6. (3 pts) What three things form the “deadly triad” – the three things that cannot be combined in the same learning situation without risking divergence? (circle three)

(a) eligibility traces
(b) bootstrapping
(c) sample backups
(d) \( \varepsilon \)-greedy action selection
(e) linear function approximation
(f) off-line updating
(g) off-policy learning
(h) exploration bonuses
The Deadly Triad
the three things that together result in instability

1. Function approximation

2. Bootstrapping

3. Off-policy training data (e.g., Q-learning, DP)

even if:

• prediction (fixed given policies)
• linear with binary features
• expected updates (as in asynchronous DP, iid)
6. **True or False:** What three things form the "deadly triad" – the three things that cannot be combined in the same learning situation without risking divergence? (circle three)
(a) eligibility traces
(b) bootstrapping
(c) sample backups
(d) \( \alpha \)-greedy action selection
(e) linear function approximation
(f) off-line updating
(g) off-policy learning
(h) exploration bonuses

7. **True or False:**
For any stationary MDP, assuming a step-size (\( \alpha \)) sequence satisfying the standard stochastic approximation criteria, and a fixed policy, convergence in the prediction problem is guaranteed for
- online, off-policy TD(1) with linear function approximation
- online, on-policy TD(0) with linear function approximation
- offline, off-policy TD(0) with linear function approximation
- dynamic programming with linear function approximation
- dynamic programming with nonlinear function approximation
- gradient-descent Monte Carlo with linear function approximation
- gradient-descent Monte Carlo with nonlinear function approximation

8. **True or False:**
TD(0) with linear function approximation converges to a local minimum in the MSE between the approximate value function and the true value function \( V^\pi \).

9. **In tile-coding function approximation,**
(a) (2 pts) the primary determinate of the shape of generalization is:
(b) (2 pts) the primary determinate of the acuity (fineness, resolution) of the asymptotic approximation is the product of:

10. **True or False:**
Sample-based planning methods can be focused more precisely than full backups on particular sequences or state–action pairs of interest.
The Deadly Triad
the three things that together result in instability

1. Function approximation
   • linear or more with proportional complexity
   • state aggregation ok; ok if “nearly Markov”

2. Bootstrapping
   • $\lambda=1$ ok, ok if $\lambda$ big enough (problem dependent)

3. Off-policy training
   • may be ok if “nearly on-policy”
   • if policies very different, variance may be too high anyway
Off-policy learning

• Learning about a policy different than the policy being used to generate actions

• Most often used to learn optimal behaviour from a given data set, or from more exploratory behaviour

• Key to ambitious theories of knowledge and perception as continual prediction about the outcomes of many options
Baird’s counter-example

• $P$ and $d$ are not linked
• $d$ is all states with equal probability
• $P$ is according to this Markov chain:

\[ V_k(s) = \theta(7) + \theta(1) \]
\[ V_k(s) = \theta(7) + \theta(2) \]
\[ V_k(s) = \theta(7) + \theta(3) \]
\[ V_k(s) = \theta(7) + \theta(4) \]
\[ V_k(s) = \theta(7) + \theta(5) \]

\[ r = 0 \]
on all transitions
TD can diverge: Baird’s counter-example

Parameter values, $\theta_k(i)$ (log scale, broken at ±1)

$\alpha = 0.01 \quad \gamma = 0.99 \quad \theta_0 = (1, 1, 1, 1, 10, 1)^T$ deterministic updates
TD(0) can diverge: A simple example

\[ \delta = r + \gamma \theta^\top \phi' - \theta^\top \phi \]
\[ = 0 + 2\theta - \theta \]
\[ = \theta \]

TD update: \[ \Delta \theta = \alpha \delta \phi \]
\[ = \alpha \theta \]
Diverges!

TD fixpoint: \[ \theta^* = 0 \]
Previous attempts to solve the off-policy problem

- Importance sampling
- With recognizers
- Least-squares methods, LSTD, LSPI, iLSTD
- Averagers
- Residual gradient methods
Desiderata: We want a TD algorithm that

- Bootstraps (genuine TD)
- Works with linear function approximation (stable, reliably convergent)
- Is simple, like linear TD — $O(n)$
- Learns fast, like linear TD
- Can learn off-policy (arbitrary $P$ and $d$)
- Learns from online causal trajectories (no repeat sampling from the same state)
A little more theory

\[ \Delta \theta \propto \delta \phi = (r + \gamma \theta^T \phi' - \theta^T \phi) \phi \]
\[ = \theta^T (\gamma \phi' - \phi) \phi + r \phi \]
\[ = \phi (\gamma \phi' - \phi)^T \theta + r \phi \]

\[ \mathbb{E} [\Delta \theta] \propto -\mathbb{E} \left[ \phi (\phi - \gamma \phi')^T \right] \theta + \mathbb{E} [r \phi] \]

\[ \mathbb{E} [\Delta \theta] \propto -A \theta + b \]

Therefore, at the TD fixpoint:

\[ A \theta^* = b \]
\[ \theta^* = A^{-1} b \]

LSTD computes this directly

\[ -\frac{1}{2} \nabla_\theta \text{MSPBE} = -A^T C^{-1} (A \theta - b) \]

Always pos. def.

\[ C = \mathbb{E} \left[ \phi \phi^T \right] \]

covariance matrix
TD(0) Solution and Stability

\[
\theta_{t+1} = \theta_t + \alpha \left( R_{t+1} \phi(S_t) - \phi(S_t) \left( \phi(S_t) - \gamma \phi(S_{t+1}) \right) \right)^\top \theta_t
\]

\[
= \theta_t + \alpha (b_t - A_t \theta_t)
\]

\[
= (I - \alpha A_t) \theta_t + \alpha b_t.
\]

\[
\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha (b - A \bar{\theta}_t)
\]

\[
\theta_* = A^{-1} b
\]
LSTD(0)

Ideal:

\[
A = \lim_{t \to \infty} \mathbb{E}_\pi [\phi_t (\phi_t - \gamma \phi_{t+1})^\top]
\]

\[
b = \lim_{t \to \infty} \mathbb{E}_\pi [R_{t+1} \phi_t]
\]

\[\theta_* = A^{-1} b\]

Algorithm:

\[
A_t = \sum_k \rho_k \phi_k (\phi_k - \gamma \phi_{k+1})^\top
\]

\[
b_t = \sum_k \rho_k R_k \phi_k
\]

\[
\lim_{t \to \infty} A_t = A
\]

\[
\lim_{t \to \infty} b_t = b
\]

\[
\theta_t = A_t^{-1} b_t
\]

\[
\lim_{t \to \infty} \theta_t = \theta_*
\]
LSTD($\lambda$)

**Ideal:**

\[
A = \lim_{t \to \infty} \mathbb{E}_\pi \left[ e_t (\phi_t - \gamma \phi_{t+1})^\top \right]
\]

\[
b = \lim_{t \to \infty} \mathbb{E}_\pi [R_{t+1} e_t]
\]

\[
\theta_* = A^{-1} b
\]

**Algorithm:**

\[
A_t = \sum_k \rho_k e_k \left( \phi_k - \gamma \phi_{k+1} \right)^\top
\]

\[
b_t = \sum_k \rho_k R_k e_k
\]

\[
\lim_{t \to \infty} A_t = A
\]

\[
\lim_{t \to \infty} b_t = b
\]

\[
\theta_t = A_t^{-1} b_t
\]

\[
\lim_{t \to \infty} \theta_t = \theta_*
\]