Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 6

Solution by : Arshia Gharooni

Q1

Variance reduction techniques are crucial for stabilizing policy gradient updates. Which of the following statements correctly describe how variance reduction is achieved?

A) The causality trick removes rewards from earlier time steps in a trajectory when computing the gradient to prevent unnecessary dependencies.
B) Discounting future rewards with γ^t reduces variance by giving smaller weights to long-term rewards, which are inherently more uncertain.
C) The baseline function should ideally be a function of both state and action to ensure maximum variance reduction.
D) Bootstrapping methods, such as using a learned Q-function instead of Monte Carlo return estimation, introduce bias but significantly lower variance.
E) Variance can be eliminated entirely by normalizing rewards across episodes.

Correct Answers: A, B, D

Explanation: The causality trick avoids introducing unnecessary variance from past rewards. Bootstrapping methods trade off some bias for significantly lower variance. A baseline is typically a function of state only (not state-action) to maintain unbiasedness. Normalizing rewards reduces variance but does not eliminate it.

Q2

Actor-Critic methods combine value-based and policy-based learning. Which of the following statements correctly capture their theoretical properties?

A) The critic provides an estimate of the value function, which helps in reducing the variance of policy gradient updates.
B) Using an advantage function instead of Q-values results in lower variance but higher bias.
C) If the critic is perfectly learned, the actor updates will converge to the optimal policy under the policy gradient theorem.
D) The critic's function approximation must be compatible with the policy network to ensure convergence.
E) Actor-critic methods can always outperform value-based methods like Q-learning because they directly optimize the policy.

Correct Answers: A, C, D

Explanation: The critic helps lower variance by providing a more stable learning signal. Using an advantage function does not necessarily increase bias but does reduce variance. If the critic is perfect, policy updates converge optimally. Compatibility between function approximators ensures stable learning. Actor-critic methods are not necessarily superior to Q-learning in all cases.

Q3

When adding a baseline $b(s_t)$ to the policy gradient objective:

A) The baseline must be a constant (independent of the state) for the gradient to remain unbiased.
B) Subtracting $b(s_t)$ from the returns can significantly reduce variance of the gradient estimate.
C) Choosing $b(s_t) = V^\pi(s_t)$ often provides a good trade-off between complexity and variance reduction.
D) The baseline modifies the expected gradient, thereby introducing bias.
E) A baseline that is itself learned (e.g., a neural net for V^π) can be updated in tandem with the policy.

Correct Answers: B, C, E

- (A) is incorrect: you *can* use a state-dependent baseline as long as its expected value is zero in the gradient expression.
- (B), (C), and (E) are true: subtracting a learned value function (baseline) reduces variance, and you typically learn V^π alongside the policy.
- (D) is incorrect: a properly chosen baseline does not introduce bias in the *policy gradient*.

Q4

When using the reward-to-go version of the policy gradient, why does ignoring past rewards (before time t) help?

A) Because actions at time t do not cause the rewards received at earlier time steps.
B) Because including earlier rewards introduces unnecessary terms that have zero expected impact on the gradient.
C) Because we assume earlier rewards are typically larger and would overshadow future rewards.
D) Because focusing only on future rewards makes the policy "look ahead" more effectively.
E) Because ignoring the past lowers the variance of our gradient without changing its expectation.

Correct Answers: A, B, E

- **(A)** & **(B)** capture the notion of **causality**: an action cannot affect previously collected rewards.
- **(E)** is the true reason for the method's popularity—variance is reduced *without* bias.
- **(C)** is misleading because there is no guarantee that earlier rewards are typically larger.
- **(D)** states that the policy "looks ahead" more effectively, which is not precisely why reward-to-go is used; the main benefit is **variance reduction**.

Q5

The TRPO algorithm updates the policy by solving the following constrained optimization problem:

$$\max_{\theta} \mathbb{E}_{s\sim d^{\pi}} \left[\frac{\pi_{\theta'}(a|s)}{\pi_{\theta}(a|s)} A^{\pi}(s,a) \right] \tag{1}$$

subject to the KL-divergence constraint:

$$D_{KL}(\pi_{\theta'} || \pi_{\theta}) \leq \epsilon \tag{2}$$

What are the main reasons for using this KL constraint?

A) It prevents the new policy from diverging too much from the old one, stabilizing learning.
B) It ensures **monotonic improvement** in policy updates, avoiding catastrophic drops in performance.
C) It reduces variance in policy gradients by keeping updates within a small region.
D) It encourages more exploration by limiting the policy's certainty in actions.
E) It makes policy optimization easier by converting the problem into a convex optimization.

Correct Answers: A, B, C

- **(A)** is the **trust region** idea: keeping updates within a stable region.
- **(B)** ensures **monotonic improvement**, a key TRPO insight that prevents large drops in performance.
- **(C)** helps stabilize learning by constraining how much the policy can change in a single update.
- **(D)** is incorrect—TRPO itself doesn't encourage exploration (entropy regularization does).
- **(E)** is misleading—while TRPO makes optimization more stable, it doesn't make it fully convex.