Monte Carlo Methods
Summary

Reinforcement Learning
November 2, 2022

FH Zentralschweiz

Monte Carlo (MC) Methods

- Monte Carlo (MC) methods look at **whole episodes** and then average the complete returns
- Value estimation and policies are only changed **on the completion of an episode**
- **GPI** (Generalized Policy Iteration) can be used for the control problem
- MC generally uses **action-value estimates** in order to compute a greedy policy
- We must maintain **exploration** to update all action-value estimates

Monte Carlo Prediction (first visit)

Monte Carlo Prediction for estimating v_π

Input: a policy π

Initialize:

$V(s) \in \mathbb{R}$ (arbitrarily)

$\text{Returns}(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever:

Generate episode following π: $S_0, A_0, R_0, S_1, ..., R_T$

$G \leftarrow 0$

Loop for each step of the episode, $t = T-1, T-2, ..., 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless S_t appears in $S_{t-1}, ...S_0$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow \text{average}(Returns(S_t))$

On policy vs. off policy algorithm

- **On-policy** algorithms learn a policy while following this policy in the algorithm
- **Off-policy** algorithms learn a policy **different** from the one used to generate the data

- On-policy algorithms require a *soft* policy, which means that

$$\pi(a|s) > 0, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}$$

On-policy first-visit MC control, estimates $\pi \approx \pi_*$

Input: (small) $\epsilon > 0$
Initialize:
  $\pi \leftarrow$ an arbitrary ϵ-soft policy
  $Q(s, a) \in \mathbb{R}$ (arbitrarily, for example $= 0$)
  $\mathit{Returns}(s, a) \leftarrow$ empty list
Loop forever: (for each episode)
  Generate an episode following π: $S_0, A_0, R_0, S_1, ..., R_T$
  $G \leftarrow 0$
  Loop for each step of the episode, $t = T-1, T-2, ..., 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    Unless (S_t, A_t) appears in $(S_{t-1}, A_{t-1}), ..., (S_0, A_0)$:
      Append G to $\mathit{Returns}(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow \text{average}(\mathit{Returns}(S_t, A_t))$
      $A^* \leftarrow \text{argmax}_a Q(S_t, a)$, ties broken arbitrarily
      For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \epsilon + \epsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \epsilon/|\mathcal{A}(S_t)| & \text{otherwise} \end{cases}$$

Importance Sampling

- Importance sampling *corrects* the value of the return by the probability ratio that this state would be visited by the policies

$$
\begin{aligned}
\rho_{t:T} &\doteq \frac{\prod_{k=1}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=1}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)} \\
&= \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}
\end{aligned}
$$

Importance Sampling

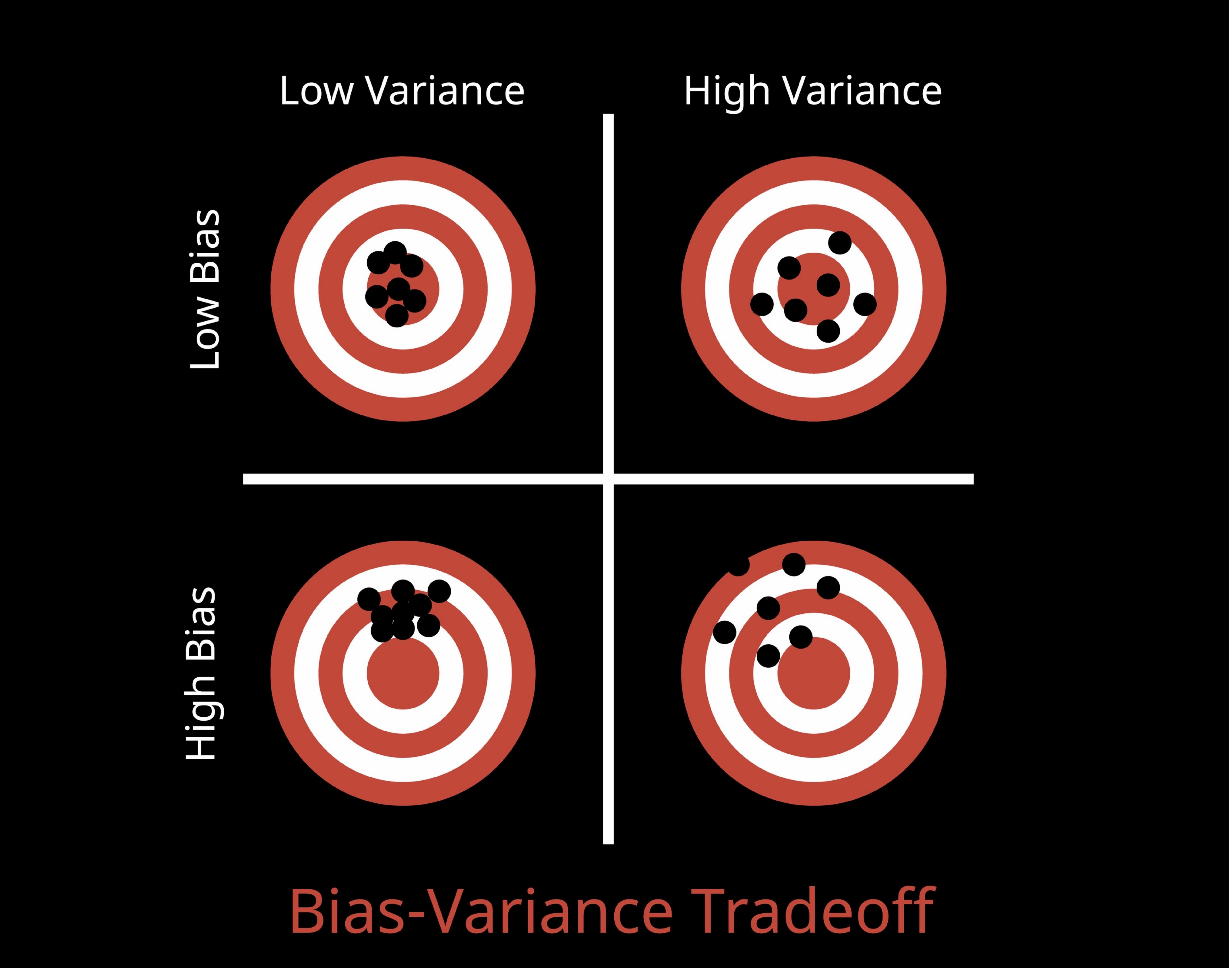
Ordinary importance sampling averages the results:

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|}$$

while weighted importance sampling uses a weighted average:

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

Formally, ordinary importance sampling is unbiased but has a high (unbounded) variance, while weighted importance sampling is biased, but has a low (converging to zero) variance.


Low Variance
High Variance
Low Bias
High Bias
Bias-Variance Tradeoff

Off-policy MC control, estimates $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}, a \in \mathcal{A}$:

$Q(s,a) \in \mathbb{R}$ (arbitrarily, for example $= 0$)

$C(s,a) \leftarrow 0$

$\pi(s) \leftarrow \text{argmax}_a Q(s,a)$ (with ties broken consistently)

Loop forever (for each episode):

$b \leftarrow$ any soft policy

Generate an episode following b: $S_0, A_0, R_0, S_1, ..., R_T$

$G \leftarrow 0$

$W \leftarrow 1$

Loop for each step of the episode, $t = T-1, T-2, ..., 0$:

$G \leftarrow \gamma G + R_{t+1}$

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t]$

$\pi(a|S_t) \leftarrow \text{argmax}_a Q(S_t, A_t)$ (with ties broken consistently)

If $A_t \neq \pi(S_t)$ then exit inner Loop (next episode)

$W \leftarrow W \frac{1}{b(A_t|S_t)}$