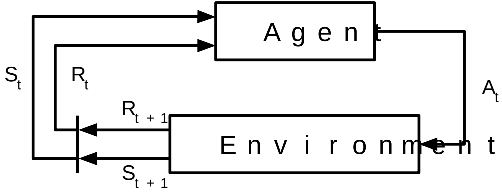
*   *

Reinforcement Learning and Artificial Intelligence La

Reinforcement learning considers an a
environment:



The function the agent uses to pick act
policy. Often the challenge is to find a

In reinforcement learning the return

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

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

Temporal-difference (TD) methods hav
tackling reinforcement learning prob
predictions to update predictions.

One of the most straightforward TD met (λ):

$$\delta_t = R_{t+1} + \gamma_{t+1} w_t^T x_{t+1} - w_t^T x_t$$

$$z_t = \gamma_t \lambda_t z_{t-1} + x_t$$

$$w_{t+1} = w_t + \alpha_{t+1} \delta_t z_t$$

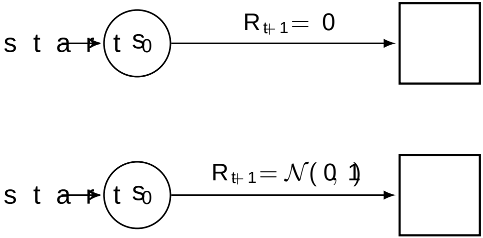
Recall what the return is:

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

We're not limited to learning only it'

learn more parts of its distributions variance

The variance might tell us things abou
expected value can't. Sometimes interesting. For e
example it could differentiate these



The variance can give useful information about the
can tell us how risky an action is to ta

Humans take risk into decisions and do
that maximizes the expected value.

We can use an estimate of the variance. example here is an algorithm that uses 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

$\text{errsq} = (\bar{g}_{t+1} - \mathbf{x}_{t+1}^\top \mathbf{w}_t)^2$

$\text{varg} = \max(0, \mathbf{x}_{t+1}^\top \mathbf{w}^{\text{sq}} - (\bar{g}_{t+1})^2)$

$\lambda_{t+1} = \text{errsq}/(\text{varg} + \text{errsq})$

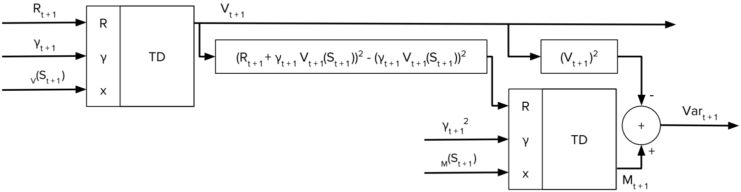
return λ_{t+1}

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 $\mathbb{E}_\pi[G_t | S_t = s]$ then we can just learn $\mathbb{E}_\pi[G_t^2 | S_t = s]$ on the side and use both our est estimate $\mathrm{Var}_\pi(G_t | S_t = s)$.

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



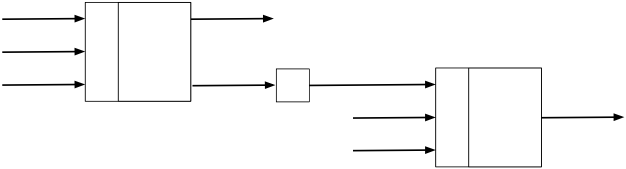
We can also use this identity:

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

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

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

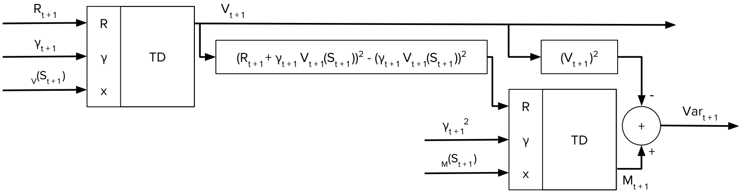
Using the identity $\mathrm{Var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$ one can estimate variance using the following structur

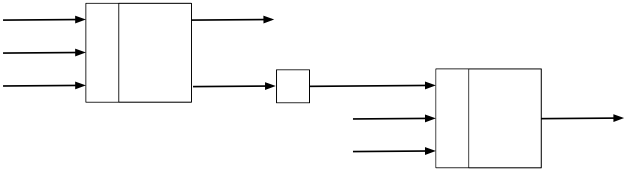


Using TD($\bar{\lambda}$) with $\bar{w}$ as the parameter vector for e
variance we obtain the following upda

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

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

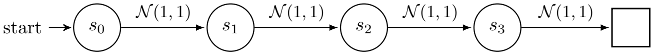

R t+1
γ t+1
V(S t+1)
R
γ
x
TD
V t+1
(R t+1 + γ t+1 V t+1(S t+1))2 - (γ t+1 V t+1(S t+1))2
(V t+1)2
-
+
+
Var t+1
γ t+1 2
M(S t+1)
R
γ
x
TD
M t+1



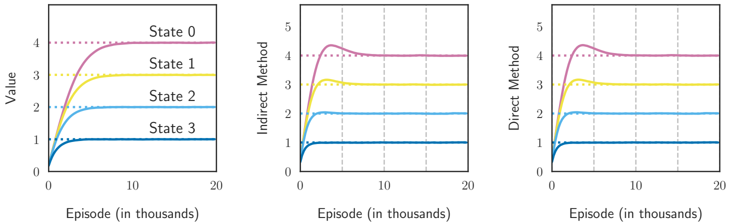
Ideally we want to know if the direct me

- is faster or slower to converge than
- is more robust or less robust to diff variance learner, and
- performs better or worse under line

We begin by comparing them on the follo
with gaussian rewards:

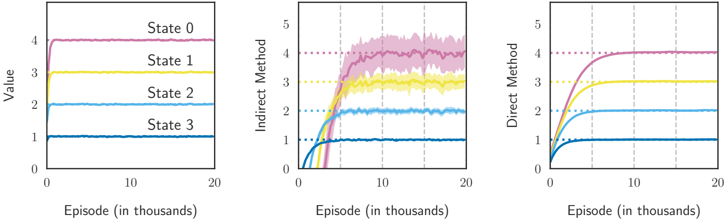


When $\alpha = \overline{\alpha} = 0.001$ both perform roughly the s



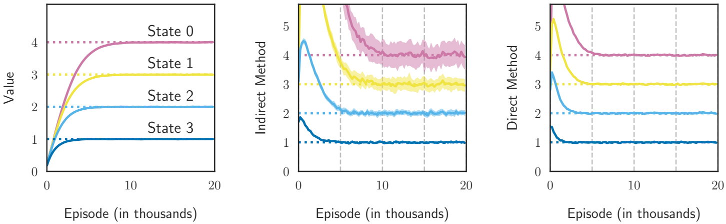
$\alpha > \overline{\alpha}$

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

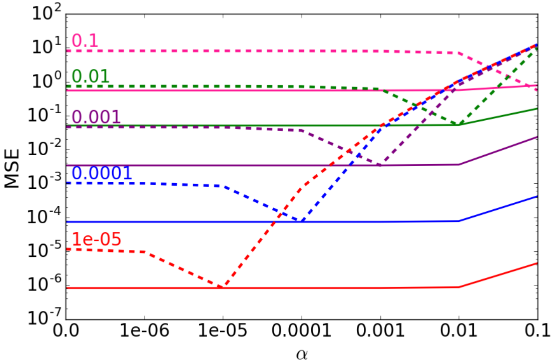


$\alpha < \overline{\alpha}$

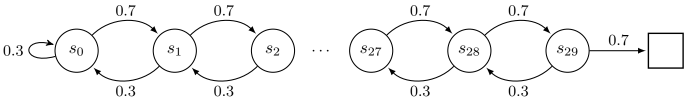
When $\alpha = 0.001$ and $\overline{\alpha} = 0.01$ the variance of the in
higher and the direct method is more st



In this domain we only see the two perfo
sizes are equal (note that the dotted l
method and the solid line represents t

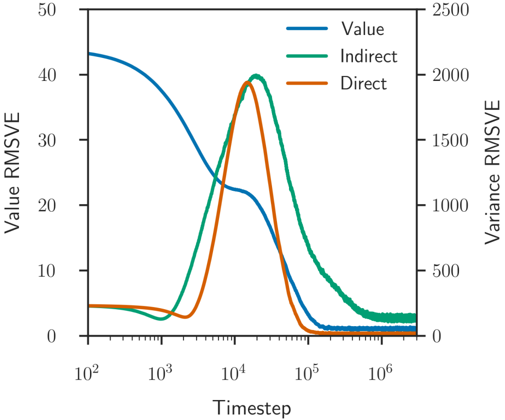


We use the following domain previousl
indirect method:

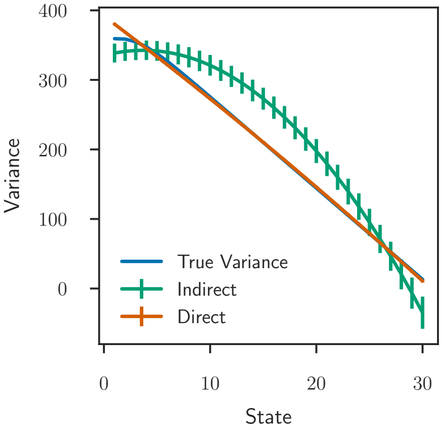


For each s_i we use $\phi(s_i) = [1, i/30]^T$ for our value est
and $\phi_2(s_i) = [1, i/30, (i/30)^2]^T$ for our variance estim

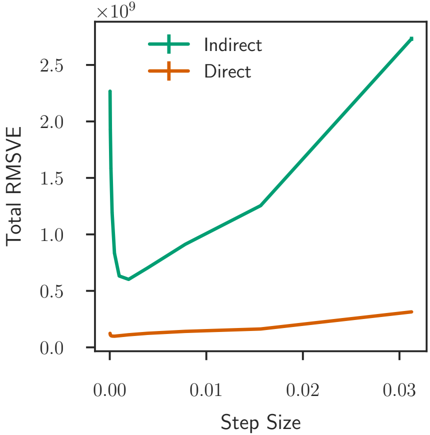
Here the direct method vastly outperf



The direct method reaches a much bette
much less variance in its variance est



In this domain, the direct method is mu
choice of step sizes than the indirect



We have described a method of directly
the return using temporal-difference
learning the variance of the return ca

- tell us interesting information about our doma
- tell us useful information about the distri
  and
- can be used to learn how to learn.

We have furthermore shown evidence th

- learns just as fast faster than the conventional l method,
- is more robust to inconsistencies in the va learner, and
- exhibits substantially better performance under li approximation.