High-dimensional continuous control using generalized advantage estimation

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan and Pieter Abbeel
Department of Electrical Engineering and Computer Science
University of California, Berkeley
{joschu,pcmoritz,levine,jordan,pabbeel}@eecs.berkeley.edu

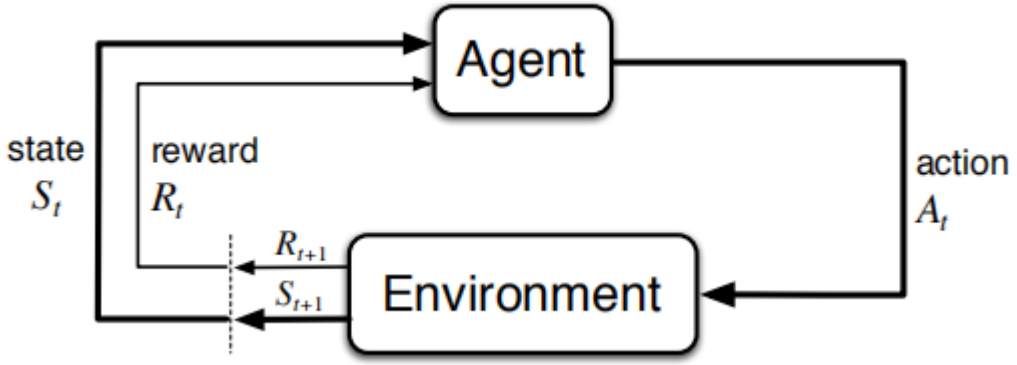
ICLR 2016

Preliminaries

- Policy $\pi_\theta(a_t \mid s_t)$
- Objective $\pi^* = \arg\max_\pi E_{\tau\sim\pi}[R(\tau)]$

$$\tau = s_1, a_1, r_1, s_1, a_2 \ldots \qquad r_t = r(s_t, a_t, s_{t+1})$$

$$R(\tau) = \sum_i r_i$$



- Optimization

$$J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[R(\tau)] = \int \pi_\theta(\tau) R(\tau)\, \mathrm{d}\tau$$

$$\nabla_\theta J(\theta) = \int \nabla_\theta \pi_\theta(\tau) \cdot R(\tau)\, \mathrm{d}\tau = \int R(\tau) \frac{\nabla_\theta \pi_\theta(\tau)}{\pi_\theta(\tau)} \pi_\theta(\tau)\, \mathrm{d}\tau$$

$$= \int R(\tau) \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau)\, \mathrm{d}\tau = E_{\tau\sim\pi_\theta(\tau)}[R(\tau) \nabla_\theta \log \pi_\theta(\tau)]$$

Preliminaries

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[R(\tau)\nabla_\theta \log \pi_\theta(\tau)]$$

Optimization

$$\pi_\theta(\tau) = P_0(s_1)\pi(a_1|s_1,\theta)P(s_2|s_1,a_1)\pi(a_2|s_2,\theta).....$$

$$\log \pi_\theta(\tau) = \log P_0(s_1) + \sum_t [\log \pi(a_t|s_t,\theta) + \log P(s_{t+1}|s_t,a_t)]$$

$$\nabla_\theta \log \pi_\theta(\tau) = \sum_t \nabla_\theta \log \pi(a_t|s_t,\theta)$$

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[R(\tau)\cdot\sum_t \nabla_\theta \log \pi(a_t|s_t,\theta)]$$

$$= E_{\tau\sim\pi_\theta(\tau)}[\sum_t r(s_t,a_t)\cdot\sum_t \nabla_\theta \log \pi(a_t|s_t,\theta)]$$

Preliminaries

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[R(\tau)\nabla_\theta \log \pi_\theta(\tau)]$$

- REINFROCE algorithm

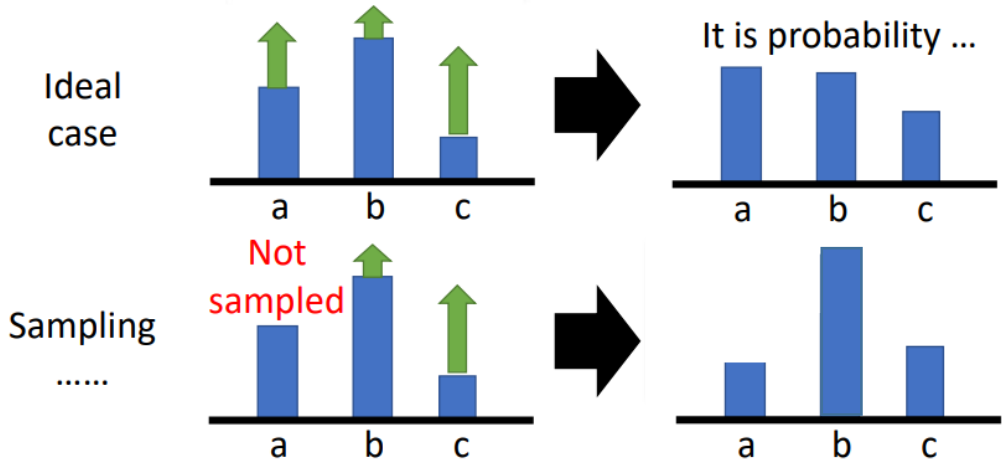
$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[\sum_t r(s_t, a_t) \cdot \sum_t \nabla_\theta \log \pi(a_t|s_t, \theta)]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}[\sum_t r(s_t^i, a_t^i) \cdot \sum_t \nabla_\theta \log \pi(a_t^i|s_t^i, \theta)]$$

MC

REINFORCE algorithm:

1. sample $\{\tau^i\}$ from $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \sum_i \left(\sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i)\right)\left(\sum_t r(\mathbf{s}_t^i, \mathbf{a}_t^i)\right)$
3. $\theta \leftarrow \theta + \alpha\nabla_\theta J(\theta)$



Preliminaries

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[\sum_t r(s_t, a_t) \cdot \sum_t \nabla_\theta \log \pi(a_t|s_t, \theta)]$$

Policy Gradient

$$g = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right], \tag{1}$$

where Ψ_t may be one of the following:

1. $\sum_{t=0}^{\infty} r_t$: total reward of the trajectory.
2. $\sum_{t'=t}^{\infty} r_{t'}$: reward following action a_t.
3. $\sum_{t'=t}^{\infty} r_{t'} - b(s_t)$: baselined version of previous formula.
4. $Q^\pi(s_t, a_t)$: state-action value function.
5. $A^\pi(s_t, a_t)$: advantage function.
6. $r_t + V^\pi(s_{t+1}) - V^\pi(s_t)$: TD residual.

The latter formulas use the definitions

$$V^\pi(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \qquad Q^\pi(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \tag{2}$$

$$A^\pi(s_t, a_t) := Q^\pi(s_t, a_t) - V^\pi(s_t), \quad \text{(Advantage function)}. \tag{3}$$

Preliminaries

Before proceeding, we will introduce the notion of a γ-just estimator of the advantage function, which is an estimator that does not introduce bias when we use it in place of $A^{\pi,\gamma}$ (which is not known and must be estimated) in Equation (6) to estimate g^γ.[1] Consider an advantage estimator $\hat{A}_t(s_{0:\infty}, a_{0:\infty})$, which may in general be a function of the entire trajectory.

Variance Reduction Parameter γ

$$V^{\pi,\gamma}(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty}\gamma^l r_{t+l}\right] \qquad Q^{\pi,\gamma}(s_t,a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty}\gamma^l r_{t+l}\right] \tag{4}$$

$$A^{\pi,\gamma}(s_t,a_t) := Q^{\pi,\gamma}(s_t,a_t) - V^{\pi,\gamma}(s_t). \tag{5}$$

The discounted approximation to the policy gradient is defined as follows:

$$g^\gamma := \mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[\sum_{t=0}^{\infty} A^{\pi,\gamma}(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right]. \tag{6}$$

Definition 1. *The estimator $\hat{A}_t$ is γ-just if*

$$\mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[\hat{A}_t(s_{0:\infty},a_{0:\infty})\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] = \mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[A^{\pi,\gamma}(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right]. \tag{7}$$

It follows immediately that if $\hat{A}_t$ is γ-just for all t, then

$$\mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[\sum_{t=0}^{\infty}\hat{A}_t(s_{0:\infty},a_{0:\infty})\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] = g^\gamma \tag{8}$$

Advantage Function Estimation

- $\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$

V : an approximate value function

$$\hat{g} = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{\infty} \hat{A}_t^n \nabla_\theta \log \pi_\theta(a_t^n \mid s_t^n)$$

n : indexes over a batch of episodes

If $V = V^{\pi,\gamma}$, then δ_t^V is a γ-just advantage estimator, and in fact, an unbiased estimator of $A^{\pi,\gamma}$:

$$\begin{aligned}\mathbb{E}_{s_{t+1}}\left[\delta_t^{V^{\pi,\gamma}}\right] &= \mathbb{E}_{s_{t+1}}\left[r_t + \gamma V^{\pi,\gamma}(s_{t+1}) - V^{\pi,\gamma}(s_t)\right] \\ &= \mathbb{E}_{s_{t+1}}\left[Q^{\pi,\gamma}(s_t, a_t) - V^{\pi,\gamma}(s_t)\right] = A^{\pi,\gamma}(s_t, a_t).\end{aligned}$$

otherwise it will yield biased policy gradient estimates.

Advantage Function Estimation

$$\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$$

- $\hat{A}_t^{(k)}$

$$\hat{A}_t^{(1)} := \delta_t^V = -V(s_t) + r_t + \gamma V(s_{t+1}) \tag{11}$$

$$\hat{A}_t^{(2)} := \delta_t^V + \gamma\delta_{t+1}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2}) \tag{12}$$

$$\hat{A}_t^{(3)} := \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 V(s_{t+3}) \tag{13}$$

$$\hat{A}_t^{(k)} := \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}^V = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{k-1} r_{t+k-1} + \gamma^k V(s_{t+k}) \tag{14}$$

$$\downarrow \; k \to \infty$$

$$\hat{A}_t^{(\infty)} = \sum_{l=0}^{\infty} \gamma^l \delta_{t+l}^V = -V(s_t) + \sum_{l=0}^{\infty} \gamma^l r_{t+l}$$

Consider $\hat{A}_t^{(k)}$ to be an estimator of the advantage function, which is only γ-just when $V = V^{\pi,\gamma}$.

However, note that the bias generally becomes smaller as $k \to \infty$.

Advantage Function Estimation

Generalized Advantage Estimator

$$
\begin{aligned}
\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} &:= (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \dots\right) \\
&= (1-\lambda)\left(\delta_t^V + \lambda(\delta_t^V + \gamma\delta_{t+1}^V) + \lambda^2(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V) + \dots\right) \\
&= (1-\lambda)\left(\delta_t^V(1 + \lambda + \lambda^2 + \dots) + \gamma\delta_{t+1}^V(\lambda + \lambda^2 + \lambda^3 + \dots)\right. \\
&\qquad \left. + \gamma^2\delta_{t+2}^V(\lambda^2 + \lambda^3 + \lambda^4 + \dots) + \dots\right) \\
&= (1-\lambda)\left(\delta_t^V\left(\frac{1}{1-\lambda}\right) + \gamma\delta_{t+1}^V\left(\frac{\lambda}{1-\lambda}\right) + \gamma^2\delta_{t+2}^V\left(\frac{\lambda^2}{1-\lambda}\right) + \dots\right) \\
&= \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V
\end{aligned}
$$

$$
\mathrm{GAE}(\gamma, 0): \quad \hat{A}_t := \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)
$$

→ biased, lower variance

$$
\mathrm{GAE}(\gamma, 1): \quad \hat{A}_t := \sum_{l=0}^{\infty} \gamma^l \delta_{t+l} = \sum_{l=0}^{\infty} \gamma^l r_{t+l} - V(s_t)
$$

→ high variance

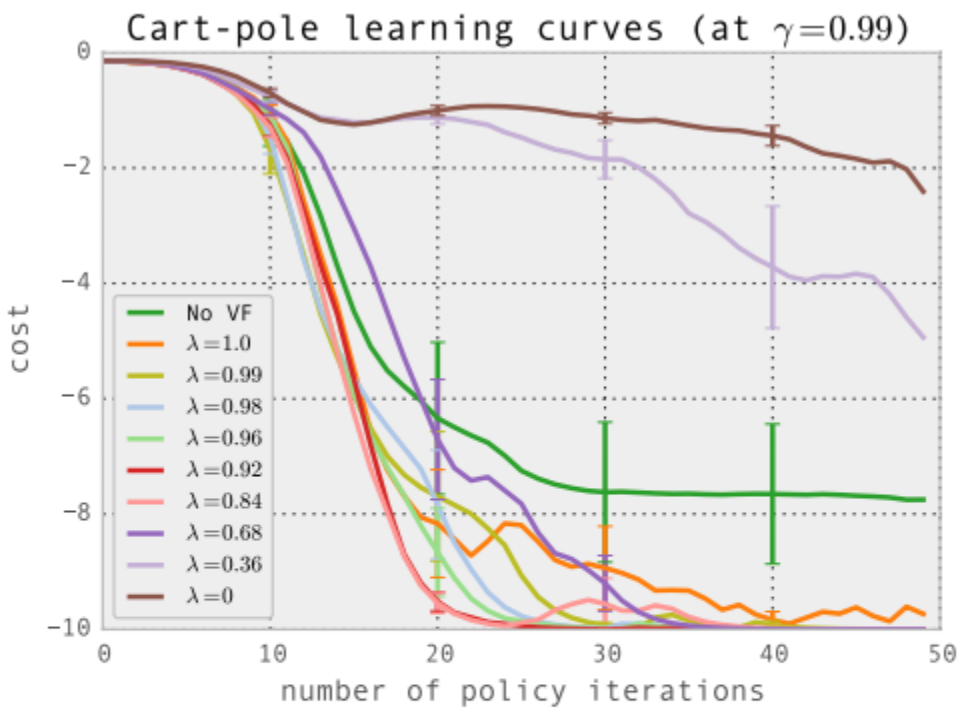
Advantage Function Estimation

- Generalized Advantage Estimator

$$g^{\gamma} \approx \mathbb{E}\left[\sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t \mid s_t) \hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)}\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t \mid s_t) \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^{V}\right]$$

Experiments

Cart-Pole



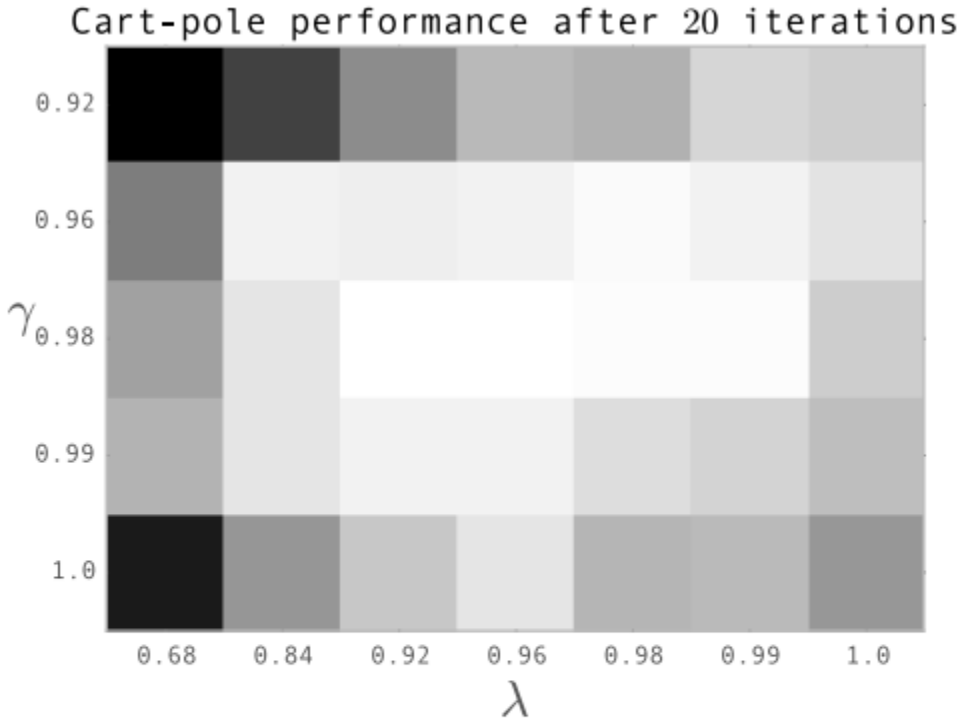


Figure 2: Left: learning curves for cart-pole task, using generalized advantage estimation with varying values of λ at $\gamma = 0.99$. The fastest policy improvement is obtain by intermediate values of λ in the range $[0.92, 0.98]$. Right: performance after 20 iterations of policy optimization, as γ and λ are varied. White means higher reward. The best results are obtained at intermediate values of both.

Experiments

- 3D biped locomotion & Quadruped locomotion

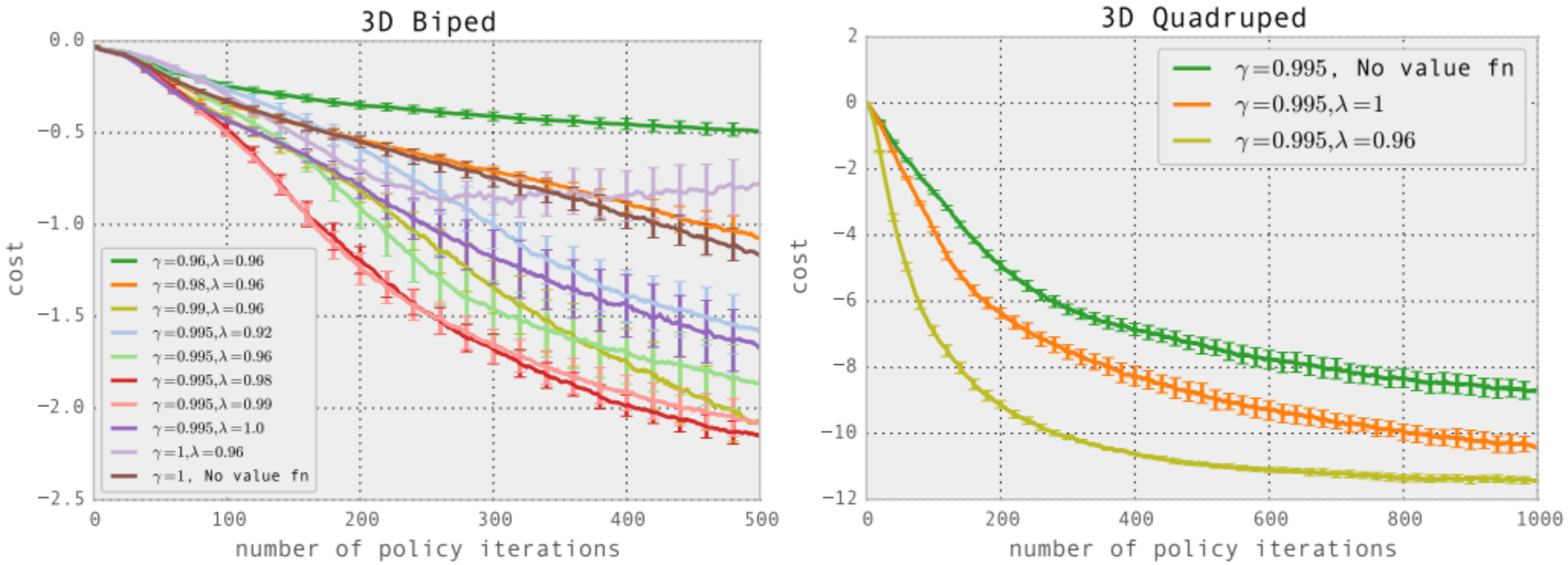
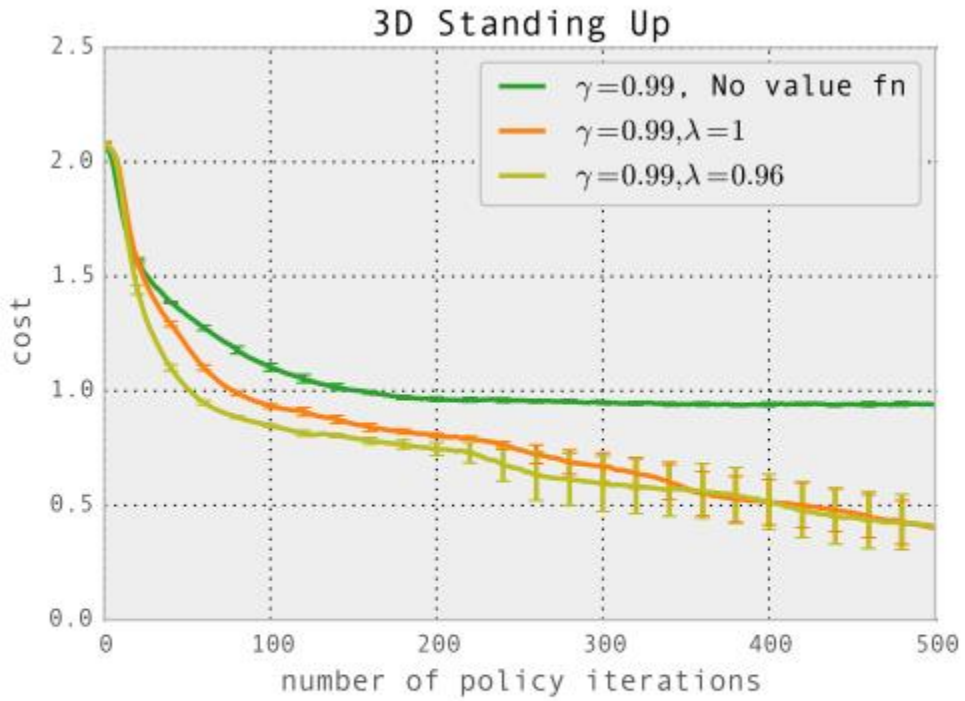


Figure 3: Left: Learning curves for 3D bipedal locomotion, averaged across nine runs of the algorithm. Right: learning curves for 3D quadrupedal locomotion, averaged across five runs.

Experiments

- Biped getting up



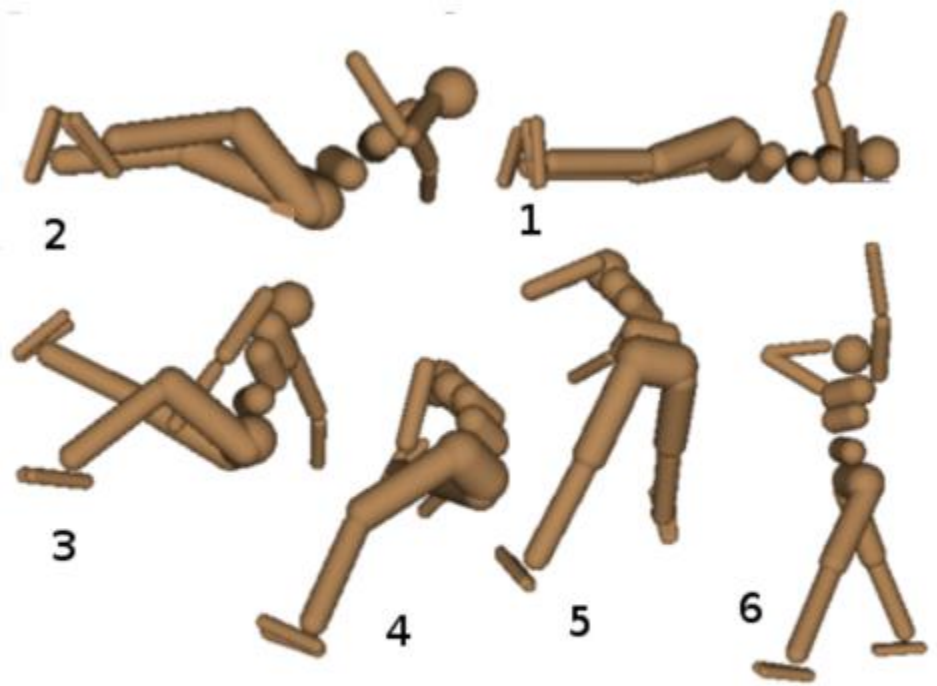


Figure 4: (a) Learning curve from quadrupedal walking, (b) learning curve for 3D standing up, (c) clips from 3D standing up.

Thanks