A Appendix

A.1 Soft Stochastic Policy Gradient Theorem

To fit the new reward function definition, the following is the process of deriving the soft version of policy gradient. Let's first start with the derivative of the soft state value function. Note that

$$
\begin{aligned}
&\nabla_\theta V_{\text{soft}}^{\pi_\theta}(s) \\
&= \nabla_\theta \sum_a \pi_\theta(a|s) Q_{\text{soft}}^{\pi_\theta}(s,a) \\
&= \sum_a \Big(\nabla_\theta \pi_\theta(a|s) Q_{\text{soft}}^{\pi_\theta}(s,a) + \pi_\theta(a|s) \nabla_\theta Q_{\text{soft}}^{\pi_\theta}(s,a) \Big) \\
&= \sum_a \Big(\nabla_\theta \pi_\theta(a|s) Q_{\text{soft}}^{\pi_\theta}(s,a) + \pi_\theta(a|s) \nabla_\theta \big(r^{\pi_\theta}(s,a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} V_{\text{soft}}^{\pi_\theta}(s') \big) \Big) \\
&= \sum_a \Big(\nabla_\theta \pi_\theta(a|s) Q_{\text{soft}}^{\pi_\theta}(s,a) + \pi_\theta(a|s) \nabla_\theta \big(r^{\pi_\theta}(s,a) + \gamma \sum_{s'} p(s'|s,a) V_{\text{soft}}^{\pi_\theta}(s') \big) \Big) \\
&= \sum_a \Big(\nabla_\theta \pi_\theta(a|s) Q_{\text{soft}}^{\pi_\theta}(s,a) + \pi_\theta(a|s) \nabla_\theta \sum_{s'} \big(r(s,a) + p(s'|s,a) \big(\alpha \mathcal{H}(\cdot|s') + \gamma V_{\text{soft}}^{\pi_\theta}(s') \big) \big) \Big) \\
&= \sum_a \Big(\nabla_\theta \pi_\theta(a|s) Q_{\text{soft}}^{\pi_\theta}(s,a) + \pi_\theta(a|s) \sum_{s'} p(s'|s,a) \big(\alpha \nabla_\theta \mathcal{H}(\cdot|s') + \gamma \nabla_\theta V_{\text{soft}}^{\pi_\theta}(s') \big) \Big) \\
&\quad \text{recursively replace } V_{\text{soft}}^{\pi_\theta}(\cdot) \text{ by the right side expression} \\
&= \sum_x \sum_{t=0}^{\infty} \gamma^t P(s \to x, t, \pi_\theta) \Big(\sum_a \frac{\partial \pi_\theta(x,a)}{\partial \theta} Q_{\text{soft}}^{\pi_\theta}(x,a) + \alpha \sum_a \pi_\theta(x,a) \sum_{x'} p(x'|x,a) \nabla_\theta \mathcal{H}(\cdot|x') \Big)
\end{aligned}
$$

where $P(s \to x, t, \pi_\theta)$ is the probability of going from state s to state x in t steps under policy π_θ, and

$$
\sum_{t=0}^{\infty} \gamma^t P(s \to x, t, \pi_\theta) := d^{\pi_\theta}(s)
$$

is the stationary distribution of Markov chain for π_θ. Now let's consider the object function given by

$$
\begin{aligned}
J^{\text{soft}}(\theta) &= \mathbb{E}_{\tau \sim p(\tau|\theta)} \big[r^\pi(\tau) \big] \\
&= \mathbb{E}_{\tau \sim p(\tau|\theta)} \Big[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t, a_t) + \alpha \mathcal{H}(\cdot|s_{t+1}) \big) \Big] \\
&= \mathbb{E}_{s_0 \sim D} V_{\text{soft}}^{\pi_\theta}(s_0),
\end{aligned}
\tag{18}
$$

where D is the initial state distribution and is independent of θ. The gradient of the object function is given by

$$
\begin{aligned}
\nabla_\theta J^{\text{soft}}(\theta) &= \nabla_\theta \int_{s_0} D(s_0) V_{\text{soft}}^{\pi_\theta}(s_0) ds_0 \\
&= \int_{s_0} D(s_0) \nabla_\theta V_{\text{soft}}^{\pi_\theta}(s_0) ds_0 \\
&\propto \boldsymbol{\nabla_\theta V_{\text{soft}}^{\pi_\theta}(s_0)}.
\end{aligned}
\tag{19}
$$

The gradient is proportional to the derivative of state value $V^{\pi_\theta}_{\text{soft}}(s_0)$ with respect to θ, where $\nabla_\theta V^{\pi_\theta}_{\text{soft}}(s), \forall s \in \mathcal{S}$, has already been obtained. Hence, the soft policy gradient is

$$
\begin{aligned}
&\boldsymbol{\nabla_\theta J(\theta) \propto \nabla_\theta V^{\pi_\theta}(s_0)} \\
&= \sum_x \sum_{t=0}^{\infty} \gamma^t Pr(s_0 \to x, t, \pi_\theta) \Big(\sum_a \frac{\partial \pi_\theta(x,a)}{\partial \theta} Q^{\pi_\theta}_{\text{soft}}(x,a) + \alpha \sum_a \pi_\theta(x,a) \sum_{x'} p(x'|x,a) \nabla_\theta \mathcal{H}(\cdot|x') \Big) \\
&= \sum_s d^{\pi_\theta}(s) \Big(\sum_a \frac{\partial \pi_\theta(s,a)}{\partial \theta} Q^{\pi_\theta}_{\text{soft}}(s,a) + \alpha \sum_a \pi_\theta(s,a) \sum_{s'} p(s'|s,a) \nabla_\theta \mathcal{H}(\cdot|s') \Big) \\
&= \sum_s d^{\pi_\theta}(s) \Big(\sum_a \frac{\pi_\theta(s,a)}{\pi_\theta(s,a)} \frac{\partial \pi_\theta(s,a)}{\partial \theta} Q^{\pi_\theta}_{\text{soft}}(s,a) + \alpha \sum_a \pi_\theta(s,a) \sum_{s'} p(s'|s,a) \nabla_\theta \mathcal{H}(\cdot|s') \Big) \\
&= \mathbb{E}_{s \sim \rho^{\pi_\theta}, a \sim \pi_\theta} \Big[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}_{\text{soft}}(s,a) + \alpha \sum_{s'} p(s'|s,a) \nabla_\theta \mathcal{H}(\cdot|s') \Big] \\
&= \mathbb{E}_{(s,a,s') \sim \beta^{\pi_\theta}} \Big[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}_{\text{soft}}(s,a) + \alpha \nabla_\theta \mathcal{H}(\cdot|s') \Big) \Big]
\end{aligned}
$$

A.2 Hyperparameters for Experiments

Table 2 lists the SSPG parameters used in the comparative evaluation in Figure 1. All the environments are in MuJoCo version 2.0.

Table 2: SSPG Hyperparameters.

Environment	HalfCheetah	Ant	Hopper	Reacher	Walker2d	Swimmer	Humanoid
Q function network	Two hidden layers, hidden-dim = 256, ReLU						
Policy network	Two hidden layers, hidden-dim = 256, ReLU						
replay buffer size M	1e6						
action sample N	1						
batch size	254						
discount factor γ	0.99						
learning rate	3e-4						
target smoothing coefficient λ	0.005						
target update interval	1						
temperature α	0.1	0.1	0.2	0.1	0.1	0.2	0.1

A.3 MuJoCo Environment Specific Parameters

Table 3: Environment Specific Parameters

Environment	Space Dimensions	Action Dimensions
Swimmer-v2	8	2
Reacher-v2	11	2
Hopper-v2	11	3
Walker2d-v2	17	6
HalfCheetah-v2	17	6
Ant-v2	111	8
Humanoid-v2	376	17