Soft Actor Critic

$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \dots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$,

$V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \dots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$,

Policy Evaluation

$V_{i+1}^{\pi_k}(s) = \sum_{s'} P(s'|s, \pi_k(s))[r + \gamma V_i^{\pi_k}(s')]$

Policy Improvement

$\pi_{k+1}(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s, a)[r + \gamma V^{\pi_k}(s')]$

$\nabla J(\theta) = \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(a|s, \theta)}{\pi(a|s, \theta)} \right]$

$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a \left(q_\pi(s, a) - b(s) \right) \nabla \pi(a|s, \theta)$

$\nabla J(\theta) = \mathbb{E}_\pi [\nabla_\theta \log \pi_\theta (a_t | s_t) \hat{A}_t]$

ρ_π: Trajectory distribution induced by a policy π.

$\rho_\pi(s_t)$ & $\rho_\pi(s_t, a_t)$ state marginal of trajectory distribution induced by policy $\pi(s_t, a_t)$

Max-Entropy RL

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot | s_t))]$$

Soft action value function

$\mathcal{T}^\pi Q(s_t, a_t) \triangleq r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p}[V(s_{t+1})]$

Soft state value function

$V(s_t) = \mathbb{E}_{a_t \sim \pi}[Q(s_t, a_t) - \log \pi(a_t | s_t)]$

Tabular Representation

Partition function: $Z^{\pi_{old}}(s_t)$

θ: Policy parameters

w_V: State – value parameters

w_Q: Action – value parameters

- Soft Policy Iteration
  - Soft Policy Evaluation

    $Q^{k+1} = \mathcal{T}^\pi Q^k$; Sequence Q^k converges as $k \to \infty$

    If entropy augmented reward is:

    $r_\pi(s_t, a_t) \triangleq r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim p}[\mathcal{H}(\pi(\cdot | s_{t+1}))]$

    The update rule can be rewritten as:

    $Q(s_t, a_t) \leftarrow r_\pi(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p, a_{t+1} \sim \pi}[Q(s_{t+1}, a_{t+1})]$
  - Soft Policy Improvement

    $\pi_{new}(\cdot | s_t) = \arg\min_{\pi' \in \Pi} D_{KL}\left(\pi'(\cdot | s_t) \parallel e^{Q^{\pi_{old}}(s_t, \cdot) - \log Z^{\pi_{old}}(s_t)}\right)$

    $\pi_{new}(\cdot | s_t) = \arg\min_{\pi' \in \Pi} J_{\pi_{old}}(\pi'(\cdot | s_t))$

    $J_{\pi_{old}}(\pi_{new}(\cdot | s_t)) \leq J_{\pi_{old}}(\pi_{old}(\cdot | s_t))$. Hence,

    $\mathbb{E}_{a_t \sim \pi_{new}}[\log \pi_{new}(a_t, s_t) - Q^{\pi_{old}}(s_t, a_t) + \log Z^{\pi_{old}}(s_t)] \leq$

    $\mathbb{E}_{a_t \sim \pi_{old}}[\log \pi_{old}(a_t, s_t) - Q^{\pi_{old}}(s_t, a_t) + \log Z^{\pi_{old}}(s_t)]$

    Since, partition function $Z^{\pi_{old}}$ depends only on state, the inequality becomes

    $\mathbb{E}_{a_t \sim \pi_{new}}[Q^{\pi_{old}}(s_t, a_t) - \log \pi_{new}(a_t | s_t)] \geq V^{\pi_{old}}(s_t)$

    $Q^{\pi_{old}}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p}[V^{\pi_{old}}(s_{t+1})]$

    $Q^{\pi_{old}}(s_t, a_t) \leq r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p}\left[\mathbb{E}_{a_{t+1} \sim \pi_{new}}[Q^{\pi_{old}}(s_{t+1}, a_{t+1}) - \log \pi_{new}(a_{t+1} | s_{t+1})]\right]$
  - Soft Actor – Critic

    $J_V(w_V) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\frac{1}{2}\left(V_{w_V}(s_t) - \mathbb{E}_{a_t \sim \pi_\theta}\left[Q_{w_Q}(s_t, a_t) - \log \pi_\theta(a_t | s_t)\right]\right)^2\right]$

    $J_Q(w_Q) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}}\left[\frac{1}{2}\left(Q_{w_Q}(s_t, a_t) - \hat{Q}(s_t, a_t)\right)^2\right]$

    $\hat{Q}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p}[V_{\bar{w}_v}(s_{t+1})]$

    $J_\pi(\theta) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[D_{KL}\left(\pi_\theta(\cdot | s_t) \parallel e^{Q_{w_Q}(s_t, \cdot) - \log Z_{w_Q}(s_t)}\right)\right]$

- Soft Actor – Critic

  $\hat{\nabla}_{w_V} J_V(w_V) = \nabla_{w_V} V_{w_V}(s_t) \left(V_{w_V}(s_t) - \mathbb{E}_{a_t \sim \pi_\theta}\left[Q_{w_Q}(s_t, a_t) - \log \pi_\theta(a_t | s_t)\right]\right)$

  $\hat{\nabla}_{w_Q} J_Q(w_Q) = \nabla_{w_Q} Q_{w_Q}(s_t, a_t) \left(Q_{w_Q}(s_t, a_t) - r(s_t, a_t) - \gamma V_{\bar{w}_v}(s_{t+1})\right)$

  $\hat{Q}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p}[V_{\bar{w}_v}(s_{t+1})]$

  $a_t = f_\theta(\epsilon_t; s_t)$

  $\hat{\nabla}_\theta J_\pi(\theta) = \nabla_\theta \log \pi_\theta(a_t | s_t) + \left(\nabla_{a_t} \log \pi_\theta(a_t | s_t) - \nabla_{a_t} Q(s_t, a_t)\right) \nabla_\theta f_\theta(\epsilon_t; s_t)$

Algorithm 1: Soft Actor – Critic

Initicalize parameter vectors $w_V, \bar{w}_V, w_Q, \theta$

for *iteration* $= 1, 2, \dots$ **do**

  for *timestep* $= 1, 2, \dots, N$ **do**

    $a_t \sim \pi_\theta(a_t | s_t)$

    $s_{t+1} \sim p(s_{t+1} | s_t, a_t)$

    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$

  end for

  for *timestep* $= 1, 2, \dots, N$ **do**

    $w_V \leftarrow w_V - \lambda_V \hat{\nabla}_{w_V} J_V(w_V)$

    $w_{Q_i} \leftarrow w_{Q_i} - \lambda_{Q_i} \hat{\nabla}_{w_{Q_i}} J_Q(w_{Q_i})$ for $i \in \{1, 2\}$

    $\theta \leftarrow \theta - \lambda_\pi \hat{\nabla}_\theta J_\pi(\theta)$

    $\bar{w}_V \leftarrow \tau w_V + (1 - \tau) \bar{w}_V$

  end for

end for

https://www.profkumar.co.in/

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, and Sergey Levine. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." In International conference on machine learning, pp. 1861-1870. PMLR, 2018.