x_a < y_a < \mu_1$. Let

$$\ell_a(T) = \left\lceil \frac{(1 + \varepsilon) \log T}{K(x_a, y_a)} \right\rceil, \quad \varepsilon > 0.$$

Here $\ell_a(T)$ is a deterministic allowance for the first $\ell_a(T)$ pulls of arm a ; after those pulls, each additional pull of arm a must be charged to one of the two posterior-sampling failure events displayed below. All $o(\log T)$ statements in this question are for fixed K , fixed gaps, fixed x_a, y_a , and fixed $\varepsilon > 0$ as $T \rightarrow \infty$.

(a) Prove the decomposition

$$\mathbb{E}[N_{a,T+1}] \leq \ell_a(T) + \sum_{t=K+1}^T \mathbb{P}(\theta_{1,t} \leq y_a) + \sum_{t=K+1}^T \mathbb{P}(A_t = a, N_{a,t} > \ell_a(T), \theta_{1,t} > y_a).$$

(b) Show that the last probability in part (a) is bounded by $\mathbb{P}(N_{a,t} > \ell_a(T), \theta_{a,t} > y_a)$. Use the quantile envelope from the previous question and KL concentration to prove that its sum over $t \leq T$ is $o(\log T)$; strict versus non-strict tie inequalities do not change the asymptotic bound because the Thompson samples are continuous.

(c) Prove that the optimal-arm failure term

$$\sum_{t=K+1}^T \mathbb{P}(\theta_{1,t} \leq y_a)$$

is $o(\log T)$. Your proof should make explicit where posterior consistency of the optimal arm enters.

(d) Combine the previous parts to show that

$$\mathbb{E}[N_{a,T+1}] \leq \frac{(1 + \varepsilon) \log T}{K(x_a, y_a)} + o(\log T).$$

(e) Let $x_a \downarrow \mu_a$, $y_a \uparrow \mu_1$, and $\varepsilon \downarrow 0$. Deduce the asymptotic pull-count bound

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[N_{a,T+1}]}{\log T} \leq \frac{1}{K(\mu_a, \mu_1)}.$$

Then derive the corresponding logarithmic regret upper bound.

Question 3: Thompson Sampling versus UCB

[4 Points]

- (a) Compare the KL envelope $U_{a,t}$ with the deterministic index used by KL-UCB. Explain the sense in which posterior quantiles behave like random confidence bounds.
- (b) Give a precise comparison between temporally randomized posterior sampling and deterministic optimism. Your answer should address exploration, tie-breaking, and behavior when two arms have nearly equal means.
- (c) State one setting where Thompson Sampling is preferable to UCB and one setting where UCB-style optimism may be preferable. Justify both claims using the regret proof ingredients above.

2 Action Elimination for Multi-Armed Bandits

Consider a stochastic K -armed bandit with rewards in $[0, 1]$. Arm a has mean μ_a , arm a^* is uniquely optimal, and $\Delta_a = \mu_{a^*} - \mu_a$ for $a \neq a^*$. Let A_t be the arm pulled at round t and define the pseudo-regret

$$R_T = \sum_{t=1}^T (\mu_{a^*} - \mu_{A_t}) = \sum_{a \neq a^*} \Delta_a T_a(T).$$

Since rewards lie in $[0, 1]$, the regret on any failure event is at most T . The action-elimination algorithm maintains an active set $\mathcal{A}_\ell$ over epochs $\ell = 1, 2, \dots$. Let

$$\varepsilon_\ell = 2^{-\ell}, \quad m_\ell = \left\lceil \frac{2}{\varepsilon_\ell^2} \log \frac{4KT\ell^2}{\delta} \right\rceil.$$

In epoch ℓ , every active arm is sampled until it has m_ℓ total samples; m_ℓ is not an additional number of samples beyond previous epochs. Let $\hat{\mu}_{a,\ell}$ be its empirical mean after these samples. Eliminate arm a if

$$\hat{\mu}_{a,\ell} + \varepsilon_\ell < \max_{b \in \mathcal{A}_\ell} \hat{\mu}_{b,\ell} - \varepsilon_\ell.$$

Question 1: Elimination Event and Gap-Dependent Regret

[10 Points]

- (a) Prove that with probability at least $1 - \delta$,

$$|\hat{\mu}_{a,\ell} - \mu_a| \leq \varepsilon_\ell \quad \text{for every arm } a \text{ sampled to } m_\ell \text{ and every completed epoch } \ell$$

before the horizon T . Use a union bound over arms and epochs; epochs that are not completed before time T do not need empirical means.

- (b) On this event, prove that the optimal arm is never eliminated.
 (c) Prove that a suboptimal arm a is eliminated by the first epoch ℓ satisfying $4\varepsilon_\ell < \Delta_a$.
 (d) Derive an upper bound on the number of pulls of arm a before elimination of order

$$O\left(\frac{1}{\Delta_a^2} \log \frac{KT \log(1/\Delta_a)}{\delta}\right).$$

- (e) Combine the previous parts to obtain a high-probability gap-dependent regret bound of the form

$$R_T = O\left(\sum_{a \neq a^*} \frac{1}{\Delta_a} \log \frac{KT \log(1/\Delta_a)}{\delta}\right)$$

on the good event. Convert it into an expected-regret bound by controlling the failure event.

Question 2: Gap-Free Bound and Comparison

[8 Points]

- (a) For a threshold $\Delta_0 > 0$, split the suboptimal arms into $\{a : \Delta_a \leq \Delta_0\}$ and $\{a : \Delta_a > \Delta_0\}$. For the small-gap group, use the crude bound that their total regret is at most $T\Delta_0$. For the large-gap group, plug in the gap-dependent bound from the previous question and use $1/\Delta_a \leq 1/\Delta_0$.
 (b) Optimize the threshold Δ_0 to derive a gap-free regret bound of order

$$\tilde{O}(\sqrt{KT}),$$

where $\tilde{O}$ hides logarithmic factors in the parameters appearing in the previous gap-dependent bound.

- (c) Compare the elimination proof with the UCB proof technique. Which parts are concentration arguments, and which parts use the epoch structure?
- (d) Give an instance where action elimination quickly discards many arms but UCB continues to sample them occasionally. Explain the trade-off between permanent elimination and persistent optimism.

3 RND vs Bootstrapped DQN on DeepSea

DeepSea is an $N \times N$ sparse-reward exploration benchmark. The agent starts at the top-left cell and is forced to move down exactly one row per time step, so each episode has horizon N . In each column, one of the two row actions moves the agent one column to the right; the identity of the right-moving row action is randomized per column. For the theoretical probability calculation in this homework, reaching the treasure is modeled as requiring N independent correct guesses of these right-moving row actions. The only large reward, equal to $+1$, is obtained at the bottom-right cell. Moving right incurs a small total cost, so the always-right policy has reference return 0.99.

The reachable set is

$$\mathcal{R}_N = \{(r, c) : 0 \leq r \leq N - 1, 0 \leq c \leq r\}.$$

The noisy-TV variant appends d fresh i.i.d. noise features to the observation whenever the agent is in a designated column, by default the left wall $c = 0$. These noise features are action-independent and reward-irrelevant.

For a training run of total length T environment steps, let

$$C_{r,c}^{(t)}$$

denote the cumulative visit count of cell (r, c) after t environment steps, where $0 \leq t \leq T$. In the submitted practical run, the total number of steps is fixed at

$$T = 50,000.$$

Thus $C_{r,c}^{(50,000)}$ is the final cumulative visit count used for the final heatmap and summary metrics. When the time index is clear, write $C_{r,c}$ for $C_{r,c}^{(t)}$. The exploration coverage at time t is

$$\text{Cov}(C^{(t)}) = \frac{1}{|\mathcal{R}_N|} \sum_{(r,c) \in \mathcal{R}_N} \mathbf{1}\{C_{r,c}^{(t)} > 0\}.$$

The visitation entropy in bits is

$$H(C^{(t)}) = - \sum_{(r,c) \in \mathcal{R}_N} p_{r,c}^{(t)} \log_2 p_{r,c}^{(t)}, \quad p_{r,c}^{(t)} = \frac{C_{r,c}^{(t)}}{\sum_{(i,j) \in \mathcal{R}_N} C_{i,j}^{(t)}}.$$

Cells with zero count contribute $0 \log_2 0 := 0$ to the entropy sum. If the denominator is zero, define $H(C^{(t)}) = 0$.

Random Network Distillation uses a frozen random target network $\bar{f} : \mathcal{S} \rightarrow \mathbb{R}^m$ and a trainable predictor $f_\psi : \mathcal{S} \rightarrow \mathbb{R}^m$. Its raw novelty error is

$$e_\psi(s) = \frac{1}{2} \|f_\psi(s) - \bar{f}(s)\|_2^2.$$

RND uses the shaped reward

$$\tilde{r}_t = r_t + \beta b_t,$$

where b_t is a normalized intrinsic bonus derived from the RND prediction error.

Bootstrapped DQN uses K value heads. Head k represents

$$Q_k(s, a) = Q_{\theta_k}(s, a) + \lambda P_k(s, a),$$

where Q_{θ_k} is trainable, P_k is a frozen randomized prior, and $\lambda > 0$ is the prior scale. One head is sampled at the beginning of each episode and followed greedily for the entire episode.

Implementation 1: RND vs Bootstrapped DQN on DeepSea**[16 Points]**

Complete the practical assignment described in `README.md`. At minimum, your submission must include the following components.

- (a) Implement `coverage_entropy` in `hw/student_metric.py`. The function must ignore unreachable cells, return coverage over $\mathcal{R}_N$, and return visitation entropy in bits, with entropy 0.0 when there are no reachable visits.
- (b) Implement the RND learning core in `hw/student_rnd.py`: network construction, intrinsic bonus normalization, predictor loss, Double-DQN TD target on shaped rewards, and the RND Q-loss.
- (c) Implement Bootstrapped DQN with randomized priors in `hw/student_bootstrap.py`: ensemble construction, one-head-per-episode sampling, greedy action selection for a sampled head, ensemble-mean evaluation, per-head Double-DQN targets, and masked Huber losses.
- (d) Run the ordinary and noisy-TV DeepSea experiments.
- (e) Submit the completed notebook and the generated `submission.zip` containing the implementation fingerprints, run records, and generated plots.

Question 1: Empirical Analysis Questions**[8 Points]**

Base your answers on the plots and metrics generated by your submitted runs; there is no predetermined numerical ranking to assume.

- (a) On ordinary DeepSea, compare RND, Bootstrapped DQN, and the plain Double-DQN baseline. Which method finds the treasure sooner? Which method is more reliable across seeds? Support your answer using the learning curves, first-visit-to-goal bars, and IQM panel.
- (b) On noisy-TV DeepSea, describe how RND's behavior changes relative to the ordinary board. Refer to its intrinsic-bonus curve and visitation heatmap. State whether the RND bonus decays in the noisy column.
- (c) Using coverage, visitation entropy, and deepest-column-reached, describe each agent's exploration style. Can coverage look healthy while the goal is never reached?
- (d) For a task with reward-irrelevant sensory noise, such as a flickering screen, choose either RND or Bootstrapped DQN with randomized priors. Justify the choice.

4 Variance-Aware UCB

Consider a stochastic multi-armed bandit with $K \geq 2$ arms. Arm k has an unknown reward distribution ν_k supported on $[0, b]$, where $b > 0$ is known. When arm k is pulled for the s -th time, the learner observes

$$X_{k,s} \sim \nu_k.$$

For each fixed k , the sequence $(X_{k,s})_{s \geq 1}$ is i.i.d., and reward sequences of different arms are independent.

Define

$$\mu_k = \mathbb{E}[X_{k,1}], \quad \sigma_k^2 = \text{Var}(X_{k,1}), \quad \mu^* = \max_{1 \leq k \leq K} \mu_k,$$

and fix one $k^* \in \arg \max_{1 \leq k \leq K} \mu_k$. If several arms are optimal, arms with $\Delta_k = 0$ are treated as optimal and are excluded from suboptimal-arm pull-count bounds. For each arm k , define

$$\Delta_k = \mu^* - \mu_k.$$

At time $t = 1, 2, \dots, n$, the learner chooses an arm $I_t \in \{1, \dots, K\}$. Let

$$T_k(t) = \sum_{u=1}^t \mathbf{1}\{I_u = k\}$$

be the number of pulls of arm k up to time t .

The cumulative pseudo-regret is

$$R_n = \sum_{t=1}^n (\mu^* - \mu_{I_t}),$$

and the realized regret is

$$\hat{R}_n = \sum_{t=1}^n X_{k^*,t} - \sum_{t=1}^n X_{I_t, T_{I_t}(t)}.$$

Here the first sum denotes the reward sequence that would have been observed by pulling the fixed optimal arm k^* for n rounds. The term $X_{I_t, T_{I_t}(t)}$ is the reward observed at time t , indexed by the selected arm's within-arm pull count. The counterfactual and observed sums use the same arm-specific i.i.d. reward-sequence notation above, and you may apply Wald's identity separately to each arm when needed.

Question 1: Regret Decomposition

[4 Points]

(a) Prove that

$$R_n = \sum_{k=1}^K \Delta_k T_k(n).$$

(b) Prove that

$$\mathbb{E}[R_n] = \sum_{k=1}^K \Delta_k \mathbb{E}[T_k(n)].$$

(c) Prove that

$$\mathbb{E}[\hat{R}_n] = \mathbb{E}[R_n].$$

(d) Explain why controlling $\mathbb{E}[T_k(n)]$ for every suboptimal arm is enough to control expected regret.

(e) If there is only one suboptimal arm k , explain why $\Delta_k \mathbb{E}[T_k(n)]$ captures both the statistical difficulty of learning and the cost of mistakes.

4.1 Empirical Bernstein Confidence Bounds

For this subsection, fix a sample size $s \geq 1$ and a confidence parameter $x > 0$. Let $X_1, \dots, X_s$ be i.i.d. random variables supported on $[0, b]$, with mean μ . Define

$$\bar{X}_s = \frac{1}{s} \sum_{i=1}^s X_i, \quad V_s = \frac{1}{s} \sum_{i=1}^s (X_i - \bar{X}_s)^2.$$

The empirical variance uses the displayed $1/s$ normalization, not the unbiased $1/(s-1)$ normalization.

Question 2: Empirical Bernstein Inequality

[5 Points]

- (a) State Bernstein's inequality for bounded random variables in a form that controls $|\bar{X}_s - \mu|$.
 (b) Prove that, with probability at least $1 - 3e^{-x}$,

$$|\bar{X}_s - \mu| \leq \sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s}.$$

You may use the following empirical-variance comparison without proof: with probability at least $1 - e^{-x}$,

$$\sqrt{\frac{2\sigma^2 x}{s}} + \frac{bx}{3s} \leq \sqrt{\frac{2V_s x}{s}} + \frac{3bx}{s}.$$

- (c) Explain the roles of the variance-sensitive term and the linear range term:

$$\sqrt{\frac{2V_s x}{s}}, \quad \frac{3bx}{s}.$$

- (d) In a bandit algorithm, $s = T_k(t)$ is random. Explain why the confidence bound must hold uniformly over many possible sample counts.

4.2 The Variance-Aware UCB Index

For each arm k and each sample count $s \geq 1$, after s pulls define

$$\bar{X}_{k,s} = \frac{1}{s} \sum_{i=1}^s X_{k,i}, \quad V_{k,s} = \frac{1}{s} \sum_{i=1}^s (X_{k,i} - \bar{X}_{k,s})^2.$$

Let $E_{s,t} \geq 0$ be an exploration function, and let $c \geq 0$. Define

$$B_{k,s,t} = \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s}E_{s,t}}{s}} + c \frac{3bE_{s,t}}{s}, \quad B_{k,0,t} = +\infty.$$

The policy selects

$$I_t \in \arg \max_{1 \leq k \leq K} B_{k,T_k(t-1),t}.$$

The convention $B_{k,0,t} = +\infty$ is used only for initialization; confidence statements involving $B_{k,s,t}$ use $s \geq 1$.

Question 3: Variance-Aware UCB Index**[4 Points]**

- (a) Explain why $B_{k,0,t} = +\infty$ guarantees that every arm is sampled at least once.
- (b) Interpret the empirical mean term, the empirical-variance term, and the linear range term in $B_{k,s,t}$.
- (c) Let

$$\mathcal{E} = \left\{ \mu_k \leq \bar{X}_{k,s} + \sqrt{\frac{2V_{k,s}E_{s,t}}{s}} + c \frac{3bE_{s,t}}{s} \text{ for every } k, s \geq 1, t \right\}.$$

Show that, on $\mathcal{E}$, $B_{k,s,t}$ is an upper confidence bound on μ_k for every sampled arm and sample count.

- (d) Compare this index to a range-only UCB index. When two suboptimal arms have the same gap but very different variances, which arm should be sampled fewer times because its index falls below competitive arms sooner?
- (e) Explain the role of $E_{s,t}$. What goes wrong if it grows too slowly or too quickly?

4.3 Counting Pulls of a Suboptimal Arm

Fix a suboptimal arm k with $\Delta_k > 0$. Let τ be a deterministic threshold and let $u \geq 1$ be a deterministic integer. The existential quantifiers over s below account for the fact that the realized sample counts $T_k(t-1)$ and $T_{k^*}(t-1)$ are random.

Question 4: Bounding Pulls of a Suboptimal Arm**[5 Points]**

- (a) Show that after the first K rounds, every arm has been pulled at least once.
- (b) Prove that if arm k is selected at time t and $T_k(t-1) \geq u$, then

$$B_{k,T_k(t-1),t} \geq B_{k^*,T_{k^*}(t-1),t}.$$

- (c) Show that the previous inequality implies that at least one of the following events occurs:

$$B_{k,T_k(t-1),t} > \tau, \quad B_{k^*,T_{k^*}(t-1),t} \leq \tau.$$

- (d) Deduce the following bound; the lower limit $u + K - 1$ only accounts for initialization and off-by-one conventions and may be changed by constants without affecting the regret order:

$$\begin{aligned} T_k(n) \leq u + \sum_{t=u+K-1}^n \mathbf{1} \{ \exists s \in \{u, \dots, t-1\} : B_{k,s,t} > \tau \} \\ + \sum_{t=u+K-1}^n \mathbf{1} \{ \exists s \in \{1, \dots, t-1\} : B_{k^*,s,t} \leq \tau \}. \end{aligned}$$

- (e) Take expectations and use a union bound to obtain an upper bound on $\mathbb{E}[T_k(n)]$.
- (f) Explain the two error events: the suboptimal arm looks too good, and the optimal arm looks too bad. Why is $\tau = \mu^*$ the natural threshold?

4.4 Expected Regret with a General Exploration Function

Assume that $E_{s,t} = E_t$, where $(E_t)_{t \geq 1}$ is nondecreasing, so $E_t \leq E_n$ for every $t \leq n$. Write $c \vee 1 = \max\{c, 1\}$. For each suboptimal arm k , define

$$u_k = \left\lceil 8(c \vee 1) \left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{2b}{\Delta_k} \right) E_n \right\rceil.$$

Question 5: Expected Regret for General Exploration

[6 Points]

- (a) Show that if $s \geq u_k$ and $t \leq n$, then the threshold choice, using $E_t \leq E_n$, makes both scale terms

$$\sqrt{\frac{\sigma_k^2 E_t}{s}} \quad \text{and} \quad \frac{b E_t}{s}$$

at most constant multiples of Δ_k . State the constants you obtain.

- (b) Prove that for all $s \geq u_k$ and $t \leq n$,

$$\mathbb{P}(B_{k,s,t} > \mu^*) \leq 2 \exp \left(-\frac{s \Delta_k^2}{8\sigma_k^2 + \frac{4}{3}b\Delta_k} \right).$$

- (c) Use the previous counting argument with $\tau = \mu^*$ to show the explicit double-sum bound

$$\begin{aligned} \mathbb{E}[T_k(n)] &\leq u_k + \sum_{t=u_k+K-1}^n \sum_{s=u_k}^{t-1} \mathbb{P}(B_{k,s,t} > \mu^*) \\ &\quad + \sum_{t=u_k+K-1}^n \sum_{s=1}^{t-1} \mathbb{P}(B_{k^*,s,t} \leq \mu^*). \end{aligned}$$

These two double sums are the confidence-failure terms.

- (d) Derive the leading-order contribution

$$\mathbb{E}[T_k(n)] = O \left(\left(\frac{\sigma_k^2}{\Delta_k^2} + \frac{b}{\Delta_k} \right) E_n \right),$$

up to the lower-order confidence-failure terms represented by the two double sums in part (c).

- (e) Multiply by Δ_k and obtain the general regret form

$$\mathbb{E}[R_n] \lesssim \sum_{k: \Delta_k > 0} \left(\frac{\sigma_k^2}{\Delta_k} + b \right) E_n + \text{lower-order terms}.$$

- (f) Compare this leading term with the range-only UCB scaling

$$\sum_{k: \Delta_k > 0} \frac{b^2}{\Delta_k} \log n.$$

4.5 Logarithmic Exploration

Take

$$E_t = \zeta \log t, \quad \zeta > 1, \quad c = 1.$$

Question 6: Logarithmic Regret**[4 Points]**

- (a) Explain why E_t should be proportional to $\log t$ in order to obtain logarithmic expected regret.
- (b) Prove that there exists a constant $C_\zeta > 0$, depending only on ζ and universal numerical constants, and not on n , K , gaps, or variances, such that

$$\mathbb{E}[R_n] \leq C_\zeta \sum_{k: \Delta_k > 0} \left(\frac{\sigma_k^2}{\Delta_k} + 2b \right) \log n.$$

- (c) Explain why $\zeta > 1$ is needed to make the time-uniform confidence-failure sums summable.
- (d) Discuss the regimes $\sigma_k^2 \approx 0$ and $\sigma_k^2 \approx b^2$. When is variance-aware UCB significantly better than range-only UCB?

Question 7: Interpretation**[2 Point]**

- (a) Give an example where variance-aware exploration should help a lot. Specify K , feasible reward distributions or feasible mean–variance values on $[0, b]$, and the corresponding μ_k , σ_k^2 , and Δ_k for each arm.
- (b) Give an example, again using feasible bounded reward distributions or feasible mean–variance values, where variance-aware exploration should not help much compared to ordinary UCB.
- (c) Explain the statement: “A good exploration strategy should not treat all uncertainty as equal.”
- (d) Distinguish between proving small expected regret and proving small high-probability regret. Why might a risk-averse decision maker care about the difference?

References

[1] Based on INF8250AE - Introduction to Reinforcement Learning. Fall 2025.