Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 12

Solution by: Behnia Soleymani

Q1

Consider a stationary multi-armed bandit problem where reward distributions have bounded variance. As the number of time steps grows very large, how does the exploration behavior induced by the Upper Confidence Bound (UCB) algorithm typically differ from that of an ϵ-Greedy algorithm using a constant $\epsilon > 0$?

A) UCB guarantees faster convergence to the optimal arm compared to ϵ-Greedy because its exploration is directed by uncertainty.

B) UCB significantly reduces the frequency of selecting suboptimal arms over time, whereas ϵ-Greedy continues to select any suboptimal arm with a non-vanishing probability.

C) ϵ-Greedy ceases exploration once value estimates stabilize, while the logarithmic term in UCB ensures exploration continues indefinitely at a constant rate.

D) UCB primarily explores arms with low estimated values to confirm they are suboptimal, while ϵ-Greedy explores purely randomly.

Correct Answer: B

Explanation:

- UCB selects arms based on $Q_t(a) + c\sqrt{\frac{\ln(t)}{N_t(a)}}$. As the optimal arm a^* is pulled, $N_t(a^*)$ grows (roughly $\propto t$), causing its uncertainty bonus to diminish faster than $\ln(t)$ grows. Suboptimal arms are pulled less often ($N_t(a')$ grows slower, ideally $\propto \ln t$), so their bonus term diminishes more slowly relative to their lower $Q_t(a')$, ensuring they are occasionally explored. However, the overall selection probability for any specific suboptimal arm tends towards zero as $t \to \infty$.
- In contrast, ϵ-Greedy with constant $\epsilon > 0$ always explores with total probability ϵ, choosing uniformly among all k arms during exploration. Thus, any specific suboptimal arm retains a minimum selection probability of ϵ/k, which is non-vanishing.
- (A) While often faster in practice, *guaranteeing* faster convergence is a strong claim depending on precise definitions and problem specifics. The key difference is the asymptotic exploration behavior.
- (C) is incorrect because constant-ϵ ϵ-Greedy never stops exploring. UCB's exploration rate diminishes for suboptimal arms.
- (D) is incorrect; UCB explores based on high uncertainty (low $N_t(a)$) or high $Q_t(a)$. ϵ-Greedy explores randomly among all arms.

Q2

When employing an incremental update rule $Q_{n+1} = Q_n + \text{StepSize} \times (R_n - Q_n)$ to estimate arm values in a non-stationary multi-armed bandit problem (where true mean rewards $q^*(a)$ can change over time), what is the primary statistical trade-off associated with using a constant step size $\alpha \in (0, 1]$ versus a decaying step size like $1/n$?

A) Trading potential bias in the estimate for reduced computational cost.
B) Trading the guarantee of converging to the true mean reward for the ability to handle arbitrary reward distributions.
C) Trading faster adaptation to changes in the underlying reward distributions for potentially higher variance in the resulting value estimates.
D) Trading simplicity of implementation for robustness against outliers in the observed rewards.

Correct Answer: C

Explanation:

- A constant step size α effectively creates an exponential moving average: $Q_{n+1} = (1 - \alpha)Q_n + \alpha R_n$. This gives greater weight to recent rewards, enabling the estimate Q_n to track changes in the true mean $q^*(a)$ more quickly (faster adaptation).
- However, because recent (potentially noisy) rewards always have a significant influence (weight α), the estimate Q_n will continuously fluctuate around the current true mean, exhibiting higher variance compared to an estimate using a $1/n$ step size (which would converge to a stable value with low variance in a stationary setting).
- (A) Computational costs are similar for both incremental updates. Both methods can be biased, especially in non-stationary settings.
- (B) Neither guarantees convergence to the true mean in non-stationary settings. The step size choice isn't primarily about the reward distribution type.
- (D) The adaptation vs. variance trade-off is the most fundamental statistical difference in this context. Robustness to outliers depends more on the magnitude of α or specific robust estimation techniques.

Q3

Thompson Sampling utilizes a Bayesian approach for action selection in multi-armed bandits. How does this algorithm inherently incorporate the concept of uncertainty about an arm's value to drive exploration?

A) By adding an explicit "uncertainty bonus" term, proportional to the variance of the belief distribution, to the mean estimate before selecting the arm.
B) By initializing arms with highly optimistic prior beliefs, which are gradually corrected downwards based on observed rewards.
C) By maintaining a full probability distribution (belief) over each arm's value parameter; arms with wider distributions (higher uncertainty) have a greater chance of producing a high sampled value, leading to their selection.
D) By reserving a fixed fraction of time steps for purely random exploration, ensuring that all arms, including uncertain ones, are eventually sampled.

Correct Answer: C

Explanation:

- Thompson Sampling operates by maintaining a posterior probability distribution (belief) for the parameter of interest (e.g., mean reward θ_a) for each arm a. At each step, it draws a sample $\tilde{\theta}_a$ from each arm's current belief distribution $P(\theta_a|\text{History})$. It then selects the arm $A_t = \text{argmax}_a \tilde{\theta}_a$.
- Uncertainty about an arm's value is represented by the spread (variance) of its belief distribution. An arm with high uncertainty (a wide distribution) has a significant probability of yielding a high sampled value $\tilde{\theta}_a$, even if its posterior mean is not the highest. If such a sample happens to be the maximum across all arms, the uncertain arm is selected (explored). This mechanism naturally balances exploration (sampling from wide distributions) and exploitation (sampling from distributions concentrated around high values).
- (A) describes the principle of UCB algorithms.
- (B) describes the Optimistic Initial Values strategy.
- (D) describes the exploration component of ϵ-Greedy algorithms.

Q4

The LinUCB algorithm for contextual bandits models the expected reward of arm a given context $x_{t,a}$ as a linear function: $E[r|x_{t,a}] = x_{t,a}^T \theta_a^*$. What is a critical vulnerability of LinUCB if this core linearity assumption is strongly violated in a practical application?

A) The algorithm may become computationally intractable due to the need to solve complex non-linear equations.
B) The uncertainty estimates derived from the linear model may become unreliable, potentially leading to insufficient exploration of truly optimal but non-linearly favored arms.
C) The algorithm will likely fail to run if the context features $x_{t,a}$ are not perfectly orthogonal.
D) The memory required to store the estimated parameter vectors $\hat{\theta}_a$ grows exponentially with the number of context features.

Correct Answer: B

Explanation:

- LinUCB's action selection relies on both the predicted linear reward $x_{t,a}^T \hat{\theta}_a$ and an uncertainty term derived from the linear regression model (related to the variance of the prediction). If the true relationship between context and expected reward is significantly non-linear, the linear model is misspecified.
- This misspecification impacts not only the prediction accuracy but also the reliability of the uncertainty quantification. The model might incorrectly estimate low uncertainty for context regions where the linear fit is poor, potentially underestimating the potential value of an arm whose true reward function behaves non-linearly. This can lead the algorithm to prematurely stop exploring arms that are optimal only in specific, non-linearly defined regions of the context space.
- (A) LinUCB assumes linearity specifically to maintain computational tractability via linear algebra, avoiding non-linear optimization.
- (C) LinUCB typically employs Ridge Regression, which handles correlated (non-orthogonal) features through regularization.
- (D) The memory for parameters $\hat{\theta}_a$ scales linearly with the number of features per arm, not exponentially.

Q5

Consider using the Optimistic Initial Values strategy (initializing $Q_0(a)$ to values known to be higher than any possible mean reward) combined with an incremental update rule using a constant step size α. How might the effectiveness of the initial optimism in driving exploration be affected, particularly in contrast to using a $1/n$ step size?

A) The constant step size α ensures the initial optimism persists longer, leading to more prolonged initial exploration compared to a $1/n$ step size.
B) The relatively rapid decay of the influence of the initial value under a constant α update might cause the exploration phase induced by optimism to terminate prematurely before sufficient information is gathered on all arms.
C) This combination is invalid; Optimistic Initial Values fundamentally requires a decaying step size ($1/n$) to guarantee that estimates eventually reflect true rewards.
D) In non-stationary environments, the constant step size α enhances the optimistic exploration by adapting quickly to potentially high initial rewards from suboptimal arms.

Correct Answer: B

Explanation:

- Optimistic Initial Values encourages exploration because untried arms retain their high initial $Q_0(a)$ value, making them appear better than arms that have been tried and whose estimates $Q_n(a)$ have decreased towards realistic values. The duration of this effect depends on how quickly the influence of $Q_0(a)$ diminishes.
- With a constant step size α, the update $Q_{n+1} = (1-\alpha)Q_n + \alpha R_n$ causes the influence of the initial value $Q_0(a)$ to decay exponentially as $(1-\alpha)^n$. If α is relatively large (e.g., $\alpha = 0.1$), this decay can be quite rapid. The initial high value might be effectively "washed out" by the first few observed rewards before the agent has been forced to try all other arms, thus potentially shortening the intended exploration phase.
- In contrast, using a sample average ($1/n$ effective step size for the n-th update for that arm) causes the influence of early data points (including the initial effective value) to diminish more slowly at the beginning, potentially allowing the optimism to drive exploration for longer.
- (A) is incorrect; constant α leads to faster decay of past influence, including the initial value.
- (C) is incorrect; the combination is algorithmically valid, but its behavior is different. No specific step size is mandated, but the interaction matters.
- (D) This statement is misleading. The primary effect is the faster decay of the initial optimism itself due to α.