A. Hyperparameters

A.1. Common Parameters

We use the following parameters for LSP policies throughout the experiments. The algorithm uses a replay pool of one million samples, and the training is delayed until at least 1000 samples have been collected to the pool. Each training iteration consists of 1000 environments time steps, and all the networks (value functions, policy scale/translation, and observation embedding network) are trained at every time step. Every training batch has a size of 128. The value function networks and the embedding network are all neural networks comprised of two hidden layers, with 128 ReLU units at each hidden layer. The dimension of the observation embedding is equal to two times the number of action dimensions. The scale and translation neural networks used in the real NVP bijector both have one hidden layer consisting of number of ReLU units equal to the number of action dimensions. All the network parameters are updated using Adam optimizer with learning rate $3 \cdot 10^{-4}$.

Table 1 lists the common parameters used for the LSP-policy, and Table 2 lists the parameters that varied across the environments.

Table 1. Shared parameters for benchmark tasks

Parameter	Value
learning rate	$3 \cdot 10^{-4}$
batch size	128
discount	0.99
target smoothing coefficient	10^{-2}
maximum path length	10^3
replay pool size	10^6
hidden layers (Q, V, embedding)	2
hidden units per layer (Q, V, embedding)	128
policy coupling layers	2

Table 2. Environment specific parameters for benchmark tasks

Parameter	Swimmer (rllab)	Hopper-v1	Walker2d-v1	HalfCheetah-v1	Ant (rllab)	Humanoid (rllab)
action dimensions	2	3	6	6	8	21
reward scale	100	1	3	1	3	3
observation embedding dimension	4	6	12	12	16	42
scale/translation hidden units	2	3	6	6	8	21

A.2. High-Level Policies

All the low-level policies in hierarchical cases (Figures 5(b), 4(a), 4(b)) are trained using the same parameters used for the corresponding benchmark environment. All the high-level policies use Gaussian action prior. For the Ant maze task, the latent sample of the high-level policy is sampled once in the beginning of the rollout and kept fixed until the next one. The same high-level action is kept fixed over three environment steps. Otherwise, all the policy parameters for the high-level policies are equal to the benchmark parameters.

The environments used for training the low-level policies are otherwise equal to the benchmark environments, except for their reward function, which is modified to yield velocity based reward in any direction on the xy-plane, in contrast to just positive x-direction in the benchmark tasks. In the Ant maze environment, the agent receives a reward of 1000 upon reaching the goal and 0 otherwise. In particular, no velocity reward nor any control costs are awarded to the agent. The environment terminates after the agent reaches the goal.