Soft Actor Critic メモ [2]

菱沼 徹

2020 年 3 月 8 日

■[2] の利点.

- 方策は，より広く探索するように奨励される.
- 方策は，最適挙動の周りの複数モードを捉えることができる.
- この目的関数で，探索を改善することができるという結果が過去（[1] のこと）に得られている.

■既存研究との比較. 既存研究（[1] のこと）は，ソフト Q 関数の Bellman 方程式を，直接解くことを提案していた．この研究（[2] のこと）では，「現在の方策の Q 関数を評価してオフ方策勾配更新を通じて方策改善する」というような方策反復定式ををを通じて，SAC を実装する方法を議論する．最大エントロピー強化学習においてオフ方策 actor critic を提案したのが新しい（らしい）.

1 最大エントロピー強化学習（[1] の表記）

1.1 定式

確率分布のエントロピーは，次のように定義される（これは一般的な定義である）.

$$\mathcal{H}(p(\cdot)) = -\int p(x)\ln p(x)dx = \mathbb{E}_{x\sim p}[-\ln p(x)] \geq 0$$

方策 $\pi(a_t|s_t)$ により誘導される軌道における状態と状態行動対の分布を $\rho_\pi(s_t)$ と $\rho_\pi(s_t, a_t)$ とおく．ソフト Q 関数（soft Q-function）を次のように定義する.

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{(s_{t+1},\cdots)\sim\rho_\pi}\left[\sum_{\ell=1}^{\infty}\gamma^\ell\left(r(s_{t+\ell}, a_{t+\ell}) + \alpha\mathcal{H}(\pi(\cdot|s_{t+\ell}))\right)\right]$$

ソフト価値関数（soft value function）を次のように定義する.

$$V^\pi(s_t) = \alpha\ln\int\exp\left(\frac{1}{\alpha}Q(s_t, a)\right)da$$

方策の評価関数を次のように定義する.

$$J(\pi) = \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[Q^\pi(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\right]$$

1.2 算数

次のように変形できる．

$$
\begin{aligned}
Q(s_t, a_t) &= r(s_t, a_t) + \mathbb{E}_{(s_{t+1}, \cdots) \sim \rho_\pi} \left[\sum_{\ell=1}^{\infty} \gamma^\ell \left(r(s_{t+\ell}, a_{t+\ell}) + \alpha \mathcal{H}(\pi(\cdot|s_{t+\ell})) \right) \right] \\
&= r(s_t, a_t) + \mathbb{E}_{(s_{t+1}, \cdots) \sim \rho_\pi} \left[\gamma r(s_{t+1}, a_{t+1}) + \gamma \alpha \mathcal{H}(\pi(\cdot|s_{t+1})) + \gamma \sum_{\ell=1}^{\infty} \gamma^\ell \left(r(s_{t+1+\ell}, a_{t+1+\ell}) + \alpha \mathcal{H}(\pi(\cdot|s_{t+1+\ell})) \right) \right] \\
&= r(s_t, a_t) + \gamma \mathbb{E}_{(s_{t+1}, a_{t+1}) \sim \rho_\pi} \left[\alpha \mathcal{H}(\pi(\cdot|s_{t+1})) + Q(s_{t+1}, a_{t+1}) \right] \\
&= r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \rho_\pi} \left[\alpha \mathcal{H}(\pi(\cdot|s_{t+1})) + \mathbb{E}_{a_{t+1} \sim \pi} \left[Q(s_{t+1}, a_{t+1}) \right] \right]
\end{aligned}
$$

また，次が成り立つ．

$$
\exp\left(\frac{1}{\alpha} V(s_t)\right) = \exp\left\{ \ln \left(\int \exp\left(\frac{1}{\alpha} Q(s_t, a)\right) da \right) \right\} = \int \exp\left(\frac{1}{\alpha} Q(s_t, a)\right) da
$$

$$
\exp\left[\frac{1}{\alpha} Q(s_t, a_t) - \frac{1}{\alpha} V(s_t)\right] = \frac{\exp \frac{1}{\alpha} Q(s_t, a_t)}{\exp\left(\frac{1}{\alpha} V(s_t)\right)} = \frac{\exp \frac{1}{\alpha} Q(s_t, a_t)}{\int \exp\left(\frac{1}{\alpha} Q(s_t, a)\right) da}
$$

■方策 $\pi(a_t|s_t) \propto \left(\frac{1}{\alpha} Q(s_t, a_t)\right)$ に対して成り立つこと．　もし方策を $\pi(a_t|s_t) = \exp\left[\frac{1}{\alpha} Q(s_t, a_t) - \frac{1}{\alpha} V(s_t)\right] \propto \left(\frac{1}{\alpha} Q(s_t, a_t)\right)$ のように表現したとすると，次が成り立つ．

$$
\mathcal{H}(\pi(\cdot|s_t)) = \mathbb{E}_{a_t \sim \pi}[-\ln \pi(a_t|s_t)] = -\mathbb{E}_{a_t \sim \pi}\left[\frac{1}{\alpha} Q(s_t, a_t) - \frac{1}{\alpha} V(s_t)\right] = -\mathbb{E}_{a_t \sim \pi}\left[\frac{1}{\alpha} Q(s_t, a_t)\right] + \frac{1}{\alpha} V(s_t)
$$

これを整理すると，

$$
V(s_t) = \mathbb{E}_{a_t \sim \pi}\left[Q(s_t, a_t)\right] + \alpha \mathcal{H}(\pi(\cdot|s_t)) = \mathbb{E}_{a_t \sim \pi}\left[Q(s_t, a_t) - \alpha \ln \pi(a_t|s_t)\right]
$$

これを用いて，

$$
Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{(s_{t+1}) \sim \rho_\pi}\left[V(s_{t+1})\right]
$$

1.3 最適解の性質

■最適方策（定理 1）．　評価関数 $J(\pi)$ を最小化する方策 π^* は，$\pi^*(a_t|s_t) \propto \exp(\frac{1}{\alpha} Q^*(s, \cdot))$ という方策で与えることができる（Q^* は方策 π^* のソフト Q 関数）．

証明は [1] の付録 A.1 を参照．証明のロジックとしては，

- 任意の方策 π に対して，同等以上の評価関数を持つ方策 $\tilde{\pi} \propto \exp(\frac{1}{\alpha} Q(s, \cdot))$ が必ず存在する事を示す．
- したがって，$\tilde{\pi} \propto \exp(\frac{1}{\alpha} Q(s, \cdot))$ のクラスに最適方策が存在する．
- Notice that for convenience, we set the entropy parameter α to 1. The theory can be easily adapted by dividing rewards by α.

■Soft Bellman 方程式（定理 2）．

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

次のようにして証明する．

- $\pi(a_t|s_t) \propto \left(\frac{1}{\alpha} Q(s_t, a_t)\right)$ に対して，$Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{(s_{t+1}) \sim \rho_\pi}\left[V(s_{t+1})\right]$ が成り立つ（上の算数）．
- 最適方策は，$\pi^*(a_t|s_t) \propto \left(\frac{1}{\alpha} Q^*(s_t, a_t)\right)$ である（定理 1）．

■Soft Q 反復（定理 3）．　次の不動点反復アルゴリズムにより，最適方策のソフト Q 関数とソフト価値関数が（理論的には）得られる．

$$Q(s_t, a_t) \leftarrow r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p}[V(s_{t+1})]$$
$$V(s_t) \leftarrow \alpha \ln \int \exp\left(\frac{1}{\alpha} Q(s_t, a)\right) da$$

ただし，この反復アルゴリズムを連続空間において実行することは実際にはできないので，適当な近似を導入して用いることになる．

2 SAC [2]

2.1 数式

■ソフト方策改善．

$$\begin{aligned}
\pi^{new} &= \arg\min_{\pi} D_{KL}\left(\pi(\cdot|s_t) \middle\| \frac{\exp\left(\frac{1}{\alpha} Q^{\pi^{old}}(s_t, \cdot)\right)}{\frac{1}{\alpha} V^{\pi^{old}}(s_t)}\right) \\
&= \arg\min_{\pi} \int \pi(a_t|s_t) \left(\ln \pi(a_t|s_t) - \ln \frac{\exp\left(\frac{1}{\alpha} Q^{\pi^{old}}(s_t, \cdot)\right)}{\frac{1}{\alpha} V^{\pi^{old}}(s_t)}\right) da_t \\
&= \arg\min_{\pi} \int \pi(a_t|s_t) \left(\ln \pi(a_t|s_t) - \frac{1}{\alpha} Q^{\pi^{old}}(s_t, \cdot) + \ln \frac{1}{\alpha} V^{\pi^{old}}(s_t)\right) da_t \\
&= \arg\max_{\pi} \int \pi(a_t|s_t) \left(Q^{\pi^{old}}(s_t, \cdot) - \alpha \ln \pi(a_t|s_t)\right) da_t \\
&= \arg\max_{\pi} \mathbb{E}_{a_t \sim \pi}[Q^{\pi^{old}}(s_t, \cdot) - \alpha \ln \pi(a_t|s_t)] = \arg\max_{\pi} \mathbb{E}_{a_t \sim \pi}[Q^{\pi^{old}}(s_t, \cdot) + \alpha \mathcal{H}(\pi(\cdot|s_t))]
\end{aligned}$$

まず，$\pi^{new} = \pi^{old}$ として選ぶことが可能であるため，次の不等式が成り立つ．

$$\begin{aligned}
\mathbb{E}_{a_t \sim \pi^{new}}[Q^{\pi^{old}}(s_t, \cdot) + \alpha \mathcal{H}(\pi^{new}(\cdot|s_t))] &= \max_{\pi} \mathbb{E}_{a_t \sim \pi}[Q^{\pi^{old}}(s_t, \cdot) + \alpha \mathcal{H}(\pi(\cdot|s_t))] \\
&\geq \mathbb{E}_{a_t \sim \pi^{old}}[Q^{\pi^{old}}(s_t, \cdot) + \alpha \mathcal{H}(\pi^{old}(\cdot|s_t))] = J(\pi^{old})
\end{aligned}$$

また，次の不等式が成り立つ（[1] の付録参照）．

$$Q^{\pi^{new}}(s_t, a_t) \geq Q^{\pi^{old}}(s_t, a_t)$$

これら 2 つの不等式を合わせると，

$$\begin{aligned}
J(\pi^{new}) &= \mathbb{E}_{a_t \sim \pi^{new}}[Q^{\pi^{new}}(s_t, \cdot) + \alpha \mathcal{H}(\pi^{new}(\cdot|s_t))] \\
&\geq \mathbb{E}_{a_t \sim \pi^{new}}[Q^{\pi^{old}}(s_t, \cdot) + \alpha \mathcal{H}(\pi^{new}(\cdot|s_t))] \\
&\geq \mathbb{E}_{a_t \sim \pi^{old}}[Q^{\pi^{old}}(s_t, \cdot) + \alpha \mathcal{H}(\pi^{old}(\cdot|s_t))] = J(\pi^{old})
\end{aligned}$$

従って，$D_{KL}\left(\pi(\cdot|s_t) \middle\| \frac{\exp\left(\frac{1}{\alpha} Q^{\pi^{old}}(s_t, \cdot)\right)}{\frac{1}{\alpha} V^{\pi^{old}}(s_t)}\right)$ を最小化する方策 π を選べば，評価指標は必ず改善される．

2.2 アルゴリズム

■近似対象．

- ソフト状態価値関数：$V_\psi(s)$

- ソフト Q 関数：$Q_\theta(s,a)$
- 方策：$\pi_\phi(a|s)$

■ソフト状態価値関数.　残差事情最小化により訓練する.

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

勾配を取ってサンプル近似する.

$$\begin{aligned}\nabla_\psi J_V(\psi) &= \mathbb{E}_{s_t\sim D}\left[\left(V_\psi(s_t) - \mathbb{E}_{a_t\sim\pi_\phi}\left[Q_\theta(s_t,a_t) - \ln\pi_\phi(a_t|s_t)\right]\right)\nabla_\psi V_\psi(s_t)\right] \\ &\approx \left(V_\psi(s_t) - \left[Q_\theta(s_t,a_t) - \ln\pi_\phi(a_t|s_t)\right]\right)\nabla_\psi V_\psi(s_t)\end{aligned}$$

注意：原理的にはソフト状態価値関数は必要ない．しかし，ソフト価値関数のための近似器を分離して持っておくことは，また他のネットワークと同時に訓練をするのに便利である.

■ソフト Q 関数.　ソフト Bellman 残差を最小化するように訓練する.

$$J_Q(\theta) = \frac{1}{2}\mathbb{E}_{(s_t,a_t)\sim D}\left[\left(Q_\theta(s_t,a_t) - \left[r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim p}[V_{\tilde\psi}(s_{t+1})]\right]\right)^2\right]$$

勾配を取ってサンプル近似すると，

$$\begin{aligned}\nabla_\theta J_Q(\theta) &= \mathbb{E}_{(s_t,a_t)\sim D}\left[\left(Q_\theta(s_t,a_t) - \left[r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim p}[V_{\tilde\psi}(s_{t+1})]\right]\right)\nabla_\theta Q_\theta(s_t,a_t)\right] \\ &\approx \left(Q_\theta(s_t,a_t) - \left[r(s_t,a_t) + \gamma V_{\tilde\psi}(s_{t+1})\right]\right)\nabla_\theta Q_\theta(s_t,a_t)\end{aligned}$$

ターゲット価値ネットワーク $V_{\tilde\psi}$ のターゲット重み $\tilde\psi$ を，現在の価値関数重みと周期的に一致するように選ぶ（[2] の付録 E 参照）

■方策.　期待 KL ダイバージェンスを最小化するように学習する.

$$\begin{aligned}J_\pi(\phi) &= \mathbb{E}_{s_t\sim D}\left[D_{KL}\left(\pi_\phi(\cdot|s_t)\middle\|\frac{\exp(Q_\theta(s_t,\cdot))}{Z_\theta(s_t)}\right)\right] \\ &= \mathbb{E}_{s_t\sim D}\left[\int\pi_\phi(a_t|s_t)\left(\ln\pi_\phi(a_t|s_t) - \ln\frac{\exp(Q_\theta(s_t,a_t))}{Z_\theta(s_t)}\right)da_t\right] \\ &= \mathbb{E}_{s_t\sim D}\left[\int\pi_\phi(a_t|s_t)\left(\ln\pi_\phi(a_t|s_t) - Q_\theta(s_t,a_t)\right)da_t + \ln Z_\theta(s_t)\right]\end{aligned}$$

ここで，分配関数 $Z_\theta(s_t) = \int\exp(Q_\theta(s_t,a_t))dt$ である．勾配を取ると，

$$\begin{aligned}\nabla_\phi J_\pi(\phi) &= \mathbb{E}_{s_t\sim D}\left[\nabla_\phi\int\pi_\phi(a_t|s_t)\left(\ln\pi_\phi(a_t|s_t) - Q_\theta(s_t,a_t)\right)da_t\right] \\ &= \mathbb{E}_{s_t\sim D}\left[\nabla_\phi\mathbb{E}_{a_t\sim\pi_\phi(\cdot|s_t)}\left[\ln\pi_\phi(a_t|s_t) - Q_\theta(s_t,a_t)\right]\right]\end{aligned}$$

再パラメータ化 $a_t = f_\phi(\epsilon_t;s_t)$ としてサンプル近似すると，

$$\begin{aligned}\nabla_\phi J_\pi(\phi) &= \mathbb{E}_{s_t\sim D}\left[\nabla_\phi\mathbb{E}_{\epsilon_t\sim\mathcal{N}}\left[\ln\pi_\phi(f_\phi(\epsilon_t;s_t)|s_t) - Q_\theta(s_t,f_\phi(\epsilon_t;s_t))\right]\right] \\ &= \mathbb{E}_{s_t\sim D}\left[\mathbb{E}_{\epsilon_t\sim\mathcal{N}}\left[\nabla_\phi\left\{\ln\pi_\phi(f_\phi(\epsilon_t;s_t)|s_t) - Q_\theta(s_t,f_\phi(\epsilon_t;s_t))\right\}\right]\right] \\ &= \mathbb{E}_{s_t\sim D}\left[\mathbb{E}_{\epsilon_t\sim\mathcal{N}}\left[\nabla_\phi\left\{\ln\pi_\phi(a_t|s_t) - Q_\theta(s_t,a_t)\right\} + (\nabla_\phi f_\phi(\epsilon_t;s_t))\nabla_{a_t}\left\{\ln\pi_\phi(a_t|s_t) - Q_\theta(s_t,a_t)\right\}\right]\right] \\ &= \mathbb{E}_{s_t\sim D}\left[\mathbb{E}_{\epsilon_t\sim\mathcal{N}}\left[\nabla_\phi\ln\pi_\phi(a_t|s_t) + (\nabla_\phi f_\phi(\epsilon_t;s_t))\left\{\nabla_{a_t}\ln\pi_\phi(a_t|s_t) - \nabla_{a_t}Q_\theta(s_t,a_t)\right\}\right]\right] \\ &\approx \nabla_\phi\ln\pi_\phi(a_t|s_t) + (\nabla_\phi f_\phi(\epsilon_t;s_t))\left\{\nabla_{a_t}\ln\pi_\phi(a_t|s_t) - \nabla_{a_t}Q_\theta(s_t,a_t)\right\}\end{aligned}$$

■2 つの Q 関数の利用． 方策改善における正のバイアス（これは価値ベース手法の性能を低下させることが知られている）を低減化するために，二つの Q 関数を使う（パラメータをそれぞれ θ_1 と θ_2 とする）．これらの Q 関数の内小さい方を，$\nabla_\psi J_V(\psi)$ と $\nabla_\phi J_\pi(\phi)$ のサンプル近似に使う．1 つの Q 関数を使っても学習はできるものの，2 つの Q 関数を使う方がより早く学習できる．

■リプレイバッファ． 現在の方策を用いた環境から経験を集め，リプレイバッファからサンプルされたバッチから確率的勾配を用いて関数近似器を更新する．実際には，1 環境ステップに対して，one or several 勾配ステップを行う．価値推定器と方策の両方が完全にオフ方策データ上で訓練できるため，リプレイバッファからオフ方策データを使うことが可能である．

参考文献

[1] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361, 2017.

[2] Tuomas Haarnoja, 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*, pages 1861–1870, 2018.