A Architecture Details

The policy network, the intrinsic reward network, the value networks are all 2-layer MLP with 128-hidden-units each. Using a network size of 256 did not make a significant difference. The last layer of intrinsic reward network is simply a linear transformation with no bias. And the 128-dim vector is used as η. CLN is applied right before the non-linear activation of the policy net's last layer, after which a gaussian layer is constructed. For most tasks considered, we found it helpful to do a tanh squashing on the action outputs, then multiplied by the action scale of the environment. We use the Xavier initializer with a normal distribution for all the weights. We use weight normalization on η by default, thus fixing the norm of η to be 1.

B Hyperparameters

The policy network has a linear decay schedule for the learning rate, which is initialized to be $5e-4$. We use 4 workers for Hopper, Walker, Ant and Humanoid. Using 8 workers for Ant and Humanoid while decreasing the horizon per worker speeds up training. Each worker collects 2200 timesteps, and during the inner loop, divides those timesteps into mini-batches of length 100. The intrinsic reward network has a learning rate of $1e-3$. For the number of iterations, $N_r = 5$, $N_p = 10$. The noise level σ is initialized to be 0.001 and then trained by back-propagating through CLN. The weighting of the policy gradient loss for intrinsic reward learning is $\lambda = 0.5$.