Advanced Topics in Reinforcement Learning

Lecture 14: Off-Policy Function Approximation

Josiah Hanna
University of Wisconsin — Madison

Announcements

- Homework 3 due Thursday @ 9:29 AM
- Begin reading chapter 11 for next week.
- Midterm survey
  - At just under 50% right now.

Function Approximation Review

- Objective with function approximation.
  - $\overline{VE}(\mathbf{w}) = \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w}) \right]^2$
- Semi-gradient TD update.

Estimate of $v_\pi(S_t)$

  - $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \alpha(U_t - \hat{v}(s, \mathbf{w}_t)) \nabla \hat{v}(S_t, \mathbf{w}_t)$
- Linear Semi-Gradient Update
  - $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \alpha(U_t - \hat{v}(s, \mathbf{w}_t))\mathbf{x}(S_t)$

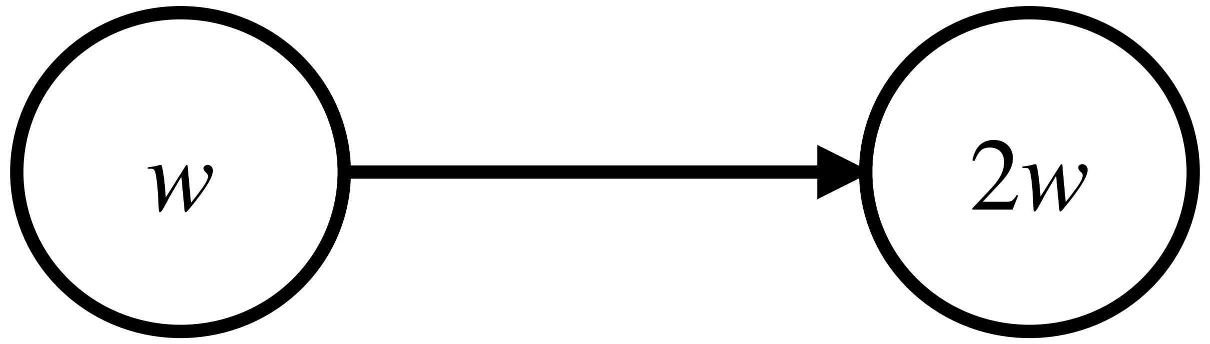
Off-Policy Prediction with Linear Function Approximation

- U_t must be an estimate of $v_\pi(S_t)$ but the return was generated by behavior policy, b.
- Recall from chapter 5, that we can correct for this by importance sampling.
  - N-step return: $G_{t:t+n} := R_{t+1} + \ldots + \gamma^{n-1} R_{t+n-1} + \gamma^n \hat{v}(S_{t+n}, \mathbf{w}_{t+n-1})$
  - For off-policy, replace $G_{t:t+n}$ with $G_{t:t+n} \cdot \rho_{t:t+n}$.
- Consider $U_t \leftarrow G_t$. Does this update minimize our $\overline{VE}$ objective?
  - No — does not adjust for state weighting.

Action-values with Linear Function Approximation

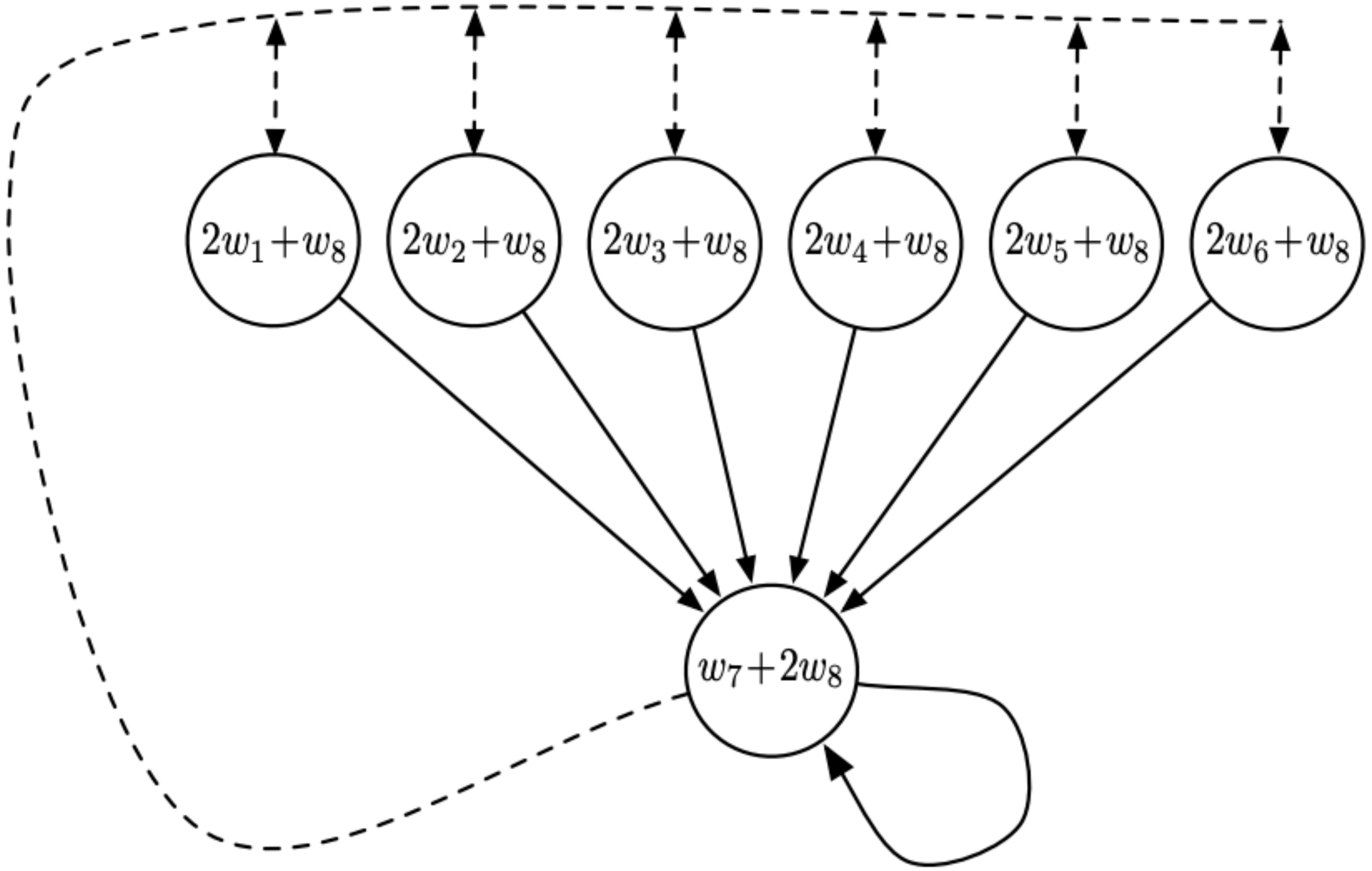
- For on- or off-policy learning, we can use Expected Sarsa:
- $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \alpha[R_{t+1} + \sum_a \pi(a \mid S_{t+1})\hat{q}(S_{t+1}, a, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t)] \nabla_{\mathbf{w}} \hat{q}(S_t, A_t, \mathbf{w}_t)$
- In off-policy case, why do we not require importance sampling?
  - We only sample A_t and it is the only action considered when estimating an action-value?

Divergence Example #1



- Initialize $w = 10$, $\gamma = 0.99$, $\alpha = 0.1$, and the transition gives zero reward.
- What happens after you've seen this transition once?
  - w increases to try and match bootstrapping target of $2\gamma w$.
- How can we fix divergence here?
  - First extend example to full MDP, then remove off-policy, bootstrapping, or function approximation.

Divergence Example #2: Baird's Counter-example



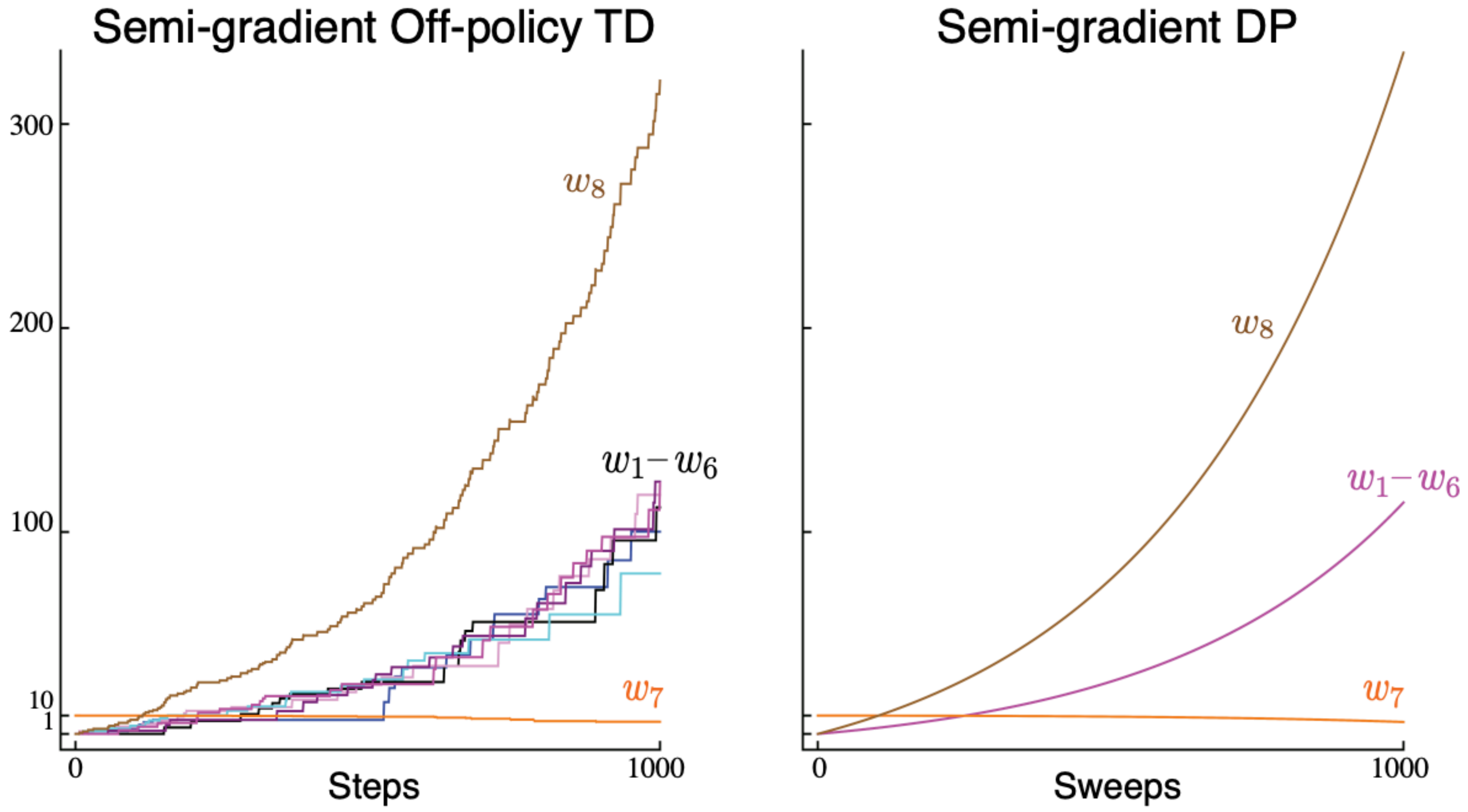
$\pi(\text{solid}|\cdot) = 1$

$b(\text{dashed}|\cdot) = 6/7$

$b(\text{solid}|\cdot) = 1/7$

$\gamma = 0.99$

Divergence Example #2: Baird's Counter-example



Off-Policy Divergence

- In general, we lack convergence or even stability results for the simplest and most practical off-policy, semi-gradient methods.
- Includes Q-learning which is one of the most widely used algorithms in RL.
  - Maybe OK if behavior and target policy are close?
  - State distributions will then be close.

The Deadly Triad

1. Function Approximation: changing the value estimate at one state affects the value estimate at other states.

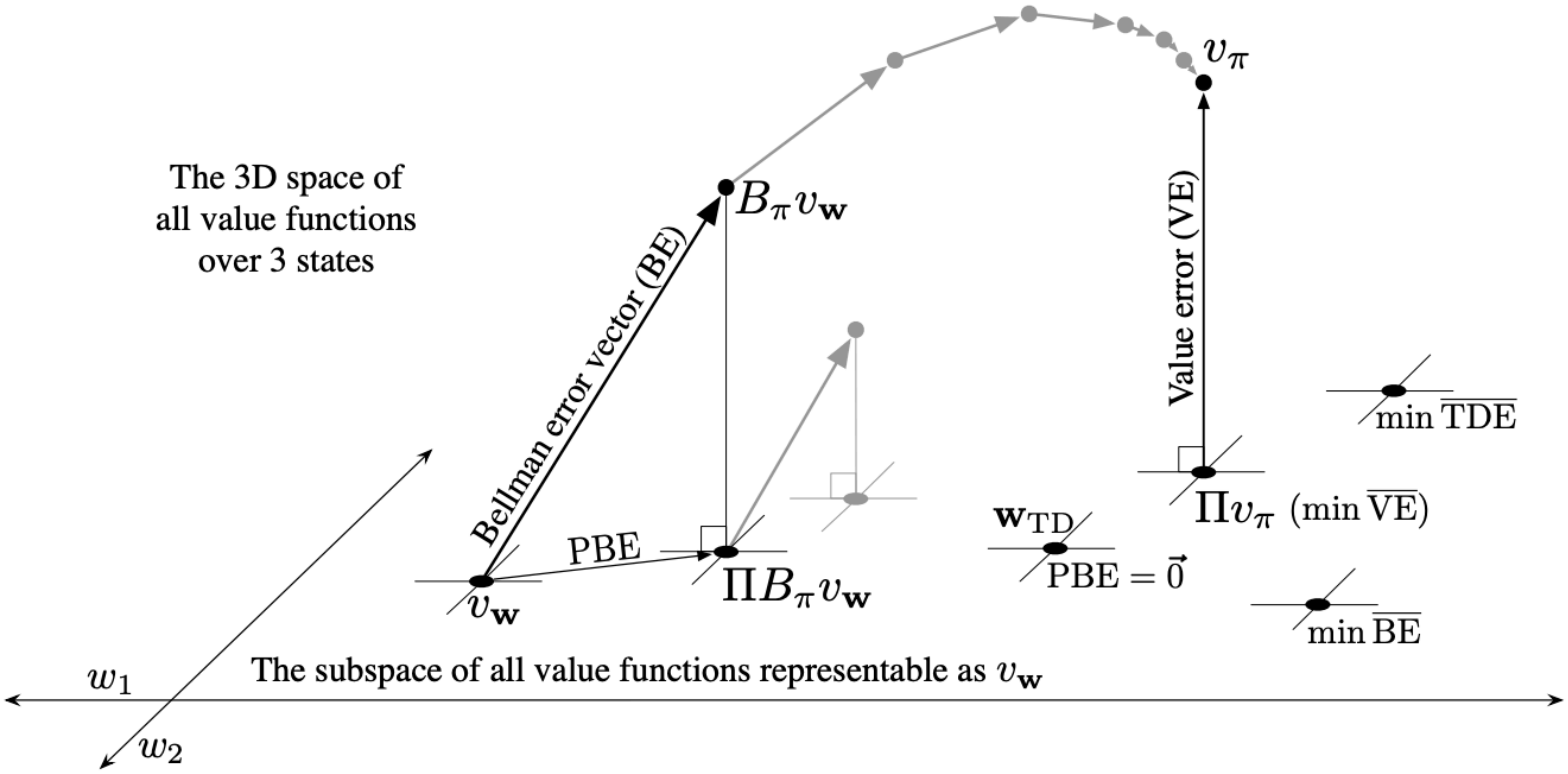
2. Bootstrapping: using existing estimated values as part of the learning target instead of only using actual returns.

3. Off-Policy Learning: using a distribution of transitions (s, a, s', r) other than that of the target policy.

Do we need the deadly triad?

- Why use function approximation?
  - Too many states to represent explicitly; need generalization.
- Why bootstrap?
  - Memory and computation requirements; learning in non-episodic tasks; faster learning.
- Why use off-policy learning?
  - Separate exploration and exploitation; general purpose learning agents must learn about multiple reward signals and target policies at the same time.

Geometric Interpretation of Value Functions



Possible Learning Objectives

- Minimum value error
  - $\overline{VE}(\mathbf{w}) = \sum_{s} \mu(s)(v_{\pi}(s) - \hat{v}(s, \mathbf{w}))^2 = ||v_{\mathbf{w}} - v_{\pi}||_{\mu}^2$
- Minimum TD-Error
  - $\overline{TDE}(\mathbf{w}) = \sum_{s} \mu(s)\mathbf{E}_{\pi}[\delta_t^2 \,|\, S_t = s, A_t \sim \pi]$
- Minimum Bellman error:
  - $\overline{BE}(\mathbf{w}) = ||\delta_{\mathbf{w}}||_{\mu}^2$
  - $\delta_w = \mathbf{E}_{\pi}[R_{t+1} + \gamma v_{\mathbf{w}}(S_{t+1}) - v_{\mathbf{w}}(S_t) \,|\, S_t = s, A_t \sim \pi]$

Andrew’s Presentation

- Slides

Summary

- Off-policy semi-gradient methods often lack stability and convergence results due to the deadly triad.
- Deadly Triad: off-policy, function approximation, and bootstrapping.
- Two paths forward:
  - Reconsider our prediction objective with function approximation.
  - Re-weight state updates.

Action Items

- Homework 3.
- Begin literature review.
- Begin reading Chapter 11.
- Midterm survey and evaluation.