# Multi-step Off-policy TD with Control Variates

Sutton & Barto - 5.8, 5.9, 7.4, 12.8, 12.9

KRIS DE ASIS

# Overview

- Importance sampling
- Control variates
- n-step TD control variates
- Back to eligibility traces
- Experiments

# Importance Sampling

With a **target policy** $\pi(a|s)$, and **behavior policy** $\mu(a|s)$:

$$\rho_t \triangleq \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$$

To correct a **trajectory**:

$$\prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\mu(A_k|S_k)p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} = \prod_{k=t}^{T-1} \rho_k$$

# Importance Sampling

Scaling the **target**:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[ \left( \prod_{k=t+1}^{T-1} \rho_k \right) \hat{G}_t - Q(S_t, A_t) \right]$$

Scaling the **TD error**:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left( \prod_{k=t+1}^{T-1} \rho_k \right) [\hat{G}_t - Q(S_t, A_t)]$$

# Importance Sampling

**Per-decision** importance sampling:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[\hat{G}_t - Q(S_t, A_t)\right]$$

$$\hat{G}_t = R_{t+1} + \gamma \rho_{t+1} R_{t+2} + \gamma^2 \rho_{t+1} \rho_{t+2} R_{t+3} + \gamma^3 \rho_{t+1} \rho_{t+2} \rho_{t+3} R_{t+4} \ldots$$

Can be written recursively:

$$\hat{G}_t = R_{t+1} + \gamma \rho_{t+1} \hat{G}_{t+1}$$

# Control Variates

Uses the estimation error of a **known quantity** to try and reduce the estimation error of an **unknown quantity**

# Control Variates

When estimating $\mathbb{E}[X]$, they're often of the form:

$$X^* = X + c(Y - \mathbb{E}[Y])$$

Where $Y$ is a random variable with a **known** expected value

$X^*$ has the following variance:

$$\mathrm{Var}(X^*) = \mathrm{Var}(X) + c^2\mathrm{Var}(Y) + 2c\mathrm{Cov}(X, Y)$$

# Control Variates

$$\mathrm{Var}(X^*) = \mathrm{Var}(X) + c^2\mathrm{Var}(Y) + 2c\mathrm{Cov}(X, Y)$$

This variance is minimized by:

$$c^* = -\frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}$$

With a resulting variance of:

$$\mathrm{Var}(X^*) = \big(1 - \mathrm{Corr}(X, Y)\big)\mathrm{Var}(X)$$

# n-Step TD Control Variates

n-step Sarsa:

$$\hat{G}_{t:h} = R_{t+1} + \gamma \rho_{t+1} \hat{G}_{t+1:h}$$
$$\hat{G}_{h:h} = Q(S_h, A_h) = \boldsymbol{Q_h}$$

Want to estimate $\mathbb{E}_\pi\left[\rho_{t+1} \hat{G}_{t+1:h}\right]$

All of these are conditioned on state

$$X^* = X + c(Y - \mathbb{E}[Y])$$

$$\mathbb{E}_\mu[\rho_{t+1} Q_{t+1}] = \mathbb{E}_\pi[Q_{t+1}]$$

$$\left(\rho_{t+1} \hat{G}_{t+1:h}\right)^* = \rho_{t+1} \hat{G}_{t+1:h} + c(\rho_{t+1} Q_{t+1} - \mathbb{E}_\pi[Q_{t+1}])$$

# n-Step TD Control Variates

$$\left(\rho_{t+1}\hat{G}_{t+1:h}\right)^* = \rho_{t+1}\hat{G}_{t+1:h} + c\left(\rho_{t+1}Q_{t+1} - \mathbb{E}_\pi[Q_{t+1}]\right)$$

$$c^* = -\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}$$

Assuming estimates are accurate, $Q_{t+1}$ is the expected value of the sampled return. In expectation, $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,Y) = \mathrm{Var}(Y)$

A reasonable choice is $c = -1$

# n-Step TD Control Variates

$$\left(\rho_{t+1}\hat{G}_{t+1:h}\right)^* = \rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}$$

Denote $\mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}$ as the **action-value control variate** (ACV)

n-step ACV Sarsa:

$$\hat{G}_{t:h} = R_{t+1} + \gamma\left(\rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}\right)$$
$$\hat{G}_{h:h} = Q_h$$

# n-Step TD Control Variates

In the one-step case:

$$\hat{G}_{t:t+1} = R_{t+1} + \gamma(\rho_{t+1}Q_{t+1} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1})$$
$$\hat{G}_{t:t+1} = R_{t+1} + \gamma\mathbb{E}_\pi[Q_{t+1}]$$

Gives **one-step Expected Sarsa**, but differs from **n-step Expected Sarsa**:

$$\hat{G}_{t:h} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \gamma^{h-t}\mathbb{E}_\pi[Q_h]$$

# n-Step TD Control Variates

$$\hat{G}_{t:h} = R_{t+1} + \gamma\left(\rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}\right)$$

First action is specified, but want the expectation across **all** sequences of actions after the first

The agent typically knows which action resulted in each reward

# n-Step TD Control Variates

$$\hat{G}_{t:h} = R_{t+1} + \gamma\left(\rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}\right)$$

$\rho_{t+1}Q_{t+1}$ is like a **guess** of what $\rho_{t+1}\hat{G}_{t+1:h}$ is going to be, since it knows which action led to the next reward

We want the expectation across all trajectories after the first reward, and $\mathbb{E}_\pi[Q_{t+1}]$ is a **guess** of this expectation

$\mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}$ is a **guess** of how far off the specific sampled reward sequence will be from the expectation across all trajectories- an **expectation correction** term

# n-Step TD Control Variates

$$\hat{G}_{t:h} = R_{t+1} + \gamma\left(\rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}\right)$$

Reduces variance due to stochasticity in the **policy**

… If the estimates are accurate

# n-Step TD Control Variates

$$\hat{G}_{t:h} = R_{t+1} + \gamma \mathbb{E}_{\pi}[Q_{t+1}] + \gamma \rho_{t+1}\left(\hat{G}_{t+1:h} - Q_{t+1}\right)$$

One-step Expected Sarsa's target, and a difference between the sampled sequence and what the agent expected to happen

Has an interpretation as **adaptive n-step** which **reduces n** based on a self-tested accuracy in the current value function estimates: $\hat{G}_{t+1:h} - Q_{t+1} \to 0$

It's doing both **per-decision** importance sampling, and scaling a **TD error** by the importance sampling ratio!

# n-Step TD Control Variates

What about state values? Following similar steps:

n-step TD (with per-decision importance sampling):

$$\hat{G}_{t:h} = \rho_t\left(R_{t+1} + \gamma \hat{G}_{t+1:h}\right)$$
$$\hat{G}_{h:h} = V(S_h) = \boldsymbol{V_h}$$

Applying the control variate:

$$\rho_t\left(R_{t+1} + \gamma \hat{G}_{t+1:h}\right)^* = \rho_t\left(R_{t+1} + \gamma \hat{G}_{t+1:h}\right) + c(\underbrace{\rho_t V_t}_{\mathbb{E}_\mu[\rho_t V_t]} - V_t)$$

# n-Step TD Control Variates

Under similar assumptions, we set $c = -1$:

$$\hat{G}_{t:h} = \rho_t\left(R_{t+1} + \gamma\hat{G}_{t+1:h}\right) + V_t - \rho_t V_t$$
$$\hat{G}_{t:h} = \rho_t\left(R_{t+1} + \gamma\hat{G}_{t+1:h}\right) + (1 - \rho_t)V_t$$

Denote $(1 - \rho_t)V_t$ as the **state-value control variate** (SCV)

$$\hat{G}_{t:h} = V_t + \rho_t\left(R_{t+1} + \gamma\hat{G}_{t+1:h} - V_t\right)$$

Also results in both **per-decision** importance sampling and scaling a **TD error** with the importance sampling ratio!

# n-Step TD Control Variates

Re-cap:

ACV: $\mathbb{E}_\pi[Q_t] - \rho_t Q_t$
SCV: $(1 - \rho_t)V_t$

SCV disappears when **on-policy** ($\rho_t = 1$), ACV does not

Of note, the derivation of the SCV also applies to action-values, as state-values can be computed through $V_t = \mathbb{E}_\pi[Q_t]$

# Back to Traces

The $\lambda$-return:

$$\hat{G}_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\hat{G}_{t:t+n}$$

# Back to Traces

$$\hat{G}_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} \hat{G}_{t:t+n}$$

Substituting n-step ACV Sarsa:

$$\hat{G}_{t:h} = R_{t+1} + \gamma\big(\rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}\big)$$
$$\hat{G}_{h:h} = Q_h$$

$$\hat{G}_t^{\lambda} = Q_t + \sum_{k=t}^{\infty} \underbrace{(R_{k+1} + \gamma\mathbb{E}_\pi[Q_{k+1}] - Q_k)}_{\text{One-step TD error}} \prod_{i=t+1}^{k} \gamma\lambda\rho_i$$

Trace-decay rate

# Back to Traces

$$\hat{G}_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\hat{G}_{t:t+n}$$

Substituting n-step ACV Sarsa:

$$\hat{G}_{t:h} = R_{t+1} + \gamma\big(\rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_\pi[Q_{t+1}] - \rho_{t+1}Q_{t+1}\big)$$
$$\hat{G}_{h:h} = Q_h$$

$$\hat{G}_t^\lambda = Q_t + \sum_{k=t}^{\infty}\left(R_{k+1} + \gamma\mathbb{E}_\pi[Q_{k+1}] - Q_k\right)\prod_{i=t+1}^{k}\gamma\lambda\rho_i$$

Sarsa($\lambda$): $$\hat{G}_t^\lambda = Q_t + \sum_{k=t}^{\infty}\left(R_{k+1} + \gamma\rho_{k+1}Q_{k+1} - Q_k\right)\prod_{i=t+1}^{k}\gamma\lambda\rho_i$$

# Back to Traces

$$\hat{G}_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\hat{G}_{t:t+n}$$

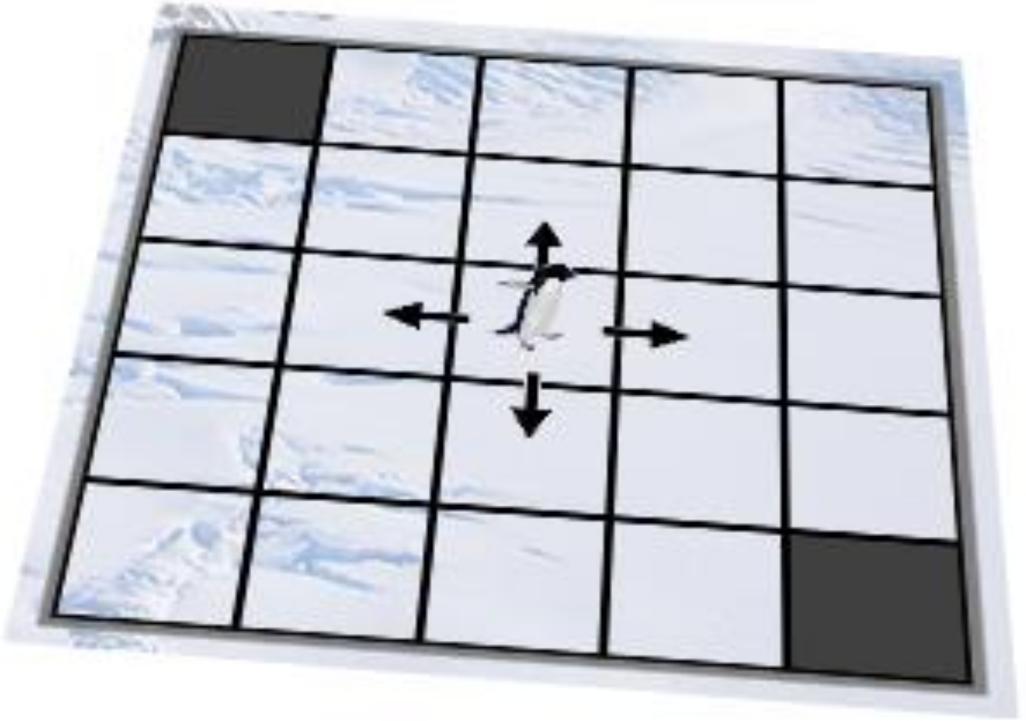
Substituting n-step SCV TD:

$$\hat{G}_{t:h} = \rho_t\left(R_{t+1} + \gamma\hat{G}_{t+1:h}\right) + (1-\rho_t)V_t$$
$$\hat{G}_{h:h} = V_h$$

$$\hat{G}_t^\lambda = V_t + \rho_t\sum_{k=t}^{\infty}(R_{k+1} + \gamma V_{k+1} - V_k)\prod_{i=t+1}^{k}\gamma\lambda\rho_i$$

Without SCV: $\hat{G}_t^\lambda = V_t + \sum_{k=t}^{\infty}(\rho_k(R_{k+1} + \gamma V_{k+1}) - V_k)\prod_{i=t+1}^{k}\gamma\lambda$

# Experiments – Off-policy Prediction

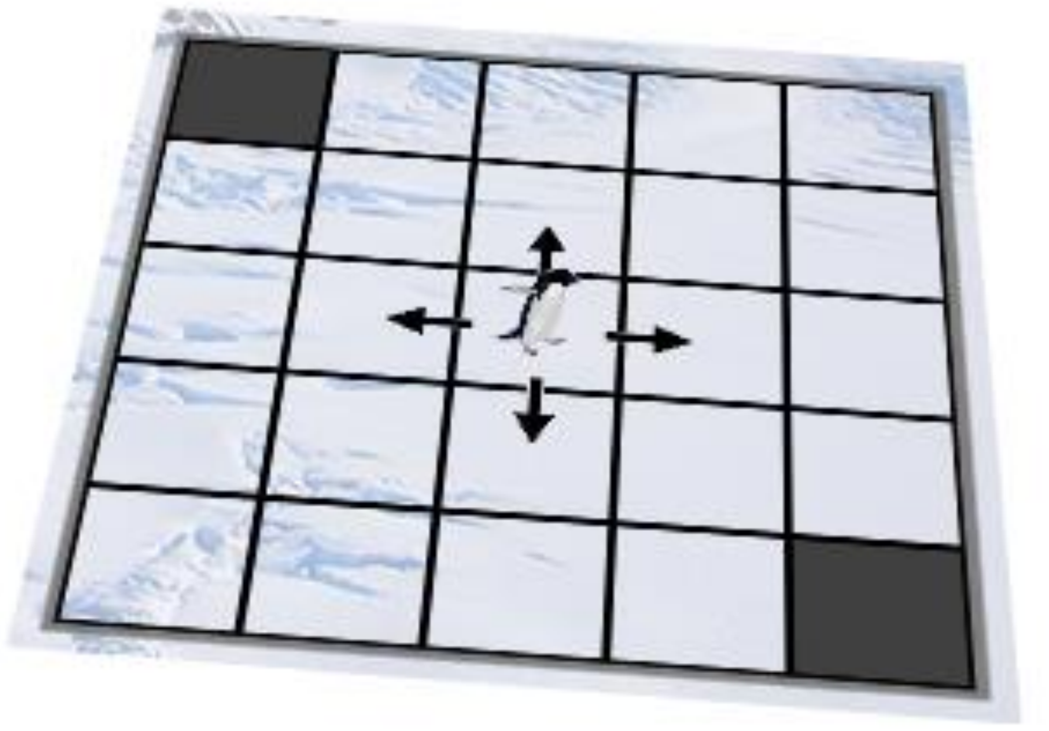


- 5×5 grid world, 2 terminal states
- 4 directional movement
- Reward of -1 at each step

- Equiprobable random **behavior policy**
- **Target policy** favors going north, all other actions equally likely

- Measured RMS erroring value estimates after 200 episodes

# Experiments – Off-policy Prediction

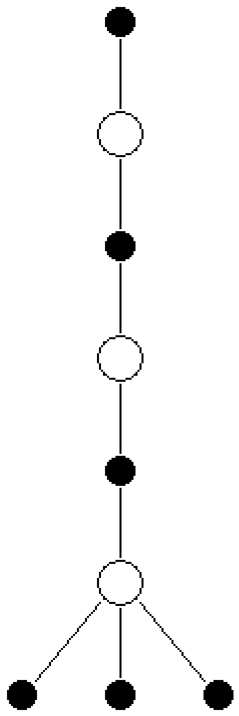


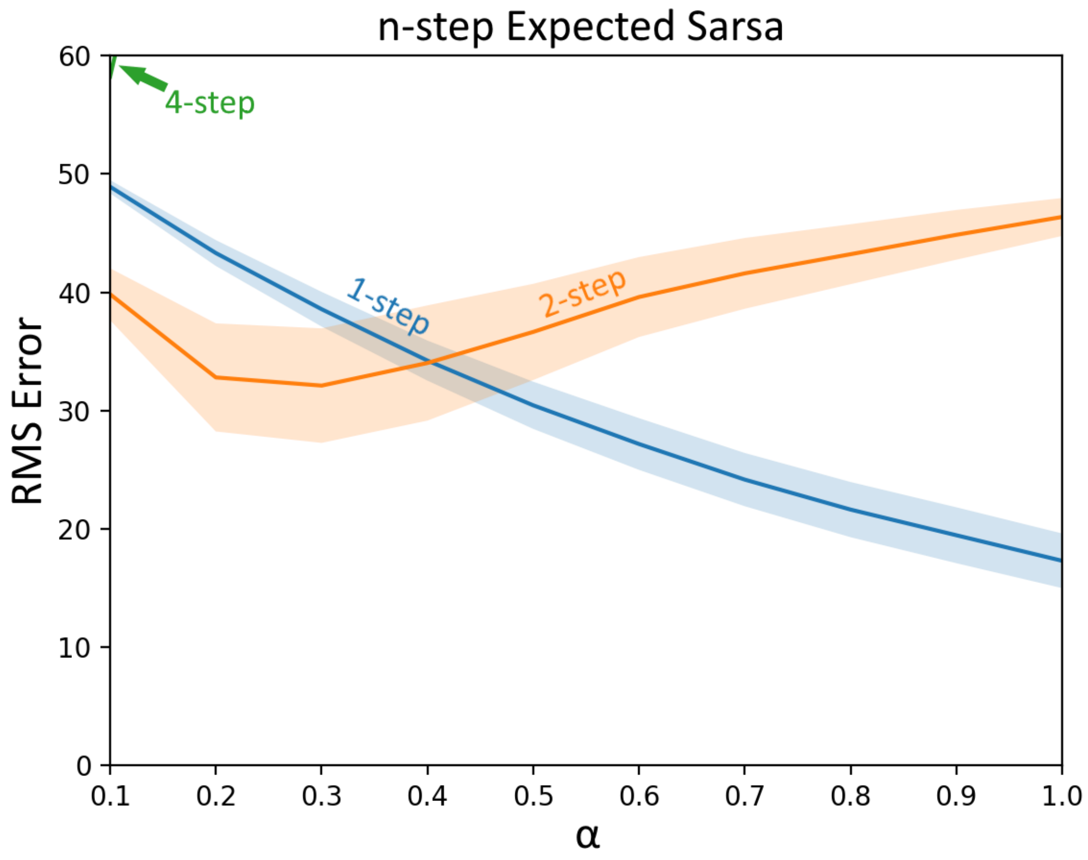
Comparing:

- n-step Expected Sarsa
- n-step SCV Sarsa $+ (1 - \rho_t)\mathbb{E}_\pi[Q_t]$
- n-step ACV Sarsa $+ \mathbb{E}_\pi[Q_t] - \rho_t Q_t$

# Experiments – Off-policy Prediction

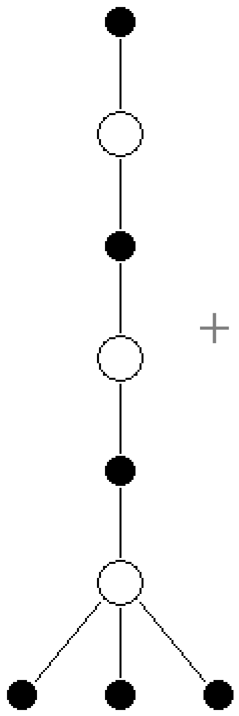
n-step Expected Sarsa



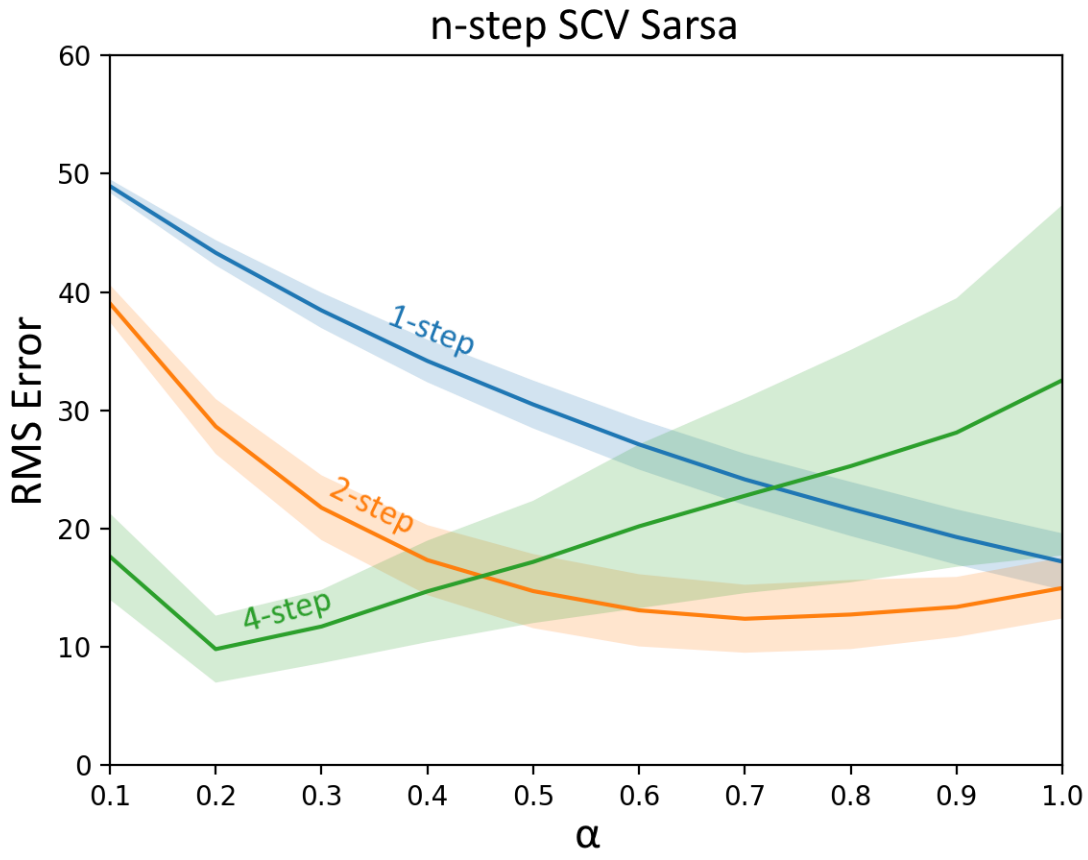


# Experiments – Off-policy Prediction

n-step SCV Sarsa

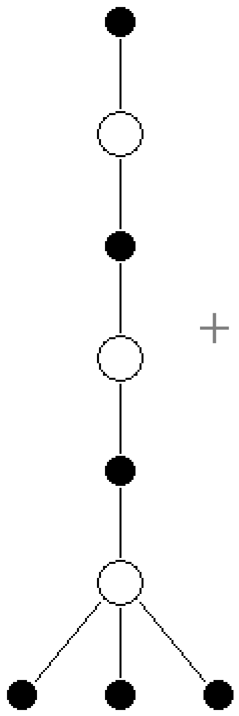


$+ (1 - \rho_t)\mathbb{E}_\pi[Q_t]$

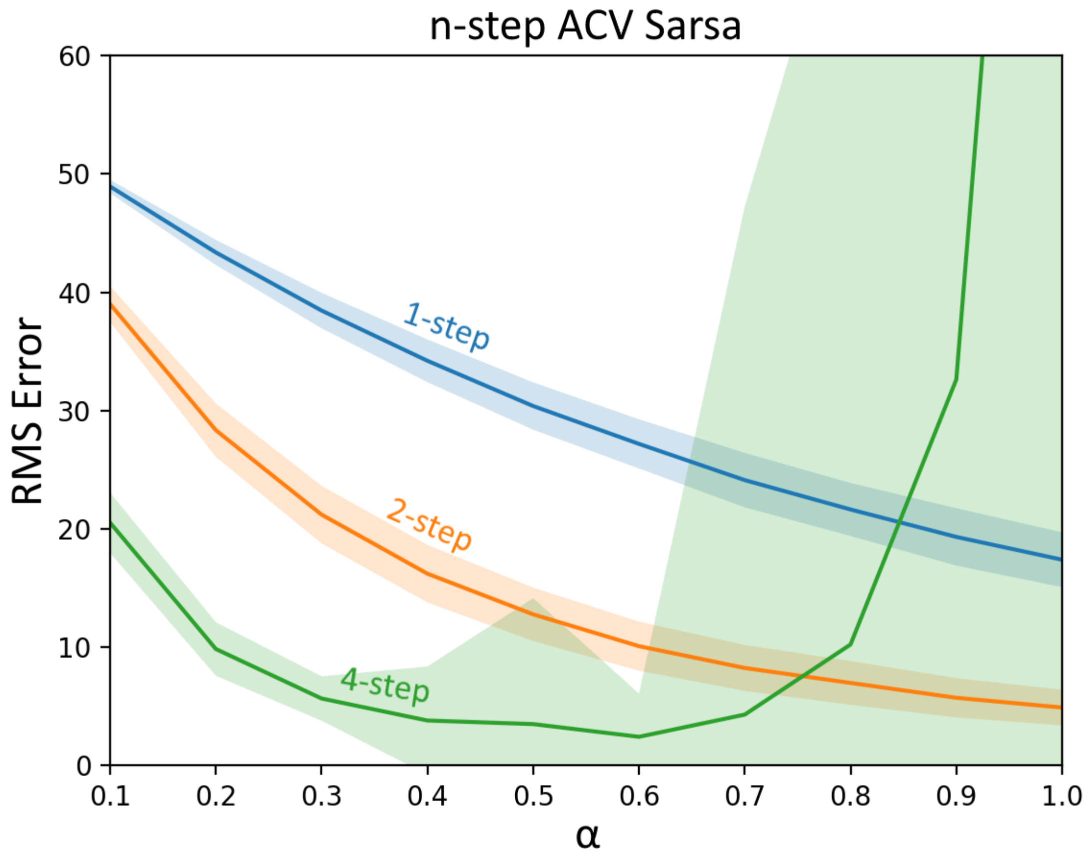


# Experiments – Off-policy Prediction

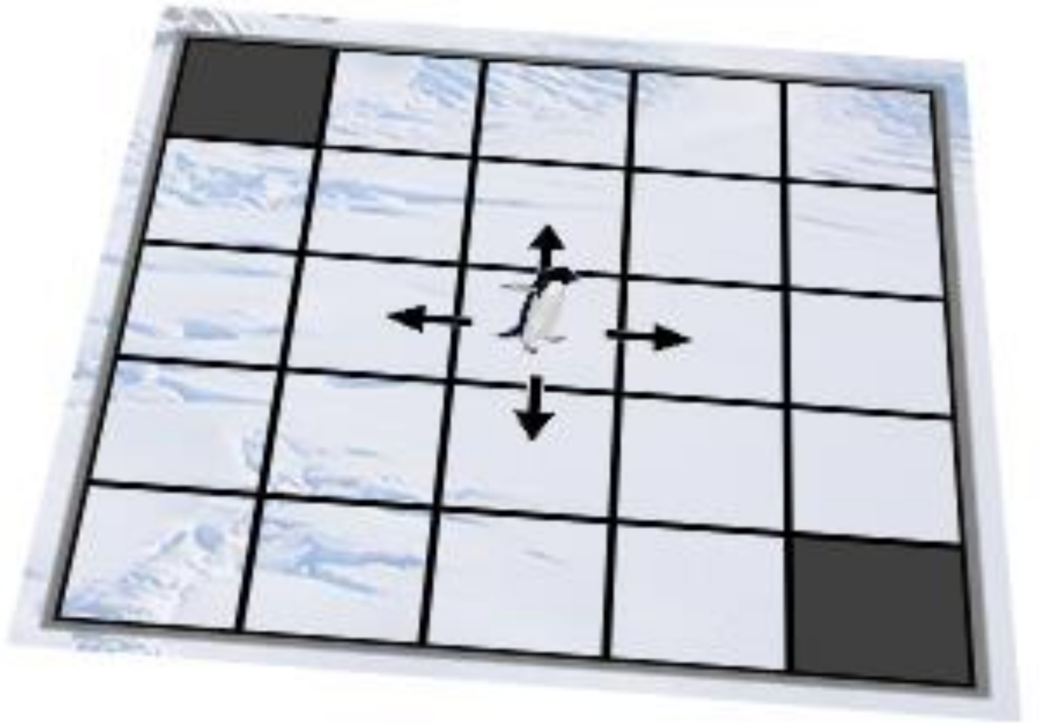
n-step ACV Sarsa



$+\mathbb{E}_\pi[Q_t] - \rho_t Q_t$



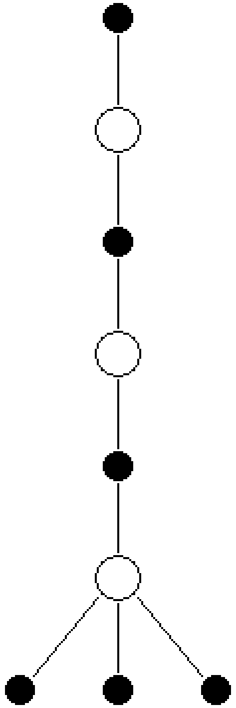
# Experiments – On-policy Prediction

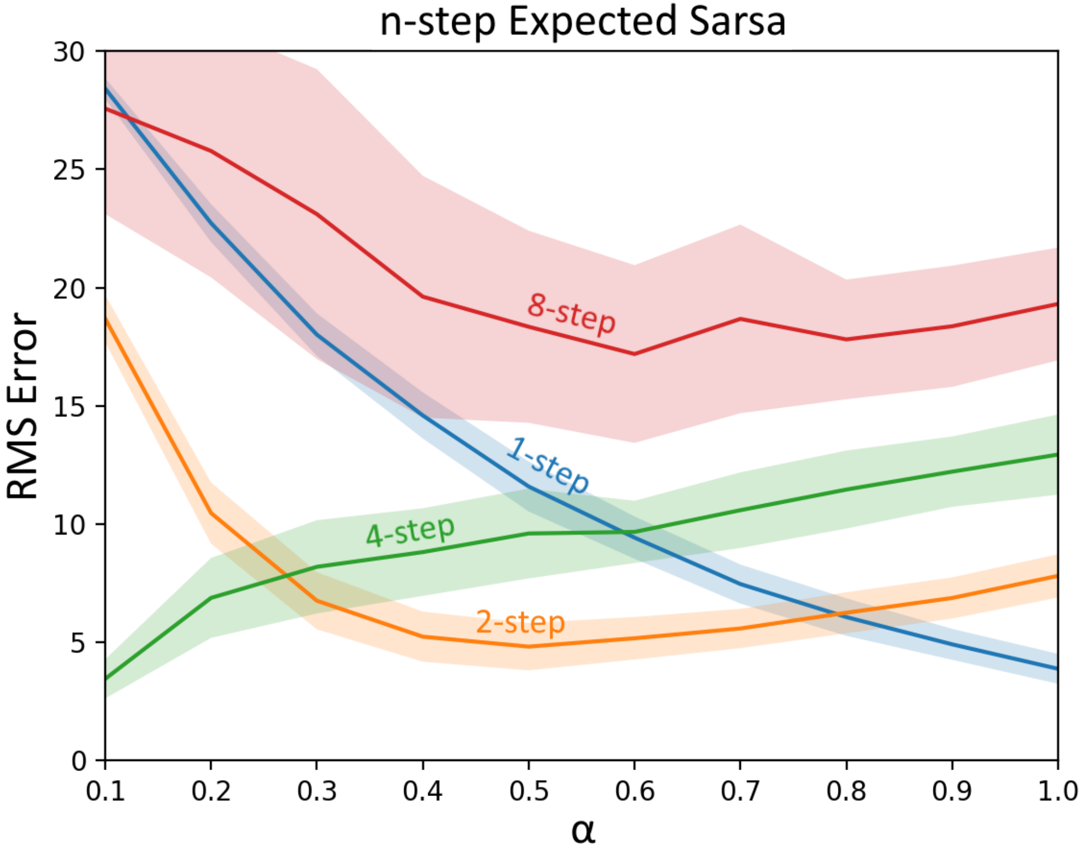


- Same environment, but **target policy** is also equiprobable random
- Did not compare n-step SCV Sarsa as the SCV disappears when on-policy

# Experiments – On-policy Prediction

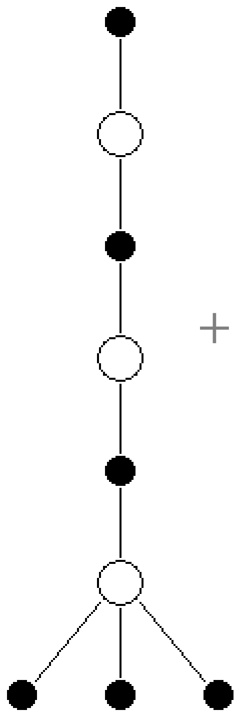
n-step Expected Sarsa



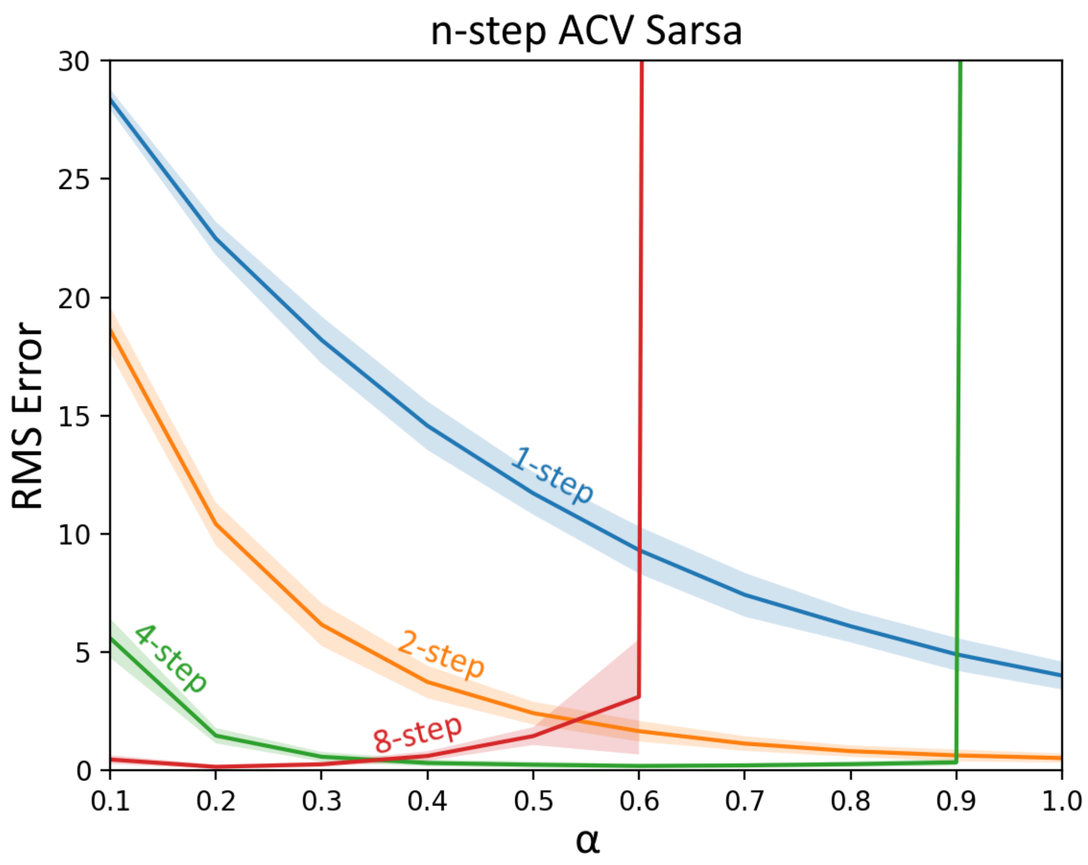


# Experiments – On-policy Prediction

n-step ACV Sarsa



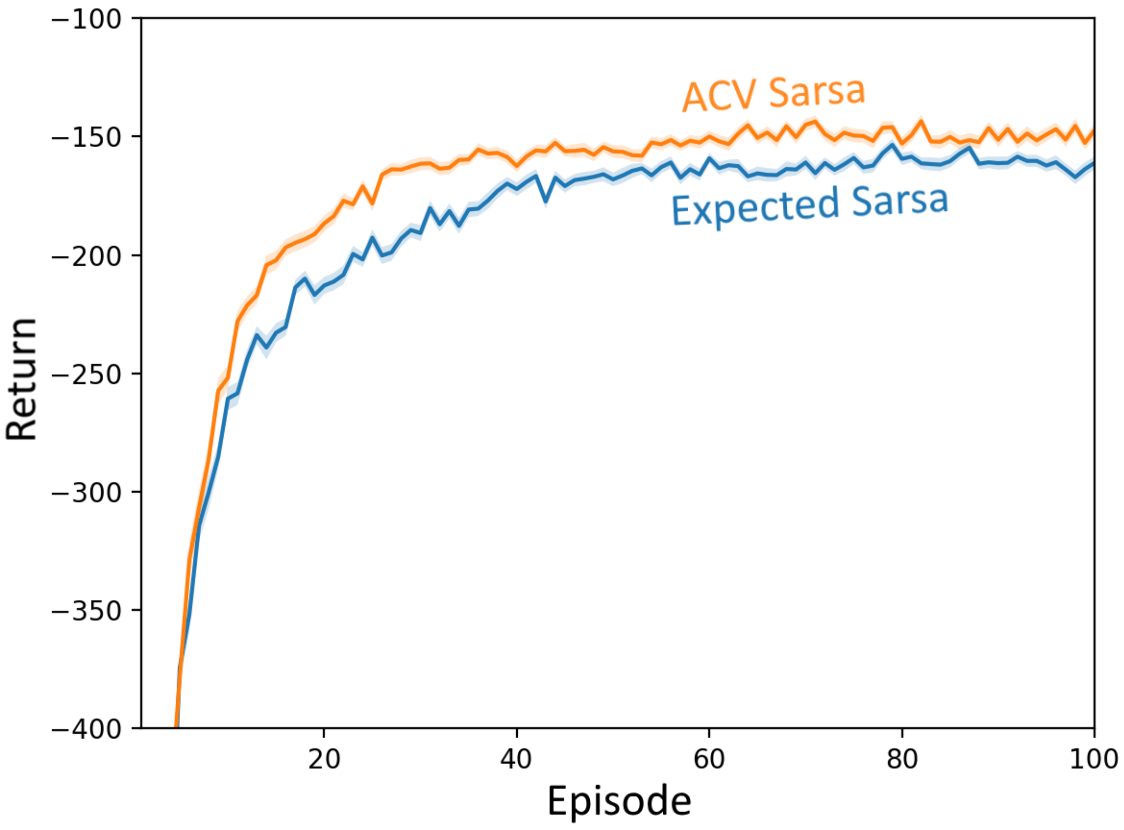
$+\mathbb{E}_\pi[Q_t] - \rho_t Q_t$



# Experiments – On-policy Control

- Mountain ~~car~~ penguin
- Starts in valley, terminates upon reaching the flag
- Reward of -1 at each step
- Linear function approximation (**Tile coding** with 16 8x8 tilings)
- $\epsilon$-greedy **on-policy control**
- Compared n-step ACV Sarsa to n-step Expected Sarsa

# Experiments – On-policy Control




- Comparing best $n$, $\alpha$ pair of each algorithm in terms of total reward over 100 episodes

- Of note, $\mathbb{E}_\pi[Q_t] - Q_t$ tends to be smaller for less-stochastic policies

# Summary

- Control variates can be used in n-step TD learning to reduce variance due to **policy stochasticity**

- In the action-value setting, they can produce an alternative multi-step generalization of **Expected Sarsa**

- They give a broader understanding of TD($\lambda$) algorithms and their underlying n-step returns

# Questions?

- De Asis, K., Sutton, R. S. (2018). Per-Decision Multi-step Temporal Difference Learning with Control Variates. *UAI 2018*.

KRIS DE ASIS