COMPARING DIRECT AND INDIRECT TEMPORAL-DIFFERENCE METHODS FOR ESTIMATING THE VARIANCE OF THE RETURN

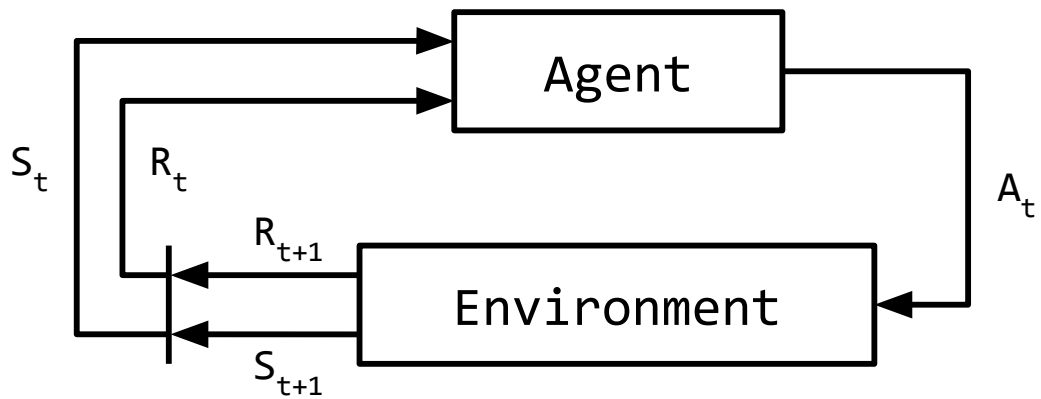
Craig Sherstan, Dylan R. Ashley*, Brendan Bennett*, Kenny Young,
Adam White, Martha White, Richard S. Sutton

Reinforcement Learning and Artificial Intelligence Laboratory, University of Alberta

BACKGROUND

WHAT IS REINFORCEMENT LEARNING?

Reinforcement learning considers an agent interacting with an environment:



The function the agent uses to pick actions in states is known as the policy. Often the challenge is to find a ”good” policy.

WHAT IS A GOOD POLICY?

In reinforcement learning the return is defined as follows:

$$G_t = R_{t+1} + \gamma_{t+1} R_{t+2} + \gamma_{t+1} \gamma_{t+2} R_{t+3} + \dots$$

Often we want to maximize the expected value of the return. But "good" does depend on what we want.

How can we learn it's expected value?

Temporal-difference (TD) methods have been fairly successful in tackling reinforcement learning problems so far. TD methods use predictions to update predictions.

One of the most straightforward TD methods is TD(λ):

$$\delta_t = R_{t+1} + \gamma_{t+1} w_t^T x_{t+1} - w_t^T x_t$$
$$z_t = \gamma_t \lambda_t z_{t-1} + x_t$$
$$w_{t+1} = w_t + \alpha_{t+1} \delta_t z_t$$

So what's this presentation about?

Recall what the return is:

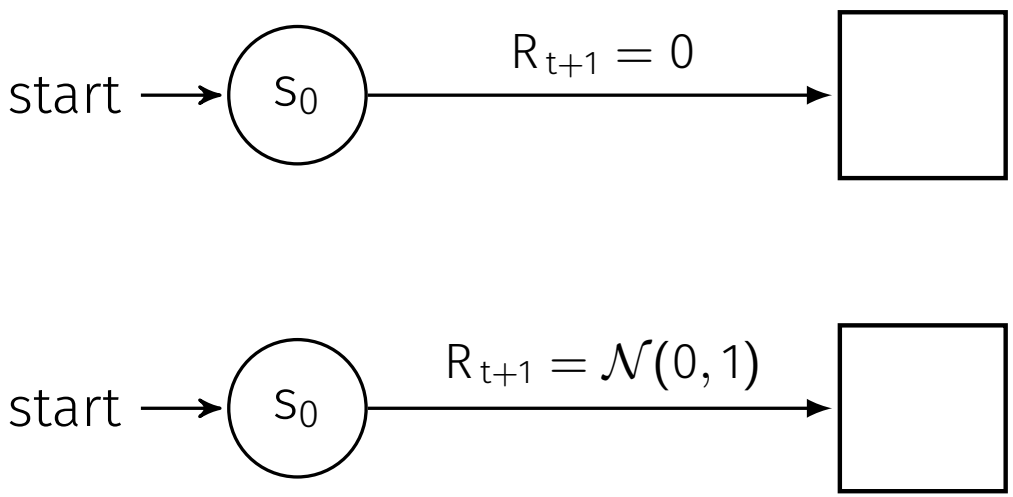
$$G_t = R_{t+1} + \gamma_{t+1} R_{t+2} + \gamma_{t+1} \gamma_{t+2} R_{t+3} + \dots$$

We're not limited to learning only it's expected value. We could also learn more parts of its distribution such as its variance.

MOTIVATION

WHY MIGHT WE WANT TO LEARN ITS VARIANCE?

The variance might tell us things about the distribution that the expected value can't. Sometimes these things are interesting. For example it could differentiate these two domains:



ANY BETTER REASONS?

The variance can give us useful information about the distribution. It can tell us how risky an action is to take in a state.

Humans take risk into decisions and don't necessarily act in a way that maximizes the expected value.

ANYTHING ELSE?

We can use an estimate of the variance to learn how to learn. For example here is an algorithm that uses an estimate of the variance to tune λ on the fly:

Algorithm 2: λ-greedy$(\mathbf{w}^{\text{err}}, \mathbf{w}^{\text{sq}}, \mathbf{w}_t, \mathbf{x}_t, \mathbf{x}_{t+1}, r_{t+1}, \rho_t)$

// Use GTD to update $\mathbf{w}^{\text{err}}$
$\bar{g}_{t+1} \leftarrow \mathbf{x}_{t+1}^\top \mathbf{w}^{\text{err}}$
$\delta_t \leftarrow r_{t+1} + \gamma_{t+1}\bar{g}_{t+1} - \mathbf{x}_t^\top \mathbf{w}^{\text{err}}$
$\bar{\mathbf{e}}_t = \rho_t(\gamma_t \bar{\mathbf{e}}_{t-1} + \mathbf{x}_t)$
$\mathbf{w}^{\text{err}} = \mathbf{w}^{\text{err}} + \alpha\delta_t\bar{\mathbf{e}}_t$
// Use VTD to update $\mathbf{w}^{\text{sq}}$
$\bar{r}_{t+1} \leftarrow \rho_t^2 r_{t+1}^2 + 2\rho_t^2\gamma_{t+1}r_{t+1}\bar{g}_{t+1}$
$\bar{\gamma}_{t+1} \leftarrow \rho_t^2\gamma_{t+1}^2$
$\bar{\delta}_t \leftarrow \bar{r}_{t+1} + \bar{\gamma}_{t+1}\mathbf{x}_{t+1}^\top \mathbf{w}^{\text{sq}} - \mathbf{x}_t^\top \mathbf{w}^{\text{sq}}$
$\bar{\mathbf{z}}_t = \bar{\gamma}_t\bar{\mathbf{z}}_{t-1} + \mathbf{x}_t$
$\mathbf{w}^{\text{sq}} = \mathbf{w}^{\text{sq}} + \alpha\bar{\delta}_t\bar{\mathbf{z}}_t$
// Compute λ estimate
$\text{errsq} = (\bar{g}_{t+1} - \mathbf{x}_{t+1}^\top \mathbf{w}_t)^2$
$\text{varg} = \max(0, \mathbf{x}_{t+1}^\top \mathbf{w}^{\text{sq}} - (\bar{g}_{t+1})^2)$
$\lambda_{t+1} = \text{errsq}/(\text{varg} + \text{errsq})$
return λ_{t+1}

LEARNING THE VARIANCE

How can we learn its variance?

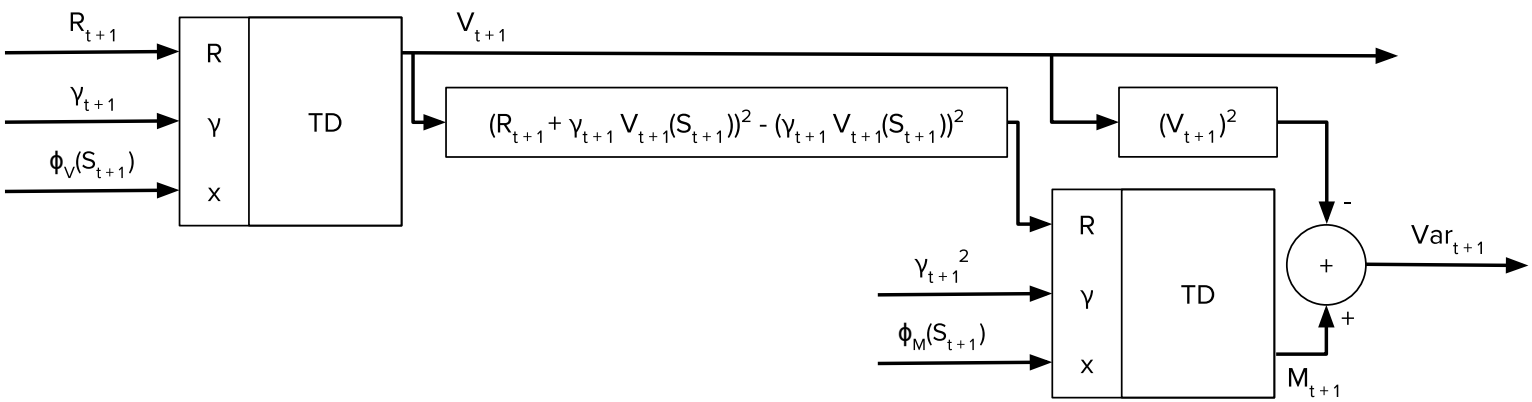
We can use this identity:

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2$$

If we are learning $\mathbb{E}_\pi[G_t \mid S_t = s]$ then we can just learn $\mathbb{E}_\pi[G_t^2 \mid S_t = s]$ on the side and use both our estimates to try to estimate $\mathrm{Var}_\pi(G_t \mid S_t = s)$.

WHAT WOULD THIS LOOK LIKE?

Using the identity $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2$ one can estimate the variance using the following structure:



IS THERE ANOTHER WAY?

We can also use this identity:

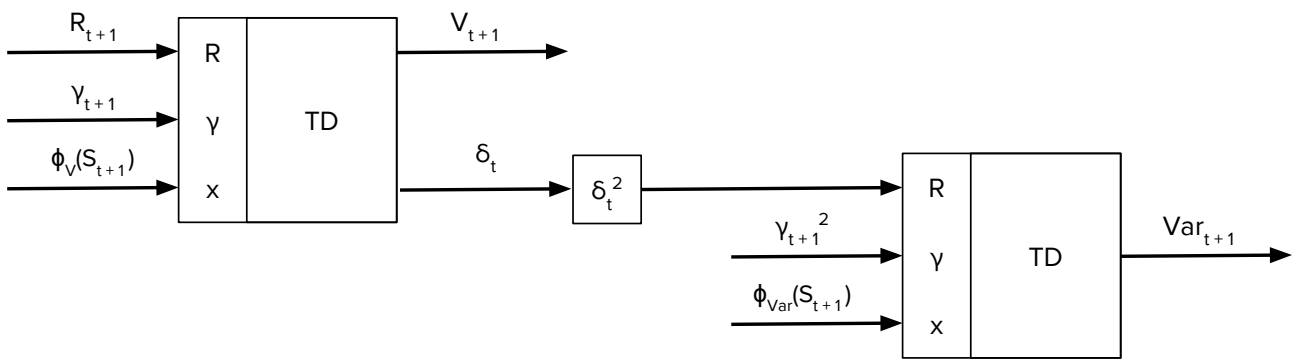
$$\mathrm{Var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]$$

If we are learning $\mathbb{E}_\pi[G_t \mid S_t = s]$ then we can approximate the variance using the following:

$$\mathrm{Var}_\pi(G_t \mid S_t = s) \approx \mathbb{E}_\pi\left[\delta_t^2 + \sum_{i=t+1}^{\infty}\left(\delta_i \prod_{j=t+1}^{i}\gamma_j\right)^2 \,\middle|\, S_t = s\right]$$

SO HOW WOULD THIS LOOK?

Using the identity $\mathrm{Var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$ one can estimate the variance using the following structure:



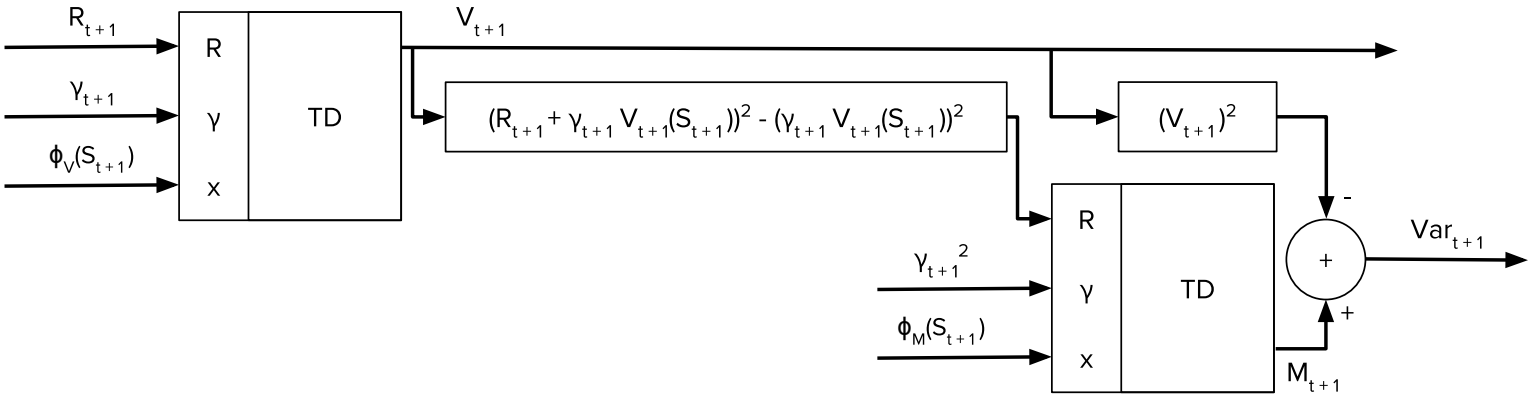
WHAT WOULD BE THE UPDATE EQUATIONS FOR THIS?

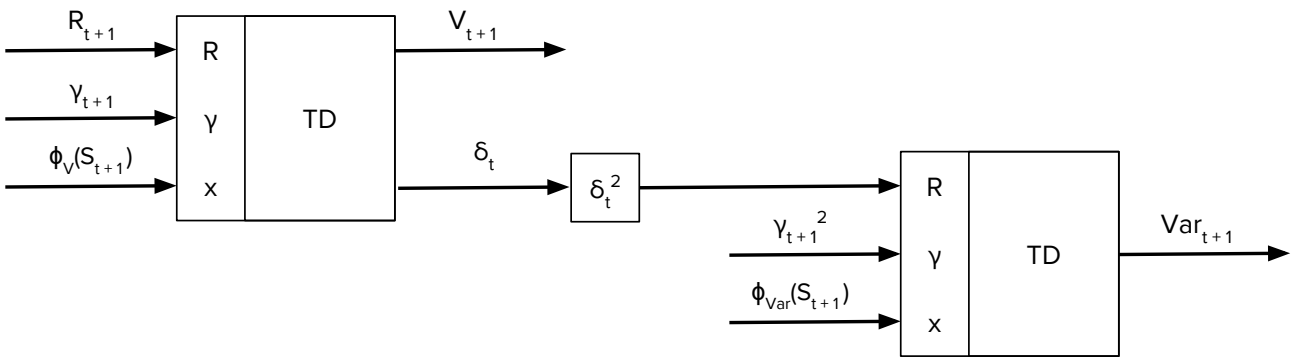
Using TD(λ) with $\overline{w}$ as the parameter vector for estimating the variance we obtain the following update equations:

$$
\begin{aligned}
\delta_t &= R_{t+1} + \gamma_{t+1} w_t^T x_{t+1} - w_t^T x_t \\
z_t &= \gamma_t \lambda_t z_{t-1} + x_t \\
w_{t+1} &= w_t + \alpha_{t+1} \delta_t z_t
\end{aligned}
$$

$$
\begin{aligned}
\overline{\delta}_t &= \delta_t^2 + \gamma_{t+1}^2 \overline{w}_t^T x_{t+1} - \overline{w}_t^T x_t \\
\overline{z}_t &= \gamma_t^2 \overline{\lambda}_t \overline{z}_{t-1} + x_t \\
\overline{w}_{t+1} &= \overline{w}_t + \overline{\alpha}_{t+1} \overline{\delta}_t \overline{z}_t
\end{aligned}
$$

HOW DO THEY COMPARE VISUALLY?





EMPIRICAL COMPARISON

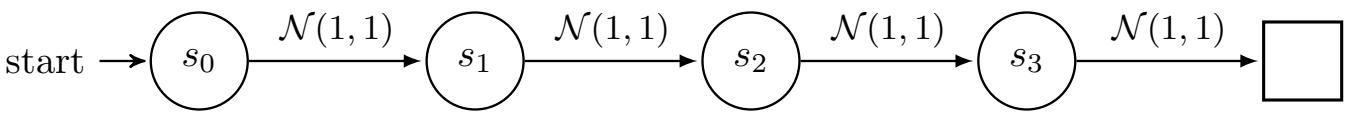
WHAT DO WE WANT TO KNOW?

Ideally we want to know if the direct method

- is faster or slower to converge than the indirect method,
- is more robust or less robust to differences in the value and variance learner, and
- performs better or worse under linear function approximation.

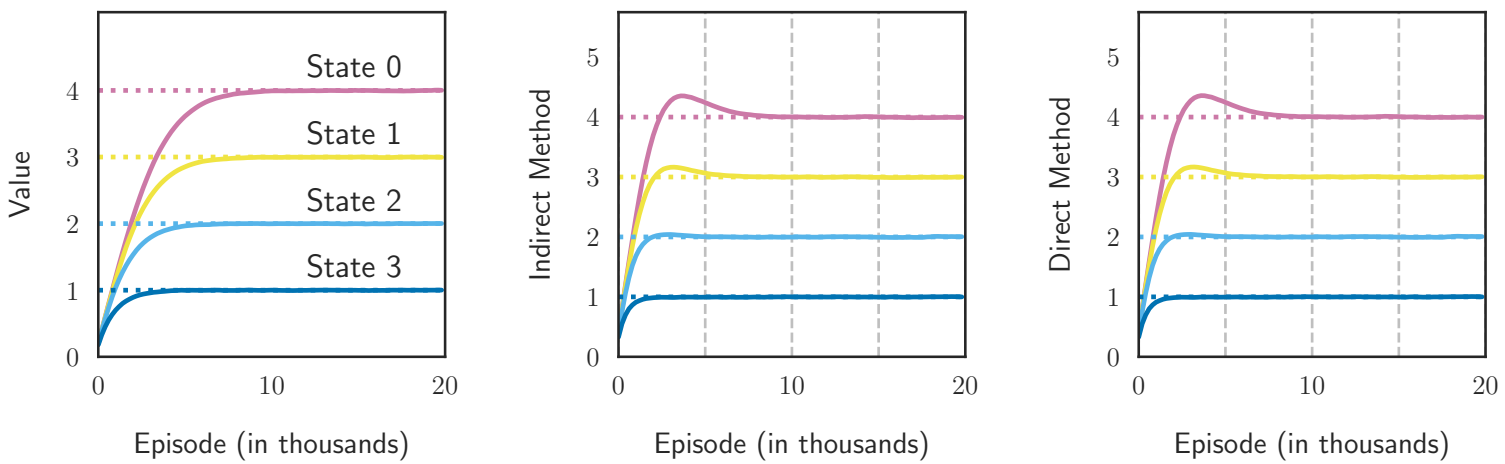
WHAT IS THE SIMPLEST DOMAIN WE CAN COMPARE THEM ON?

We begin by comparing them on the following simple Markov chain with gaussian rewards:



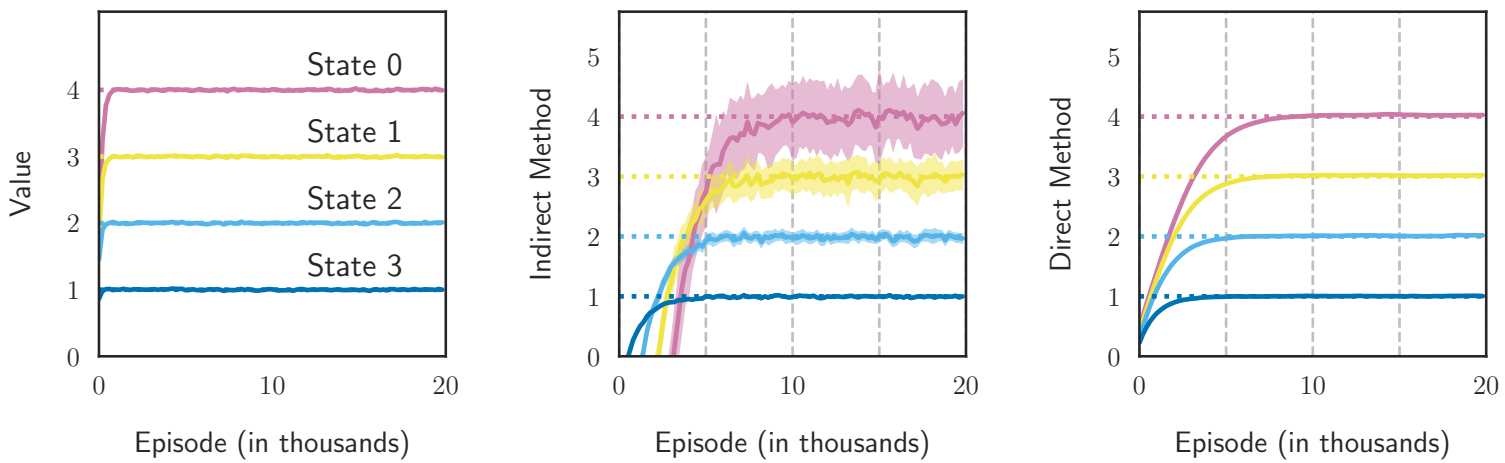
HOW DO THEY COMPARE WHEN $\alpha = \overline{\alpha}$?

When $\alpha = \overline{\alpha} = 0.001$ both perform roughly the same:



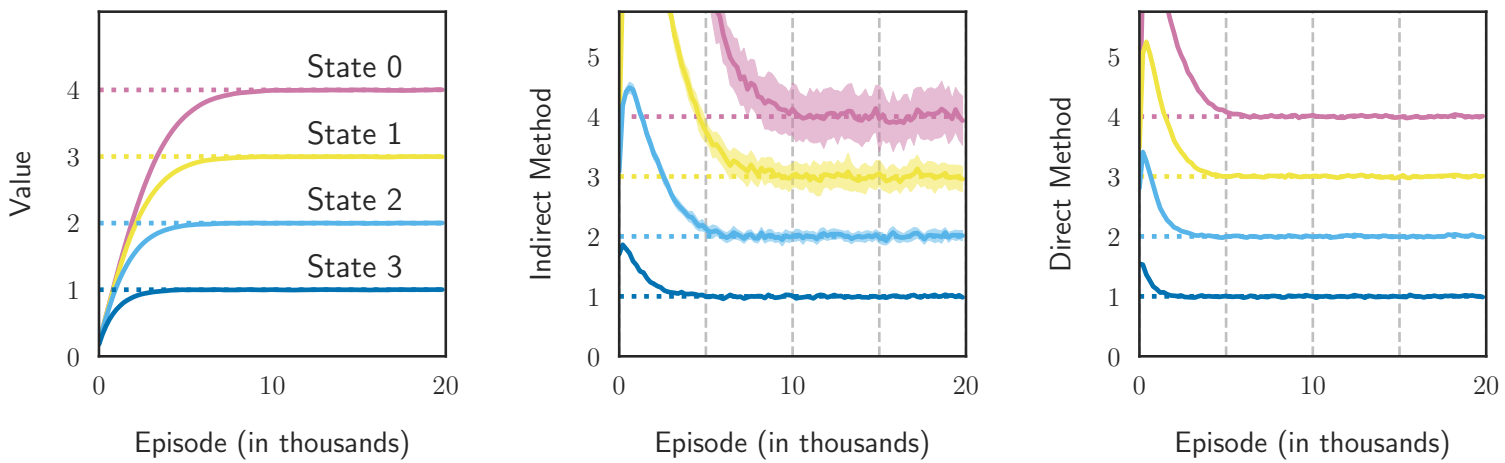
WHAT ABOUT WHEN $\alpha > \overline{\alpha}$?

When $\alpha = 0.01$ and $\overline{\alpha} = 0.001$ the variance of the indirect method is higher:



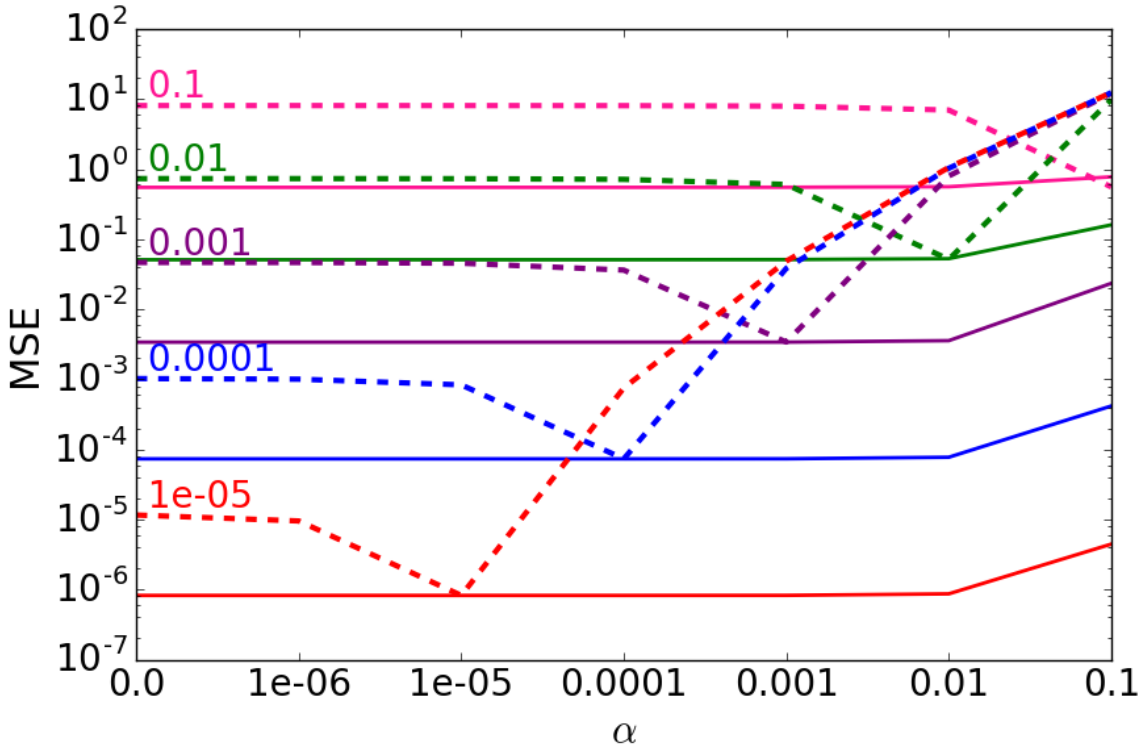
SO WHAT ABOUT WHEN $\alpha < \overline{\alpha}$?

When $\alpha = 0.001$ and $\overline{\alpha} = 0.01$ the variance of the indirect method is higher and the direct method is more stable:



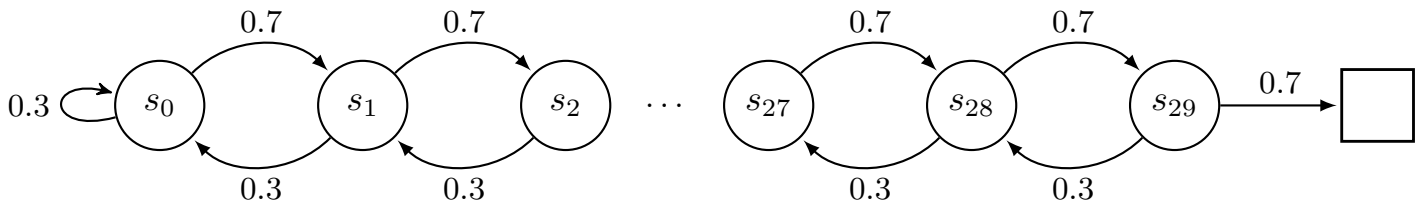
DOES THIS RESULT GENERALIZE?

In this domain we only see the two perform similarly when both step sizes are equal (note that the dotted line represents the indirect method and the solid line represents the direct method):



WHAT ABOUT UNDER FUNCTION APPROXIMATION?

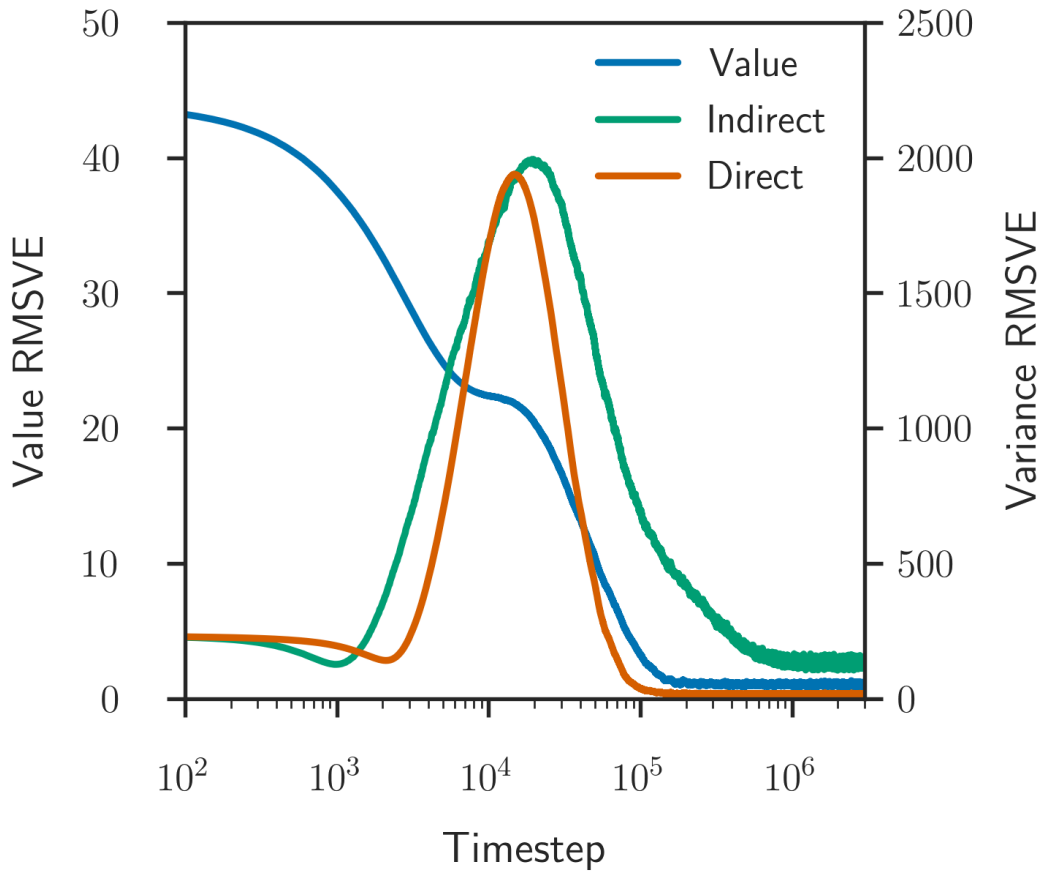
We use the following domain previously used to evaluate the indirect method:



For each state s_i we use $\phi(s_i) = [1, i/30]^T$ for our value estimator and $\phi_2(s_i) = [1, i/30, (i/30)^2]^T$ for our variance estimator.

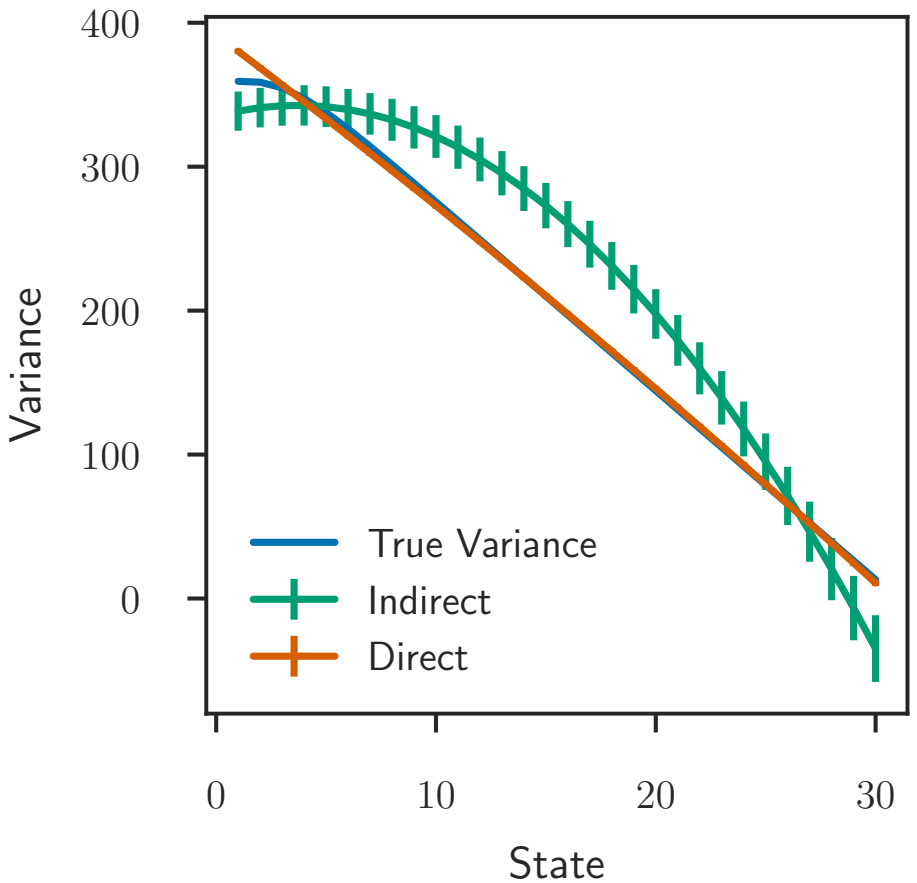
HOW DO THEY PERFORM ON THIS DOMAIN?

Here the direct method vastly outperforms the indirect method:



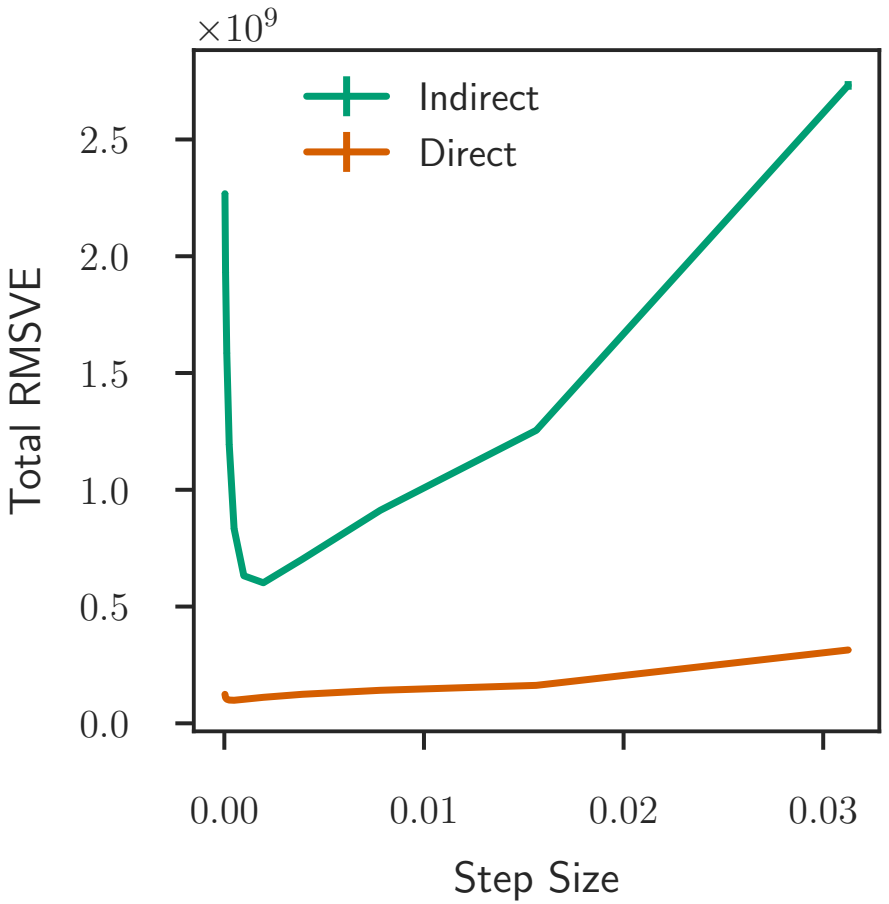
WHAT IS THE QUALITY OF THE SOLUTIONS REACHED?

The direct method reaches a much better fixed point and exhibits much less variance in its variance estimates:



WHAT IS THE PARAMETER SENSITIVITY?

In this domain, the direct method is much less sensitive to the choice of step sizes than the indirect method:



SUMMARY

We have described a method of directly estimating the variance of the return using temporal-difference methods. We have argued that learning the variance of the return can

- tell us interesting information about our domain,
- tell us useful information about the distribution of our return, and
- can be used to learn how to learn.

SUMMARY (CONTINUED)

We have furthermore shown evidence that the direct method:

- learns just as fast and occasionally faster than the indirect method,
- is more robust to inconsistencies in the value and variance learner, and
- exhibits substantially better performance under linear function approximation.

Questions?