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) ε-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)

7. **True or False:** For any stationary MDP, assuming a step-size (α) sequence satisfying the standard stochastic approximation criteria, and a fixed policy, convergence in the prediction problem is guaranteed for

   T F (2 pts) online, off-policy TD(1) with linear function approximation

   T F (2 pts) online, on-policy TD(0) with linear function approximation

   T F (2 pts) offline, off-policy TD(0) with linear function approximation

   T F (2 pts) dynamic programming with linear function approximation

   T F (2 pts) dynamic programming with nonlinear function approximation

   T F (2 pts) gradient-descent Monte Carlo with linear function approximation

   T F (2 pts) gradient-descent Monte Carlo with nonlinear function approximation

8. **True or False:** (3 pts) 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^π.

The Deadly Triad

the three things that together result in instability

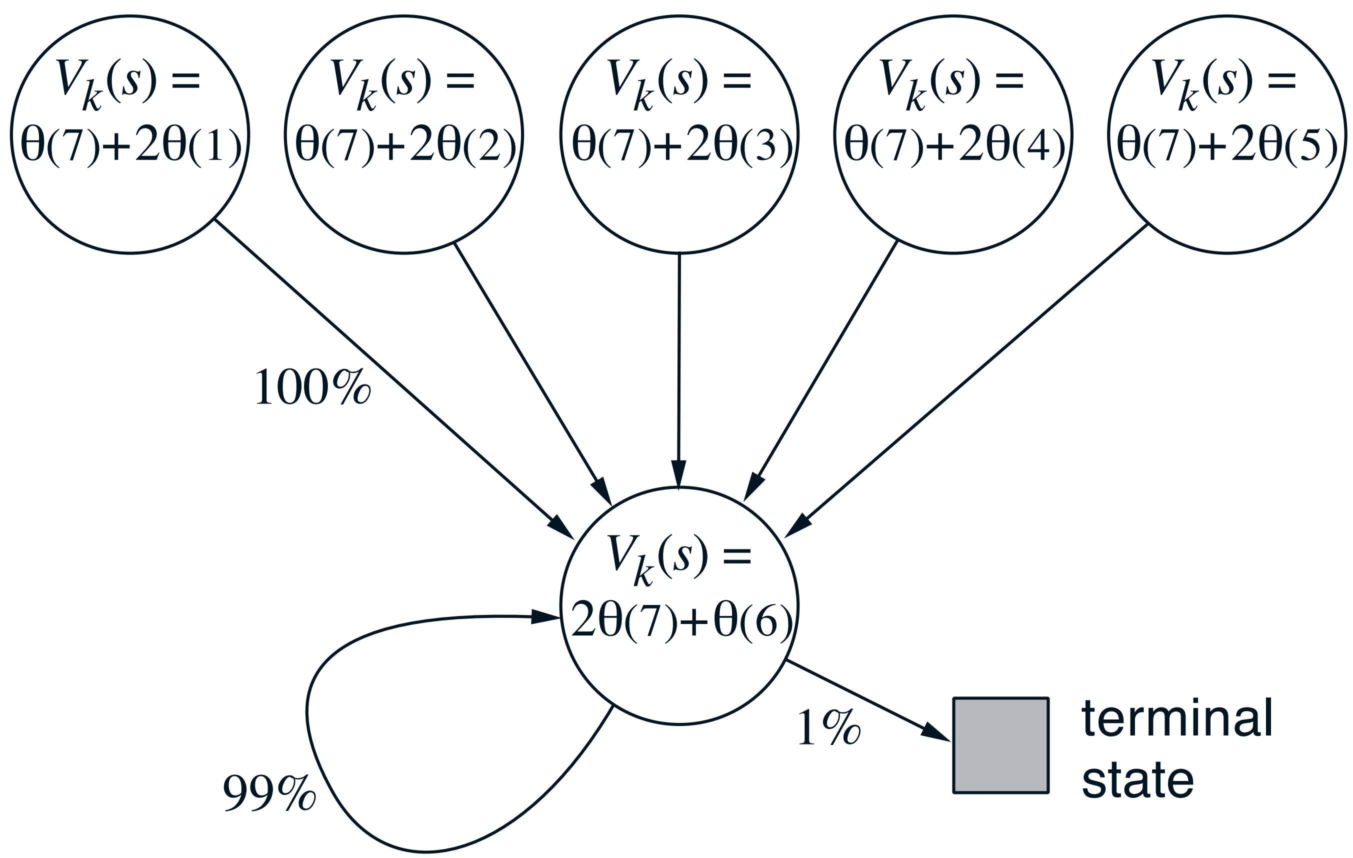
1. Function approximation
    - linear or more with proportional complexity
    - state aggregation ok; ok if “nearly Markov”
2. Bootstrapping
    - λ=1 ok, ok if λ big enough (problem dependent)
3. Off-policy training
    - may be ok if “nearly on-policy”
    - if policies very different, variance may 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

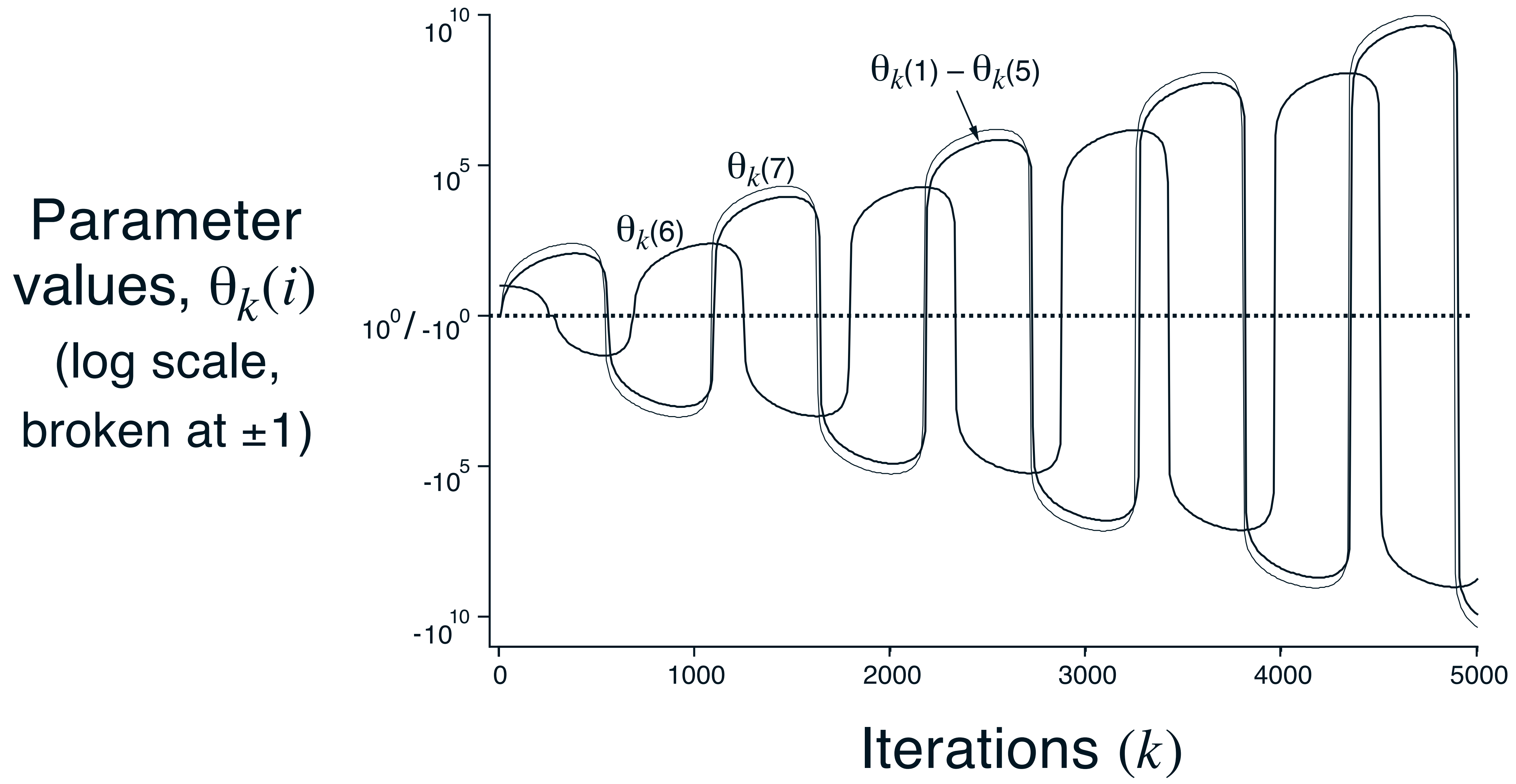
Baird's counter-example

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



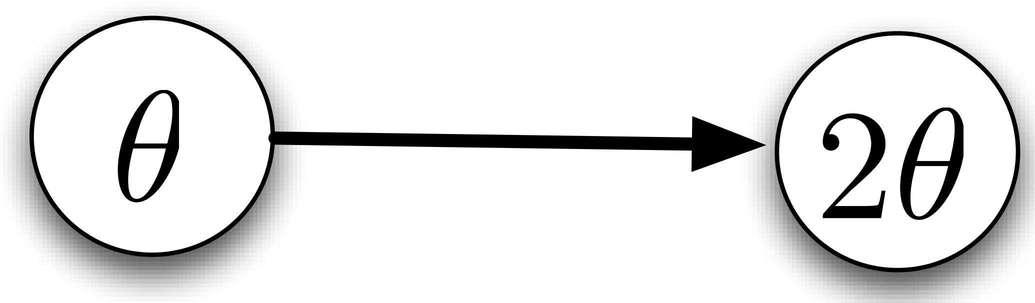
$r = 0$
on all transitions

TD can diverge: Baird's counter-example



$\alpha = 0.01$ $\gamma = 0.99$ $\theta_0 = (1, 1, 1, 1, 1, 10, 1)^\top$ deterministic updates

TD(0) can diverge: A simple example



$$
\begin{aligned}
\delta &= r + \gamma\theta^\top\phi' - \theta^\top\phi \\
&= 0 + 2\theta - \theta \\
&= \theta
\end{aligned}
$$

TD update: $$\begin{aligned} \Delta\theta &= \alpha\delta\phi \\ &= \alpha\theta \end{aligned}$$ 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

$$
\begin{aligned}
\Delta\theta \propto \delta\phi &= \left(r + \gamma\theta^\top\phi' - \theta^\top\phi\right)\phi \\
&= \theta^\top\left(\gamma\phi' - \phi\right)\phi + r\phi \\
&= \phi\left(\gamma\phi' - \phi\right)^\top\theta + r\phi \\
\mathbb{E}\left[\Delta\theta\right] &\propto -\mathbb{E}\left[\phi\left(\phi - \gamma\phi'\right)^\top\right]\theta + \mathbb{E}\left[r\phi\right]
\end{aligned}
$$

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

convergent if A is pos. def.

therefore, at the TD fixpoint:

$$
\begin{aligned}
A\theta^* &= b \\
\theta^* &= A^{-1}b
\end{aligned}
$$

LSTD computes this directly

$$
-\frac{1}{2}\nabla_\theta \mathrm{MSPBE} = -\underbrace{A^\top C^{-1}(A}_{\text{always pos. def.}}\theta - b)
$$

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

covariance matrix

TD(0) Solution and Stability

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha\Big(\underbrace{R_{t+1}\boldsymbol{\phi}(S_t)}_{\mathbf{b}_t\in\mathbb{R}^n} - \underbrace{\boldsymbol{\phi}(S_t)\left(\boldsymbol{\phi}(S_t) - \gamma\boldsymbol{\phi}(S_{t+1})\right)^\top}_{\mathbf{A}_t\in\mathbb{R}^{n\times n}}\boldsymbol{\theta}_t\Big)\\ &= \boldsymbol{\theta}_t + \alpha(\mathbf{b}_t - \mathbf{A}_t\boldsymbol{\theta}_t)\\ &= (\mathbf{I} - \alpha\mathbf{A}_t)\boldsymbol{\theta}_t + \alpha\mathbf{b}_t.\end{aligned}$$

$$\bar{\boldsymbol{\theta}}_{t+1} \doteq \bar{\boldsymbol{\theta}}_t + \alpha(\mathbf{b} - \mathbf{A}\bar{\boldsymbol{\theta}}_t)$$

$$\boldsymbol{\theta}_* = \mathbf{A}^{-1}\mathbf{b}$$

LSTD(0)

Ideal:

$$\mathbf{A} = \lim_{t\to\infty} \mathbb{E}_\pi\big[\boldsymbol{\phi}_t(\boldsymbol{\phi}_t - \gamma\boldsymbol{\phi}_{t+1})^\top\big]$$

$$\mathbf{b} = \lim_{t\to\infty} \mathbb{E}_\pi[R_{t+1}\boldsymbol{\phi}_t]$$

$$\boldsymbol{\theta}_* = \mathbf{A}^{-1}\mathbf{b}$$

Algorithm:

$$\mathbf{A}_t = \sum_k \rho_k \boldsymbol{\phi}_k\big(\boldsymbol{\phi}_k - \gamma\boldsymbol{\phi}_{k+1}\big)^\top$$

$$\mathbf{b}_t = \sum_k \rho_k R_k \boldsymbol{\phi}_k$$

$$\lim_{t\to\infty} \mathbf{A}_t = \mathbf{A}$$

$$\lim_{t\to\infty} \mathbf{b}_t = \mathbf{b}$$

$$\boldsymbol{\theta}_t = \mathbf{A}_t^{-1}\mathbf{b}_t$$

$$\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_*$$

LSTD(λ)

Ideal:

$$\mathbf{A} = \lim_{t\to\infty} \mathbb{E}_\pi\big[\mathbf{e}_t(\boldsymbol{\phi}_t - \gamma\boldsymbol{\phi}_{t+1})^\top\big]$$

$$\mathbf{b} = \lim_{t\to\infty} \mathbb{E}_\pi[R_{t+1}\mathbf{e}_t]$$

$$\boldsymbol{\theta}_* = \mathbf{A}^{-1}\mathbf{b}$$

Algorithm:

$$\mathbf{A}_t = \sum_k \rho_k \mathbf{e}_k \big(\boldsymbol{\phi}_k - \gamma\boldsymbol{\phi}_{k+1}\big)^\top$$

$$\mathbf{b}_t = \sum_k \rho_k R_k \mathbf{e}_k$$

$$\lim_{t\to\infty} \mathbf{A}_t = \mathbf{A}$$

$$\lim_{t\to\infty} \mathbf{b}_t = \mathbf{b}$$

$$\boldsymbol{\theta}_t = \mathbf{A}_t^{-1}\mathbf{b}_t$$

$$\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_*$$