TDγ

George Konidaris
gdk@csail.mit.edu

(joint work with Scott Niekum and Phil Thomas)

CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Learning and Intelligent Systems

Introduction

TD(λ): Dominant family of RL algorithms.

- Parameter λ used to mix unbiased, high-variance estimates with biased, low-variance estimates.
- λ must be set manually.
- Up until now, we did not understand what it really does.

This talk:

- Expose assumptions underlying TD(λ).
- Show that they are wrong!
- Propose another - parameter free - method, TDγ.

TD(λ)

Weighted sum:

$$R^{(1)} = r_0 + \gamma V(s_1) \qquad 1$$
$$R^{(2)} = r_0 + \gamma r_1 + \gamma^2 V(s_2) \qquad \lambda$$
$$\vdots$$
$$R^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_i + \gamma^n V(s_n) \qquad \lambda^n$$

weights

Estimator:

$$R^{\lambda}_{s_t} = (1-\lambda) \sum_{n=0}^{\infty} \lambda^n R^{(n+1)}_{s_t}$$

TD(λ)

This is called the λ-return.

- At λ=0 we get TD, at λ=1 we get MC.
- Intermediate values of λ usually best.

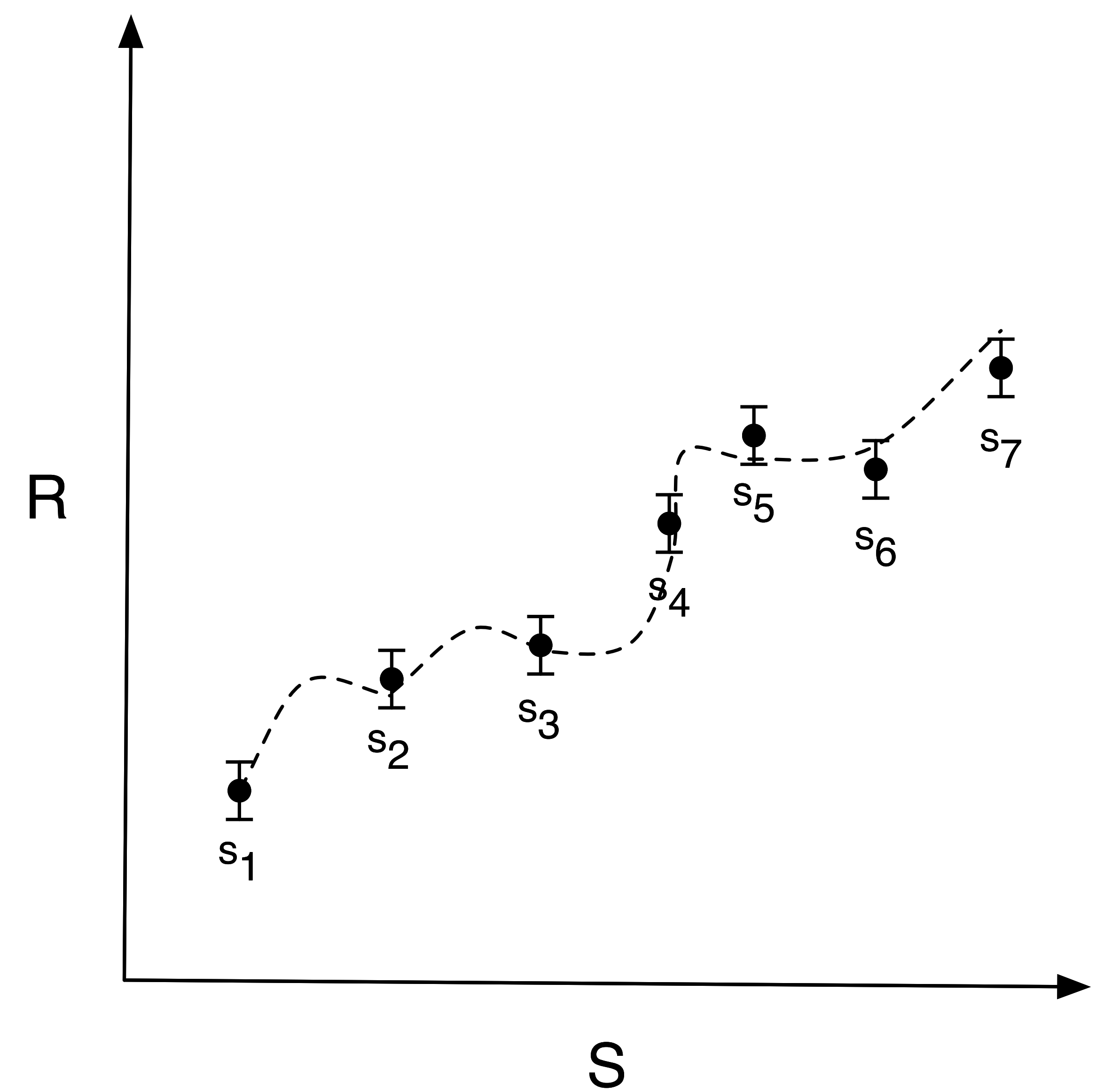
Results in a *family* of algorithms.

- Update rules via error metric using λ-return.
- Used almost exclusively, unchanged, since 1988.
- Original paper has ~3000 citations.

TD(λ)

- What are the implicit assumptions that lead to this estimator?

Linear Least-Squares



mean

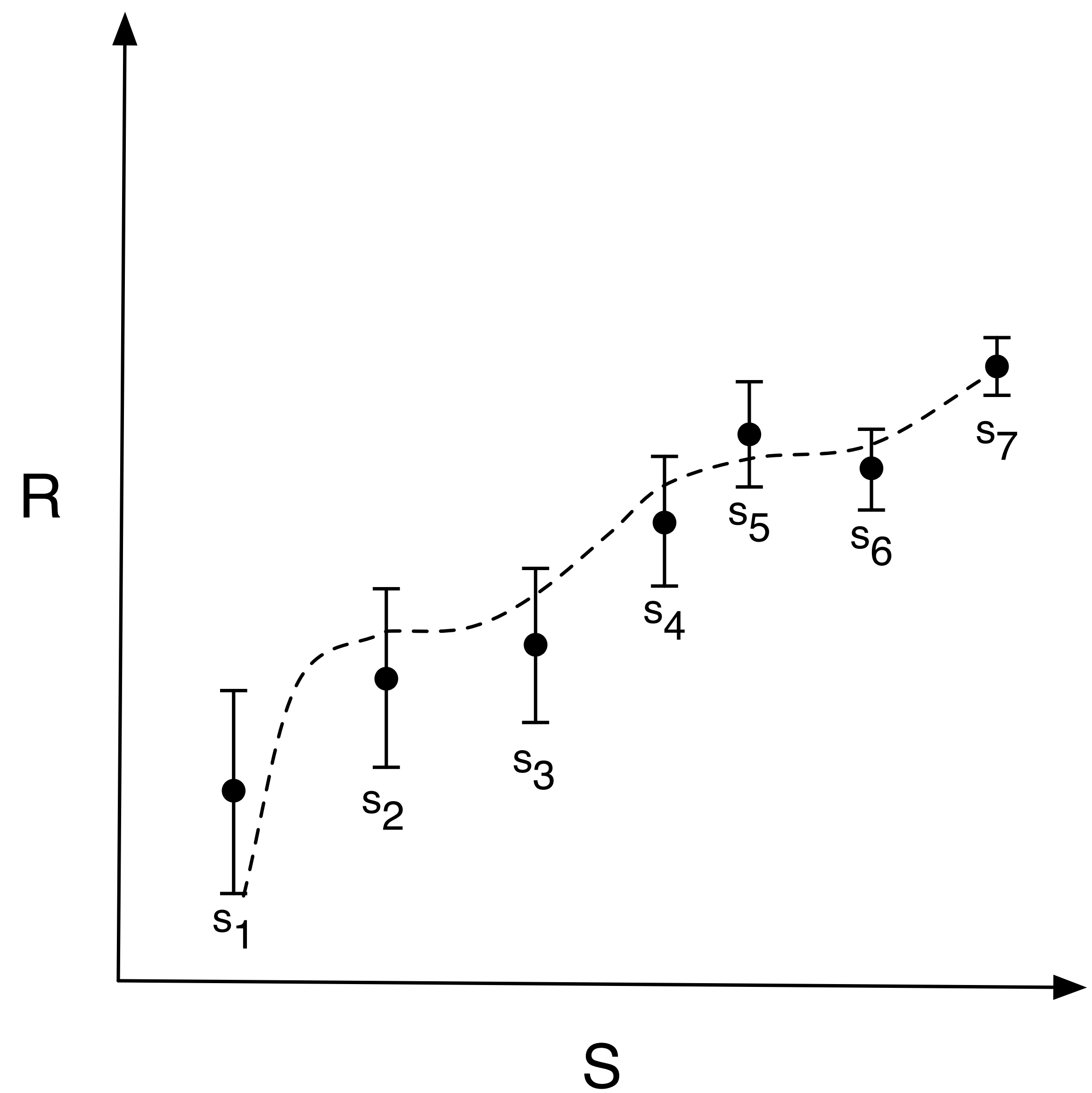
$\mathbf{w} \cdot \Phi_i(s)$

variance

β^{-1}

$$e_i = \sum_{j=1}^{m} [\mathbf{w} \cdot \Phi_i(\mathbf{s}_j) - R_j]^2$$

Weighted Linear Least-Squares



mean

$\mathbf{w} \cdot \Phi_i(s)$

variance

$(\rho^{m-i}\beta)^{-1}$

$$e_i = \sum_{j=1}^{m} \rho^{(m-j)} [\mathbf{w} \cdot \Phi_i(\mathbf{s}_j) - R_j]^2$$

TD(λ)

Consider the following assumptions:

- Each *n*-step rollout is independent.
- Each *n*-step rollout is normally distributed with mean of the true return.
- Variance of n-step rollout is $k(n)$.

TD(λ)

Likelihood:

$$\mathcal{L}(\hat{R}_{s_t} | R_{s_t}^{(1)}, \ldots, R_{s_t}^{(n)}; k) = \prod_{n=1}^{L} \mathcal{N}(R_{s_t}^{(n)} | \hat{R}_{s_t}, k(n))$$

Maximizing the log likelihood:

$$\hat{R}_{s_t} = \frac{\sum_{n=1}^{L} k(n)^{-1} R_{s_t}^{(n)}}{\sum_{n=1}^{L} k(n)^{-1}}$$

this is the λ-return, where:

$$k(n) = \lambda^{-n}$$
$$L \to \infty$$

TD(λ)

Therefore, λ-return is the estimator you get given three assumptions:

- Normal distribution of return estimates.
- Independence of rollouts.
- Variance of rollouts increases geometrically with common ratio $1/\lambda$.

All three of these assumptions are false.

On Rollout Variance

Let's let the first two slide, and consider the variance of an *n*-step sample return:

$$
\begin{aligned}
Var\left[R_{s_t}^{(n)}\right] =& Var\left[R_{s_t}^{(n-1)} - \gamma^{n-1}V(s_{t+n-1}) + \gamma^{n-1}r_{t+n-1} + \gamma^{n}V(s_{t+n})\right] \\
=& Var\left[R_{s_t}^{(n-1)}\right] + \gamma^{2(n-1)}Var\Big[V(s_{t+n-1}) - (r_{t+n-1} + \gamma V(s_{t+n}))\Big] \\
&+ 2Cov\left[R_{s_t}^{(n-1)}, -\gamma^{n-1}V(s_{t+n-1}) + \gamma^{n-1}r_{t+n-1} + \gamma^{n}V(s_{t+n})\right].
\end{aligned}
$$

First thing to notice:

- Variance increases from *n-1* to *n* **additively**.
- We assume the covariance away.

New Variance Model

We obtain:

TD error at step *n*

$$Var\left[R_{s_t}^{(n)}\right] \approx Var\left[R_{s_t}^{(n-1)}\right] + \gamma^{2(n-1)} Var\left[V(s_{t+n-1}) - (r_{t+n-1} + \gamma V(s_{t+n}))\right]$$

We assume the TD error variance is the same everywhere, set its value to κ.

Simple model of the variance of an *n*-step sample of return:

$$k(n) = \sum_{i=1}^{n} \gamma^{2(i-1)} \kappa.$$

TDγ

Resulting estimator:

$$R_{s_t}^{\gamma} = \frac{\cancel{\kappa^{-1}} \sum_{n=1}^{L} (\sum_{i=1}^{n} \gamma^{2(i-1)})^{-1} R_{s_t}^{(n)}}{\cancel{\kappa^{-1}} \sum_{n=1}^{L} (\sum_{i=1}^{n} \gamma^{2(i-1)})^{-1}} = \sum_{n=1}^{L} w(n, L) R_{s_t}^{(n)}$$

where

$$w(n, L) = \frac{(\sum_{i=1}^{n} \gamma^{2(i-1)})^{-1}}{\sum_{n=1}^{L} (\sum_{i=1}^{n} \gamma^{2(i-1)})^{-1}}$$

Parameter-free!

TDγ

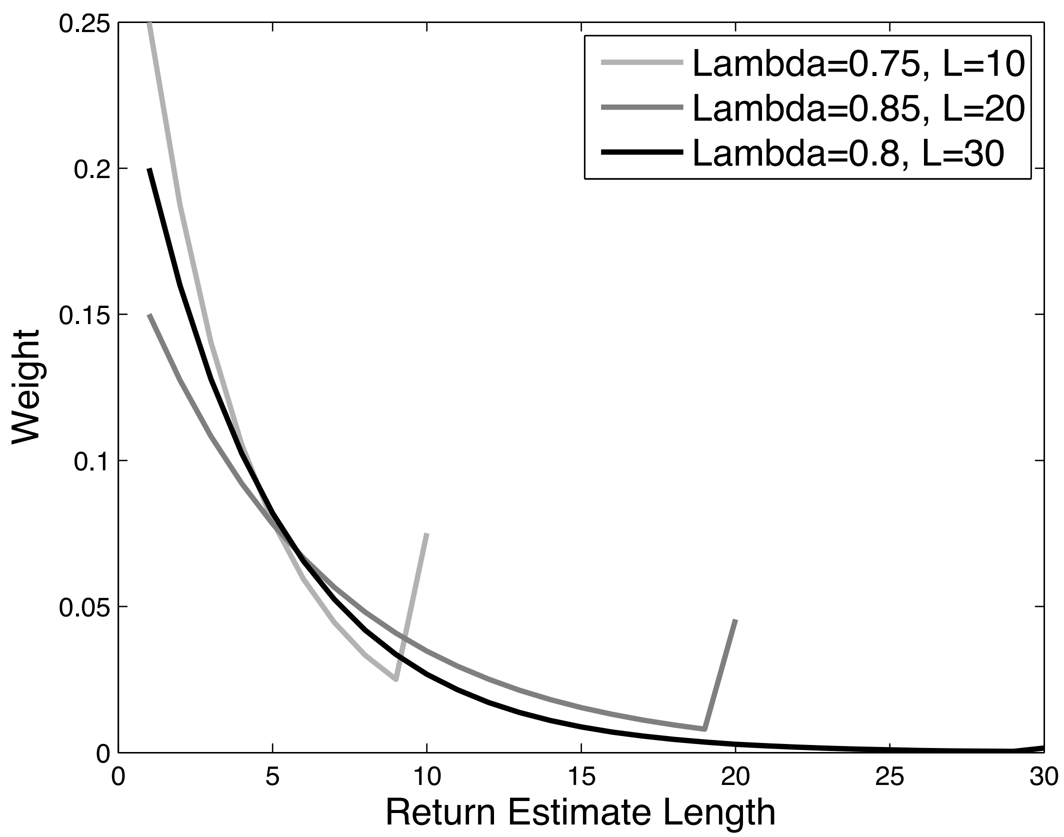
Weighted sum: **weights**

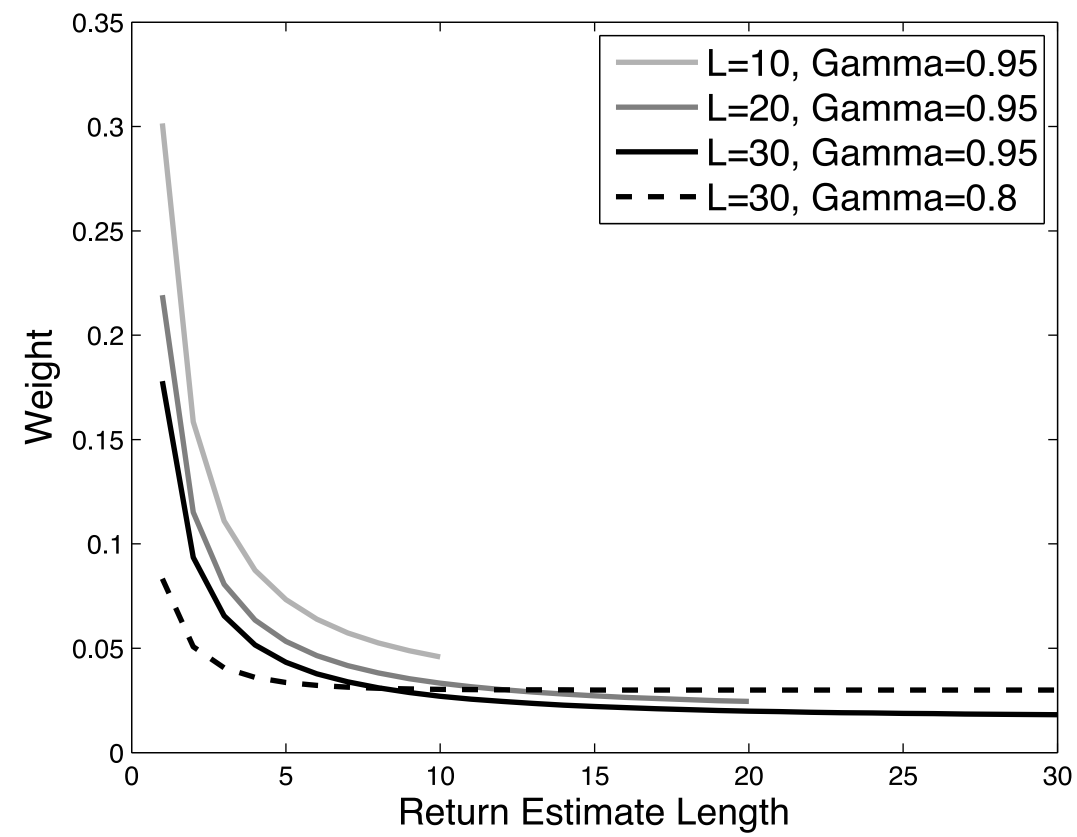
$R^{(1)} = r_0 + \gamma V(s_1)$	$\frac{1}{1}$
$R^{(2)} = r_0 + \gamma r_1 + \gamma^2 V(s_2)$	$\frac{1}{1+\gamma^2}$
$R^{(3)} = r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 V(s_3)$	$\frac{1}{1+\gamma^2+\gamma^4}$
$\vdots$	
$R^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_i + \gamma^n V(s_n)$	$\frac{1}{1+\gamma^2+\gamma^4+\ldots\gamma^{2n}}$

CSAIL

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

TDγ vs. TD(λ)





CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

The Catch

$$w(n, L) = \frac{\left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}{\sum_{n=1}^{L} \left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}$$

Normalizing constant is a function of episode length.
Differs for each state.

λ-return avoids this because it assumes episode is infinitely long, and sum of weights tends to a constant.

$$R_{s_t}^{\lambda} = (1 - \lambda) \sum_{n=0}^{\infty} \lambda^n R_{s_t}^{(n+1)}$$

So:

TDγ kills λ and replaces TD(λ) if:

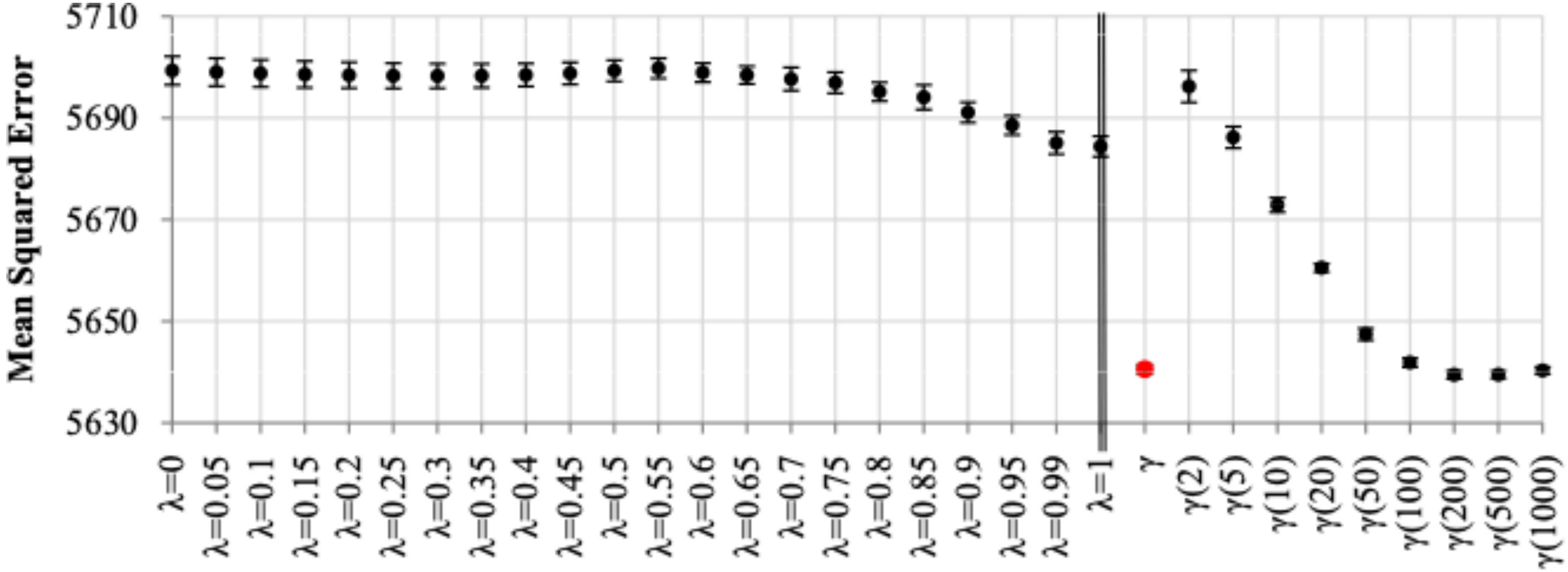
- We can be incremental episode-wise.
- We can process in a batch.

But not if we must be incremental transition-wise.

- We impose capacity *C*.
- Use the first *C* rollouts.
- Normalizer not known, except for last *C-1* steps.

Results

Acrobot



Similar for another 4 domains.

- TDγ beats TD(λ) for any value of λ (4/5)
- Intermediate values of *C* do very well.

Future Work

Better model of the variance.
Model that affords a completely incremental implementation.

Other members of the TDγ family:

- LSTDγ
- Sarsa$_\gamma$
- GQγ

Account for covariance: TD(Omega)