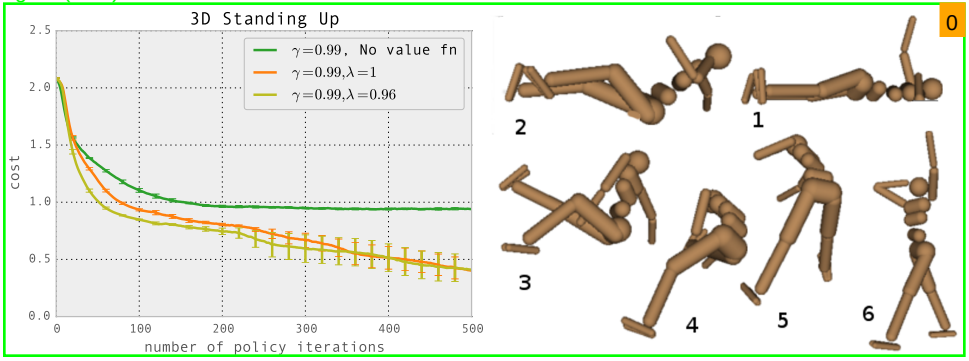
figure (0.97)



figure_caption (0.92)

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

title (0.89)

7 DISCUSSIO

plain_text (0.98)

Policy gradient methods provide a way to reduce reinforcement learning to stochastic gradient d scent, by providing unbiased gradient estimates. However, so far their success at solving difficult control problems has been limited, largely due to their high sample complexity. We have argued that the key to variance reduction is to obtain good estimates of the advantage function.

plain_text (0.98)

We have provided an intuitive but informal analysis of the problem of advantage function estimatio and justified the generalized advantage estimator, which has two parameters γ, λ which adjust the bias-variance tradeoff. We described how to combine this idea with trust region policy optimization and a trust region algorithm that optimizes a value function, both represented by neural networks. Combining these techniques, we are able to learn to solve difficult control tasks that have previously been out of reach for generic reinforcement learning methods.

plain_text (0.97)

Our main experimental validation of generalized advantage estimation is in the domain of simulat robotic locomotion. As shown in our experiments, choosing an appropriate intermediate value of λ in the range $[0.9, 0.99]$ usually results in the best performance. A possible topic for future work is how to adjust the estimator parameters γ, λ in an adaptive or automatic way.

plain_text (0.98)

One question that merits future investigation is the relationship between value function estimati error and policy gradient estimation error. If this relationship were known, we could choose an error metric for value function fitting that is well-matched to the quantity of interest, which is typically the accuracy of the policy gradient estimation. Some candidates for such an error metric might include the Bellman error or projected Bellman error, as described in Bhatnagar et al. (2009).

plain_text (0.97)

Another enticing possibility is to use a shared function approximation architecture for the policy a the value function, while optimizing the policy using generalized advantage estimation. While formulating this problem in a way that is suitable for numerical optimization and provides convergence guarantees remains an open question, such an approach could allow the value function and policy representations to share useful features of the input, resulting in even faster learning.

plain_text (0.98)

In concurrent work, researchers have been developing policy gradient methods that involve differe tiation with respect to the continuous-valued action (Lillicrap et al., 2015; Heess et al., 2015). While we found empirically that the one-step return ($\lambda = 0$) leads to excessive bias and poor performance, these papers show that such methods can work when tuned appropriately. However, note that those papers consider control problems with substantially lower-dimensional state and action spaces than the ones considered here. A comparison between both classes of approach would be useful for future work.

title (0.89)

ACKNOWLEDGEMEN

plain_text (0.97)

We thank Emo Todorov for providing the simulator as well as insightful discussions, and we th Greg Wayne, Yuval Tassa, Dave Silver, Carlos Florensa Campo, and Greg Brockman for insightful discussions. This research was funded in part by the Office of Naval Research through a Young