Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation

Josiah Hanna[1] Peter Stone[1] Scott Niekum[2]

[1]Learning Agents Research Group, UT Austin

[2]Personal Autonomous Robotics Lab, UT Austin

May 10th, 2017

Motivation

Determine a **lower bound** on the **expected performance** of an autonomous control policy given data generated from a **different** policy.

Motivation

Determine a **lower bound** on the **expected performance** of an autonomous control policy given data generated from a **different** policy.

Motivation

Determine a **lower bound** on the **expected performance** of an autonomous control policy given data generated from a **different** policy.



Preliminaries

The agent samples actions from a policy, $A_t \sim \pi(\cdot|S_t)$.

The environment responds with $S_{t+1} \sim P(\cdot|S_t, A_t)$.

$$S_0 \rightarrow A_0 \rightarrow S_1 \rightarrow A_1 \rightarrow \cdots$$

The policy and environment determine a distribution over trajectories, $H : S_1, A_1, S_2, A_2, ..., S_L, A_L$

- $H \sim \pi$.
- $V(\pi) = \mathbf{E}\left[\sum_{t=1}^{L} r(S_t, A_t) \middle| H \sim \pi\right]$ is the expected return of π.

Confidence Intervals for Off-Policy Evaluation

Given:

- Trajectories generated by a *behavior* policy, π_b, $\{H, \pi_b\} \in \mathcal{D}$.
- An *evaluation* policy, π_e.
- $\delta \in [0, 1]$ is a confidence level.

Confidence Intervals for Off-Policy Evaluation

Given:

- Trajectories generated by a *behavior* policy, π_b, $\{H, \pi_b\} \in \mathcal{D}$.
- An *evaluation* policy, π_e.
- $\delta \in [0, 1]$ is a confidence level.

Determine a lower bound $\hat{V}_{\text{lb}}(\pi_e, \mathcal{D})$ such that $V(\pi_e) \geq \hat{V}_{\text{lb}}(\pi_e, \mathcal{D})$ with probability $1 - \delta$.

Existing Methods

- Exact confidence intervals Thomas et al. [2015a].
- Clip importance weights Bottou et al. [2013]
- Bootstrap importance-sampling Thomas et al. [2015b].

Existing Methods

- Exact confidence intervals Thomas et al. [2015a].
- Clip importance weights Bottou et al. [2013]
- Bootstrap importance-sampling Thomas et al. [2015b].

Our work

Data-Efficient Confidence Intervals

We draw on two ideas to reduce the number of trajectories required for tight confidence bounds.

- Replace exact confidence bounds with bootstrap confidence intervals.
- Use learned models of the environment's transition function to reduce variance.

Data-Efficient Confidence Intervals

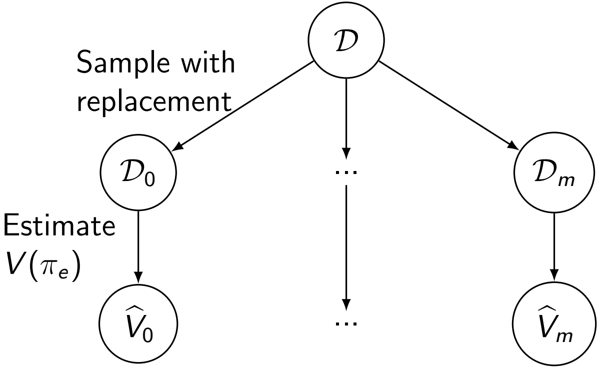
We draw on two ideas to reduce the number of trajectories required for tight confidence bounds.

- Replace exact confidence bounds with bootstrap confidence intervals.
- Use learned models of the environment's transition function to reduce variance.

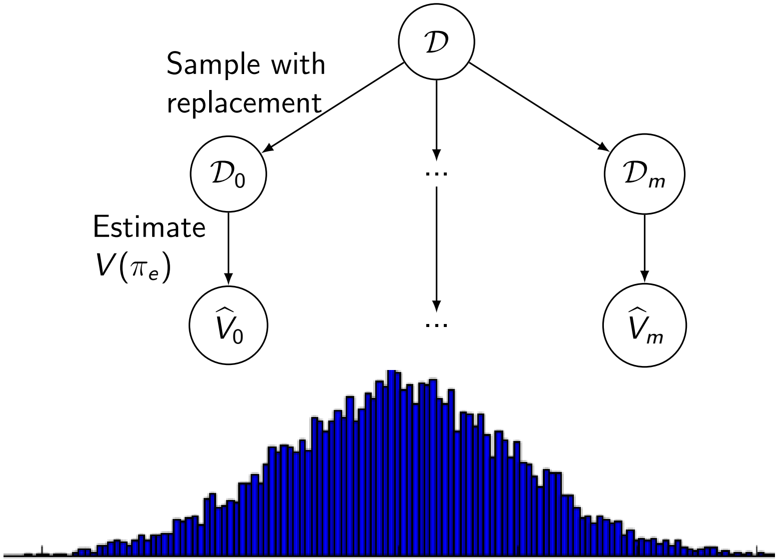
Contributions:

1. Two bootstrap methods that incorporate models for approximate high confidence policy evaluation.
2. Theoretical bound on model bias.
3. Empirical evaluation of proposed methods.

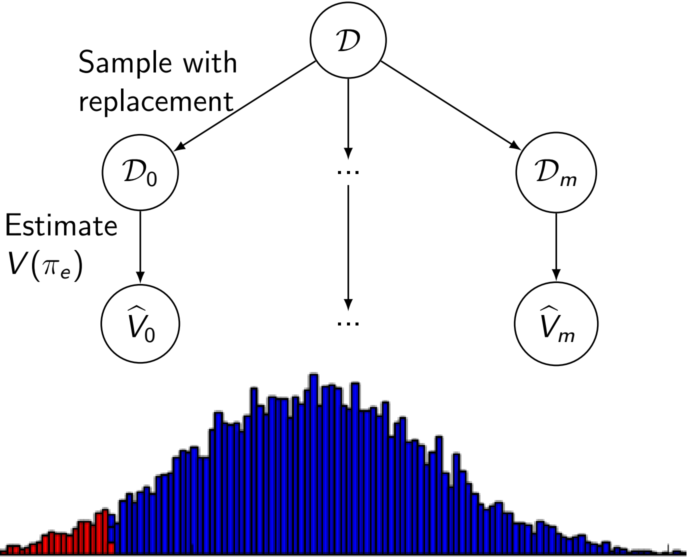
Bootstrap Confidence Intervals



Bootstrap Confidence Intervals



Bootstrap Confidence Intervals



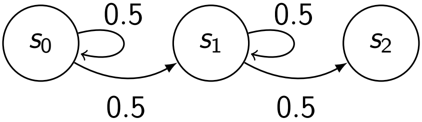
Data-Efficient Confidence Intervals

We draw on two ideas to reduce the number of trajectories required for tight confidence bounds.

- ✓ Replace exact confidence bounds with bootstrap confidence intervals.
- ■ Use learned models of the environment's transition function to reduce variance.

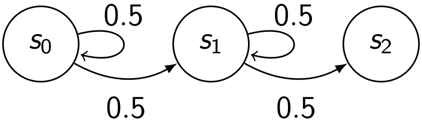
Model Based Off-Policy Evaluation

Trajectories are generated from an MDP, $M = \langle \mathcal{S}, \mathcal{A}, P, r \rangle$.

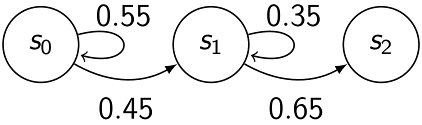


Model Based Off-Policy Evaluation

Trajectories are generated from an MDP, $M = \langle \mathcal{S}, \mathcal{A}, P, r \rangle$.



Model Based off-policy estimator use all trajectories to estimate the *unknown* transition function, P.



Model-Based off-policy estimator: $\widehat{V}(\pi_e) := V_{\widehat{M}}(\pi_e)$ where $\widehat{M} = \langle \mathcal{S}, \mathcal{A}, \widehat{P}, r \rangle$ where $\widehat{P}$ is the learned transition function.

Model-Bias

Model-Based approaches may have high bias.

1. **Lack of Data:** When we lack data for a particular (S, A) pair then we must make assumptions about the transition probability, $P(\cdot|S, A)$.
2. **Model Representation:** The true function P may be outside the class of models we consider.

Model-Bias

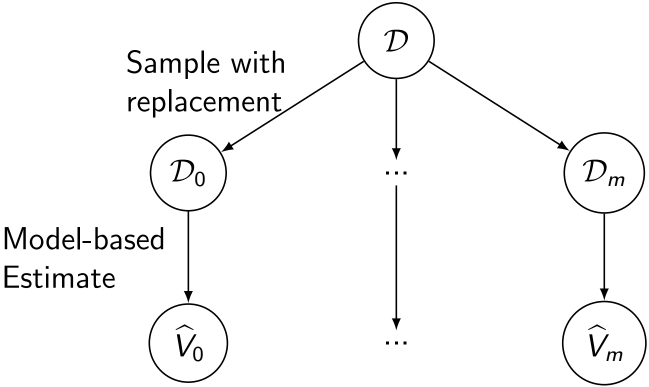
Model-Based approaches may have high bias.

1. **Lack of Data:** When we lack data for a particular (S, A) pair then we must make assumptions about the transition probability, $P(\cdot|S, A)$.
2. **Model Representation:** The true function P may be outside the class of models we consider.

We show theoretically that model bias depends on:

- The *importance-sampled* train / test error when building the model.
- The horizon length.
- The maximum reward.

Model-Based Bootstrap



Existing Methods

- Importance-sampling based methods.
- Bootstrap importance-sampling

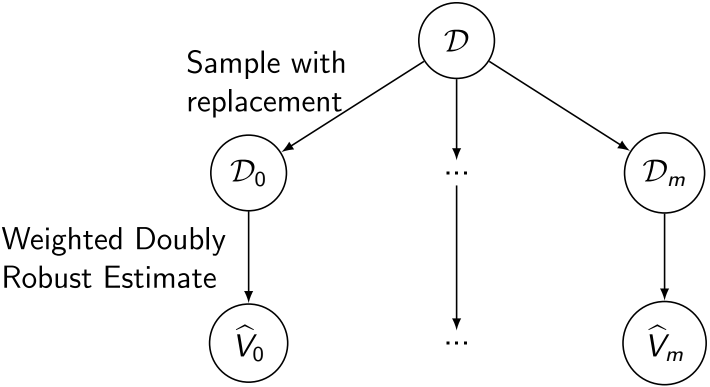
- MB-BOOTSTRAP (ours)

Doubly Robust Estimator [Jiang and Li, 2016, Thomas and Brunskill, 2016]

$$\mathrm{DR}(\mathcal{D}) := \underbrace{\mathrm{PDIS}(\mathcal{D})}_{\text{Unbiased estimator}} - \underbrace{\sum_{i=1}^{n}\sum_{t=0}^{L} w_t^i \hat{q}^{\pi_e}(S_t^i, A_t^i) - w_{t-1}^i \hat{v}^{\pi_e}(S_t^i)}_{\text{Zero in Expectation}}$$

- $\hat{v}^{\pi}(S) := \mathbb{E}_{A\sim\pi, S'\sim\hat{P}(\cdot|S,A)}\left[r(S,A) + \hat{v}(S')\right]$
  - State value function.
- $\hat{q}^{\pi}(S,A) := r(S,A) + \mathbb{E}_{S'\sim P(\cdot|S,A)}\left[\hat{v}(S')\right]$
  - State-action value function.
- w_t is the importance weight of the first t time-steps.

Weighted Doubly Robust Bootstrap



Bootstrapping with Models

MB-Bootstrap (Model-Based Bootstrap)

- **Advantages:** Low variance.
- **Disadvantages:** Potentially high bias.

WDR-Bootstrap (*Weighted* Doubly Robust Bootstrap)

- **Advantages:** Low bias.
- **Disadvantages:** Potentially higher variance.

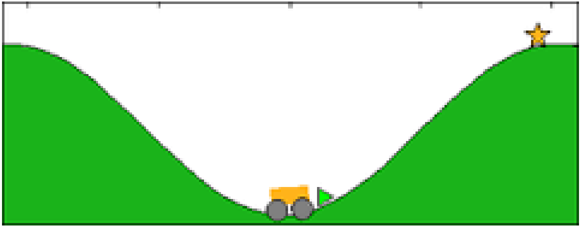
Existing Methods

- Importance-sampling based methods.
- Bootstrap importance-sampling

- WDR-BOOTSTRAP (ours)

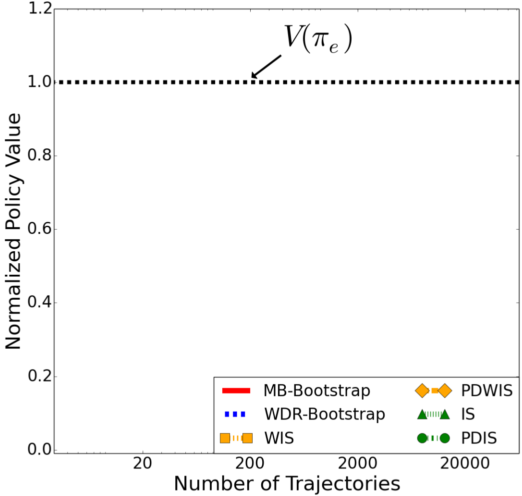
- MB-BOOTSTRAP (ours)

MountainCar Domain

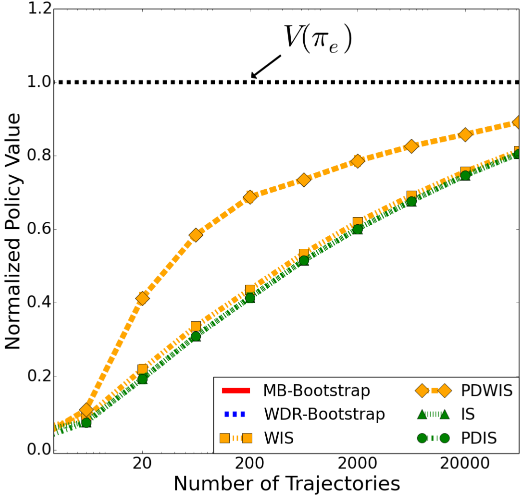


- State and action spaces are discretized.
- Models use a tabular representation.

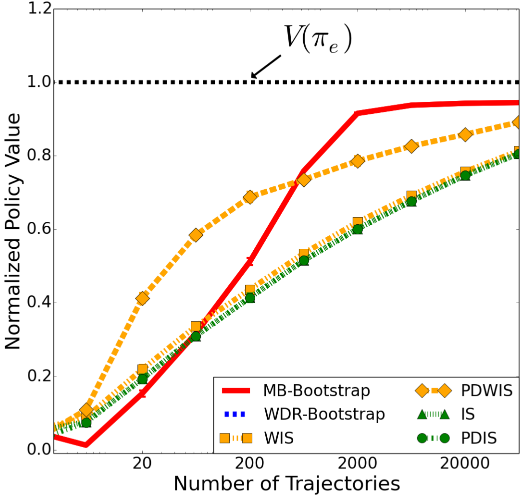
Mountain Car Domain



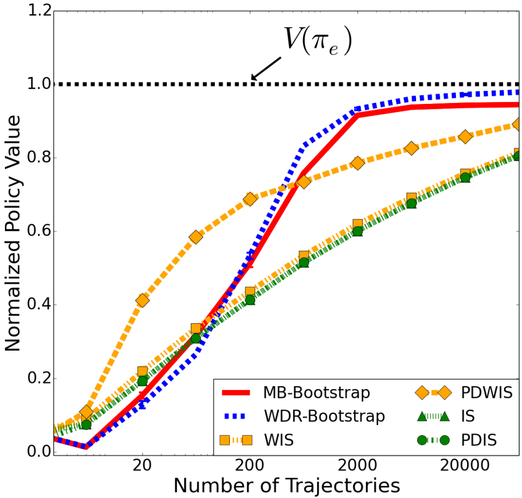
Mountain Car Domain



Mountain Car Domain



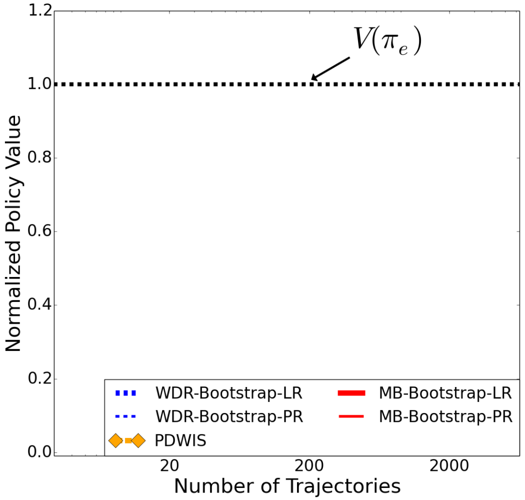
Mountain Car Domain



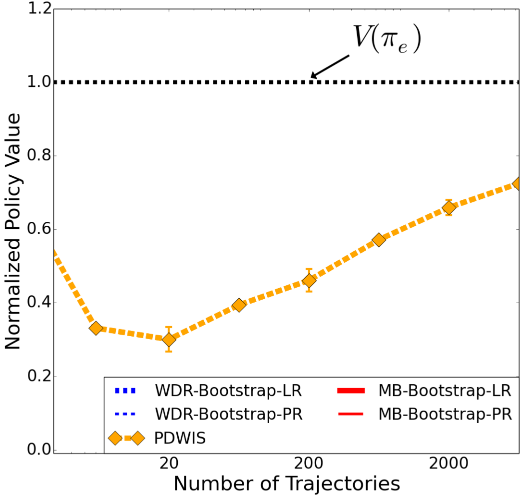
Cliffworld Domain

- Agent must cross a narrow path to reach a goal.
- State is cartesian position and velocity. The agent moves by selecting acceleration.
- Linear Gaussian dynamics.
- Models are learned with linear and polynomial regression.

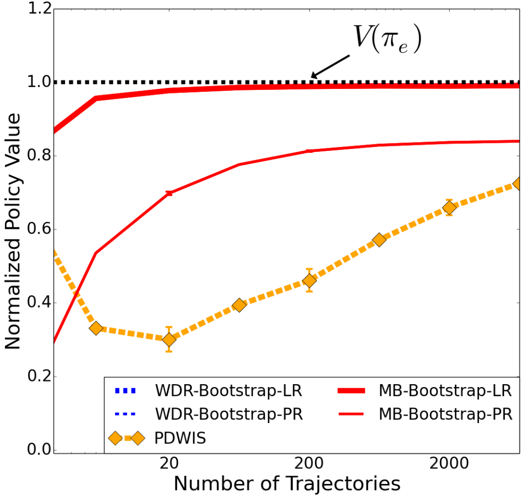
Cliffworld Domain



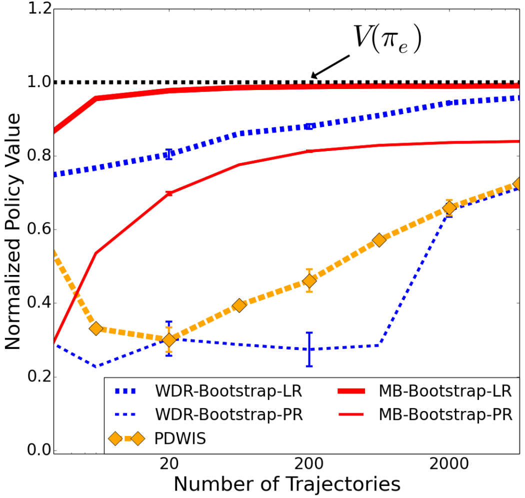
Cliffworld Domain



Cliffworld Domain



Cliffworld Domain



Conclusion

1. Two bootstrap methods that incorporate models for approximate high confidence policy evaluation.
2. Theoretical bound on model bias.
3. Empirical evaluation of proposed methods.

Future Work

- Investigate ways to “blend” MB-Bootstrap and WDR-Bootstrap for further improvements.

Future Work

- Investigate ways to "blend" MB-Bootstrap and WDR-Bootstrap for further improvements.
- Application to evaluating policies learned in simulation.

- Importance-sampling methods.
- WDR-BOOTSTRAP
- MB-BOOTSTRAP

Learning Agents Research Group
The University of Texas at Austin

Personal Autonomous Robotics Lab

Thanks for your attention!
Questions?

Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML*, 2016.

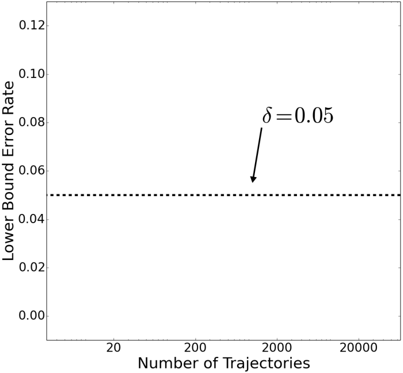
Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

P. S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence off-policy evaluation. In *Association for the Advancement of Artificial Intelligence, AAAI*, 2015a.

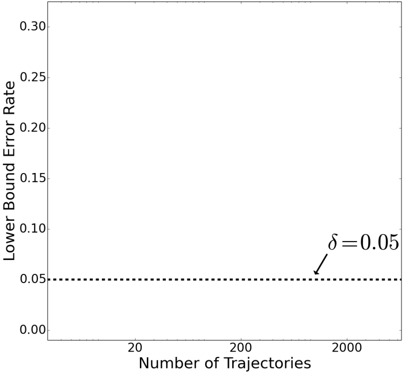
P. S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, 2015b.

P.S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML*, 2016.

Lower Bound Error

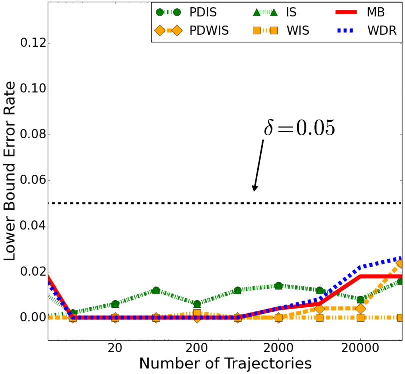


Mountain Car

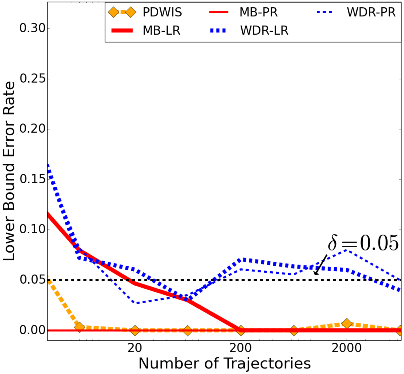


Cliffworld

Lower Bound Error



Mountain Car



Cliffworld

Prior Work: Importance Sampling [Precup et al., 2000]

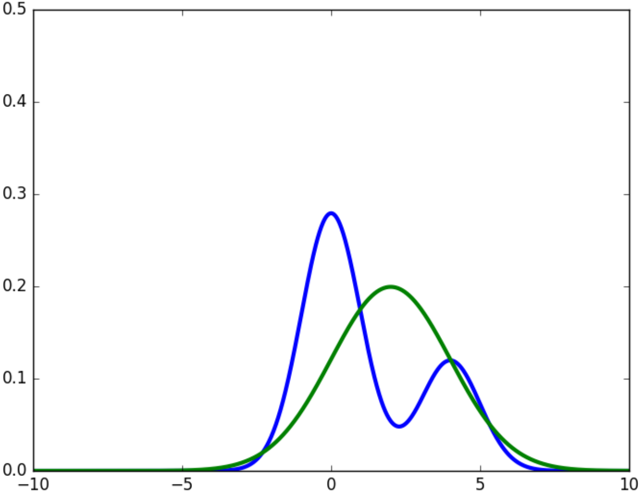
Re-weight return according to their relative likelihood:

$$\mathrm{IS}(\pi_e, H, \pi_b) := \underbrace{\prod_{t=0}^{L-1} \frac{\pi_e(A_i|S_i)}{\pi_b(A_i|S_i)}}_{\text{Importance weight}} \underbrace{\sum_{t=0}^{L-1} r(S_t, A_t)}_{\text{Observed Return}}$$

Mean of re-weighted returns is an unbiased estimate of $V(\pi_e)$:

$$\mathrm{IS}(\mathcal{D}) := \sum_{H \in \mathcal{D}} \mathrm{IS}(\pi_e, H, \pi_b)$$

Prior Work: Importance Sampling



Prior Work: Importance Sampling