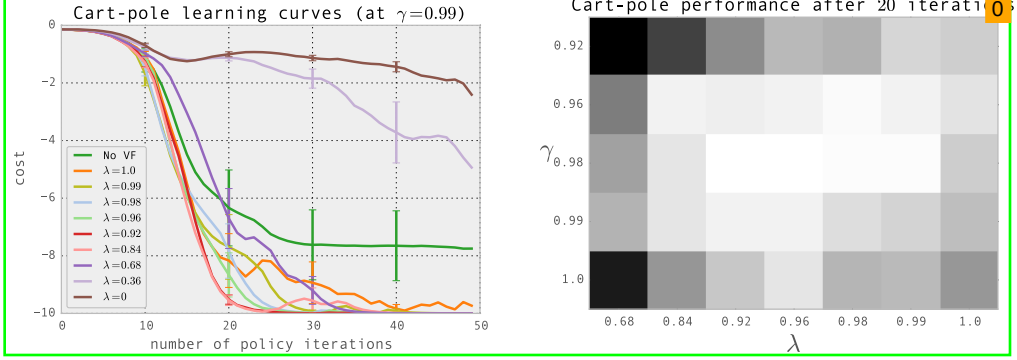
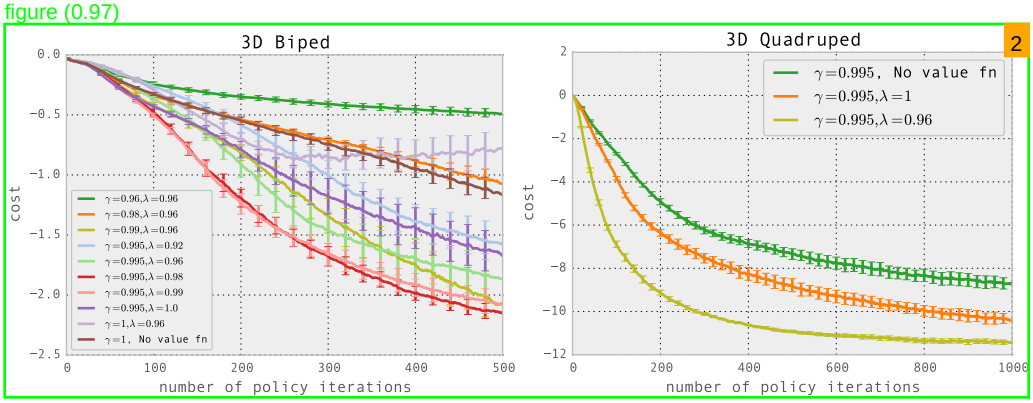
figure (0.53)



figure_caption (0.93)

Figure 2: Left: learning curves for cart-pole task, using generalized advantage estimation wi 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.

figure (0.97)



figure_caption (0.94)

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

title (0.88)

6.3.2 3D BIPEDAL LOCOMOTIO

plain_text (0.98)

Each trial took about 2 hours to run on a 16-core machine, where the simulation rollouts were para lelized, as were the function, gradient, and matrix-vector-product evaluations used when optimizing the policy and value function. Here, the results are averaged across 9 trials with different random seeds. The best performance is again obtained using intermediate values of $\gamma \in [0.99, 0.995], \lambda \in [0.96, 0.99]$. The result after 1000 iterations is a fast, smooth, and stable gait that is effectively completely stable. We can compute how much "real time" was used for this learning process: $0.01\,^{\text{seconds}}/_{\text{timestep}} \times 50000\,^{\text{timesteps}}/_{\text{batch}} \times 1000\,\text{batches}/3600 \cdot 24\,^{\text{seconds}}/_{\text{day}} = 5.8\,\text{days}$. Hence, it is plausible that this algorithm could be run on a real robot, or multiple real robots learning in parallel, if there were a way to reset the state of the robot and ensure that it doesn't damage itself.

title (0.91)

6.3.3 OTHER 3D ROBOT TASI

plain_text (0.98)

The other two motor behaviors considered are quadrupedal locomotion and getting up off the grou for the 3D biped. Again, we performed 5 trials per experimental condition, with different random seeds (and initializations). The experiments took about 4 hours per trial on a 32-core machine. We performed a more limited comparison on these domains (due to the substantial computational resources required to run these experiments), fixing $\gamma = 0.995$ but varying $\lambda = \{0, 0.96\}$, as well as an experimental condition with no value function. For quadrupedal locomotion, the best results are obtained using a value function with $\lambda = 0.96$ Section 6.3.2. For 3D standing, the value function always helped, but the results are roughly the same for $\lambda = 0.96$ and $\lambda = 1$.