Investigator Award and under grant number N00014-11-1-0688, DARPA through a Young Facul
Award, by the Army Research Office through the MAST program.

A FREQUENTLY ASKED QUESTION

A.1 WHAT'S THE RELATIONSHIP WITH COMPATIBLE FEATURE

Compatible features are often mentioned in relation to policy gradient algorithms that make u of a value function, and the idea was proposed in the paper *On Actor-Critic Methods* by Konda & Tsitsiklis (2003). These authors pointed out that due to the limited representation power of the policy, the policy gradient only depends on a certain subspace of the space of advantage functions. This subspace is spanned by the compatible features $\nabla_{\theta_i} \log \pi_\theta(a_t|s_t)$, where $i \in \{1, 2, \ldots, \dim \theta\}$. This theory of compatible features provides no guidance on how to exploit the temporal structure of the problem to obtain better estimates of the advantage function, making it mostly orthogonal to the ideas in this paper.

The idea of compatible features motivates an elegant method for computing the natural policy gra ent (Kakade, 2001a; Peters & Schaal, 2008). Given an empirical estimate of the advantage function $\hat{A}_t$ at each timestep, we can project it onto the subspace of compatible features by solving the following least squares problem:

$$\underset{\mathbf{r}}{\text{minimize}} \sum_t \|\mathbf{r} \cdot \nabla_\theta \log \pi_\theta(a_t \mid s_t) - \hat{A}_t\| \tag{32}$$

If $\hat{A}$ is γ-just, the least squares solution is the natural policy gradient (Kakade, 2001a). Note th any estimator of the advantage function can be substituted into this formula, including the ones we derive in this paper. For our experiments, we also compute natural policy gradient steps, but we use the more computationally efficient numerical procedure from Schulman et al. (2015), as discussed in Section 6.

A.2 WHY DON'T YOU JUST USE A Q-FUNCTIO

Previous actor critic methods, e.g. in Konda & Tsitsiklis (2003), use a Q-function to obtain pote tially low-variance policy gradient estimates. Recent papers, including Heess et al. (2015); Lillicrap et al. (2015), have shown that a neural network Q-function approximator can used effectively in a policy gradient method. However, there are several advantages to using a state-value function in the manner of this paper. First, the state-value function has a lower-dimensional input and is thus easier to learn than a state-action value function. Second, the method of this paper allows us to smoothly interpolate between the high-bias estimator ($\lambda = 0$) and the low-bias estimator ($\lambda = 1$). On the other hand, using a parameterized Q-function only allows us to use a high-bias estimator. We have found that the bias is prohibitively large when using a one-step estimate of the returns, i.e., the $\lambda = 0$ estimator, $\hat{A}_t = \delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$. We expect that similar difficulty would be encountered when using an advantage estimator involving a parameterized Q-function, $\hat{A}_t = Q(s, a) - V(s)$. There is an interesting space of possible algorithms that would use a parameterized Q-function and attempt to reduce bias, however, an exploration of these possibilities is beyond the scope of this work.

B PROOI

Proof of Proposition 1: First we can split the expectation into terms involving Q an

$$\begin{aligned}
&\mathbb{E}_{s_{0:\infty}, a_{0:\infty}} \left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)(Q_t(s_{0:\infty}, a_{0:\infty}) - b_t(s_{0:t}, a_{0:t-1}\right. \\
&\quad = \mathbb{E}_{s_{0:\infty}, a_{0:\infty}} \left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)(Q_t(s_{0:\infty}, a_{0:\infty}))\right] \\
&\quad - \mathbb{E}_{s_{0:\infty}, a_{0:\infty}} \left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)(b_t(s_{0:t}, a_{0:t-1}))\right]
\end{aligned} \tag{33}$$