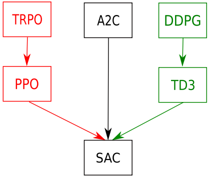
From Policy Gradient to Actor-Critic methods

Soft Actor Critic

Olivier Sigaud
with the help of Thomas Pierrot

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Soft Actor Critic: The best of two worlds



- TRPO and PPO: π_{θ} stochastic, on-policy, low sample efficiency, stable
- DDPG and TD3: π_{θ} deterministic, replay buffer, better sample efficiency, unstable
- SAC: "Soft" means "entropy regularized", π_{θ} stochastic, replay buffer
- Adds entropy regularization to favor exploration (follow-up of several papers)
- Attempt to be stable and sample efficient
- Three successive versions

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A. Abbeel, P. et al. (2018) Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018) Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*

Haarnoja, T. Tang, H., Abbeel, P. and Levine, S. (2017) Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*

Soft Actor-Critic

SAC learns a **stochastic** policy π^* maximizing both rewards and entropy:

$$\pi^* = \arg\max_{\pi_{\boldsymbol{\theta}}} \sum_t \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_{\pi_{\boldsymbol{\theta}}}} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)) \right]$$

- The entropy is defined as: $\mathcal{H}(\pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)) = \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[-\log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \right]$
- SAC changes the traditional MDP objective
- Thus, it converges toward different solutions
- Consequently, it introduces a new value function, the soft value function
- As usual, we consider a policy $\pi_{\boldsymbol{\theta}}$ and a soft action-value function $\hat{Q}^{\pi_{\boldsymbol{\theta}}}$

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. (2016) Asynchronous methods for deep reinforcement learning. *arXiv preprint arXiv:1602.01783*

Soft policy evaluation

- Usually, we define $\hat{V}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t, \mathbf{a}_t) \right]$
- In soft updates, we rather use:

$$\begin{aligned}
\hat{V}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t) &= \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t, \mathbf{a}_t) \right] + \alpha \mathcal{H}(\pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)) \\
&= \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t, \mathbf{a}_t) \right] + \alpha \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[- \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \right] \\
&= \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t, \mathbf{a}_t) - \alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \right]
\end{aligned}$$

Critic updates

- We define a standard Bellman operator:

$$\mathcal{T}^{\pi}\hat{Q}^{\pi\boldsymbol{\theta}}_{(}\mathbf{s}_t,\mathbf{a}_t) = r(\mathbf{s}_t,\mathbf{a}_t) + \gamma V^{\pi\boldsymbol{\theta}}_{(}\mathbf{s}_{t+1})$$

$$= r(\mathbf{s}_t,\mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_{t+1})}\left[\hat{Q}^{\pi\boldsymbol{\theta}}_{(}\mathbf{s}_{t+1},\mathbf{a}_t) - \alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_{t+1})\right]$$

Critic parameters can be learned by minimizing:

$$J_Q(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t,\mathbf{s}_{t+1})\sim\mathcal{D}}\left[\left(r(\mathbf{s}_t,\mathbf{a}_t) + \gamma \hat{V}^{\pi\boldsymbol{\theta}}_{(}\mathbf{s}_{t+1}) - \hat{Q}^{\pi\boldsymbol{\theta}}_{(}\mathbf{s}_t,\mathbf{a}_t)\right)^2\right]$$

where $V^{\pi\boldsymbol{\theta}}_{(}\mathbf{s}_{t+1}) = \mathbb{E}_{\mathbf{a}\sim\pi_{\boldsymbol{\theta}}(.|\mathbf{s}_{t+1})}\left[\hat{Q}^{\pi\boldsymbol{\theta}}_{(}\mathbf{s}_{t+1},\mathbf{a}) - \alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathbf{s}_{t+1})\right]$

- Similar to DDPG update, but with entropy

Actor updates

- Update policy such as to become greedy w.r.t to the soft Q-value
- Choice: update the policy towards the exponential of the soft Q-value

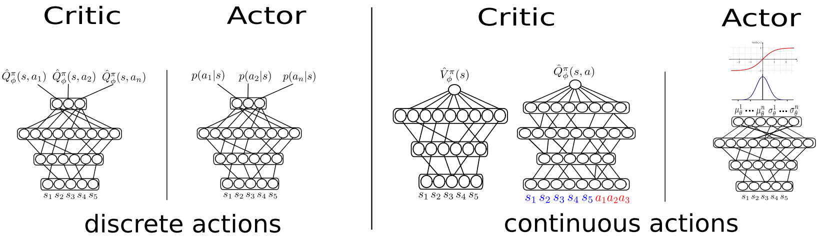
$$J_\pi(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}[KL(\pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t))||\frac{\exp(\frac{1}{\alpha}\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t,.))}{Z_{\boldsymbol{\theta}}(\mathbf{s}_t)}].$$

- $Z_{\boldsymbol{\theta}}(\mathbf{s}_t)$ is just a normalizing term to have a distribution
- SAC does not minimize directly this expression but a surrogate one that has the same gradient w.r.t $\boldsymbol{\theta}$

The policy parameters can be learned by minimizing:

$$J_\pi(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)}\left[\alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) - \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(}\mathbf{s}_t, \mathbf{a}_t)\right]\right]$$

Continuous vs discrete actions setting



- SAC works in both the discrete action and the continuous action setting
- Discrete action setting:
  - The critic takes a state and returns a Q-value per action
  - The actor takes a state and returns probabilities over actions
- Continuous action setting:
  - The critic takes a state and an action vector and returns a scalar Q-value
  - Need to choose a distribution function for the actor
  - SAC uses a squashed Gaussian: $\mathbf{a} = \tanh(n)$ where $n \sim \mathcal{N}(\mu, \sigma)$

Continuous vs discrete actions setting

- In $J_\pi(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[\alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) - \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(} \mathbf{s}_t, \mathbf{a}_t) \right] \right]$
- SAC updates require to estimate an expectation over actions sampled from the actor,
- That is $\mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|s)} \left[F(\mathbf{s}_t, \mathbf{a}_t) \right]$ where F is a scalar function.

- In the discrete action setting, $\pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)$ is a vector of probabilities
  - $\mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[F(\mathbf{s}_t, \mathbf{a}_t) \right] = \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)^T F(\mathbf{s}_t, .)$

- In the continuous action setting:
  - The actor returns $\mu_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$
  - Re-parameterization trick: $\mathbf{a}_t = \tanh(\mu_{\boldsymbol{\theta}} + \epsilon.\sigma_{\boldsymbol{\theta}})$ where $\epsilon \sim \mathcal{N}(0, 1)$
  - Thus, $\mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[F(\mathbf{s}_t, \mathbf{a}_t) \right] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \left[F(\mathbf{s}_t, \tanh(\mu_{\boldsymbol{\theta}} + \epsilon \sigma_{\boldsymbol{\theta}})) \right]$
  - This trick reduces the variance of the expectation estimate
  - And allows to backprop through the expectation w.r.t $\boldsymbol{\theta}$

Critic update improvements (from TD3)

- As in TD3, SAC uses two critics $\hat{Q}_1^{\pi_{\boldsymbol{\theta}}}$ and $\hat{Q}_2^{\pi_{\boldsymbol{\theta}}}$
- The TD-target becomes:

$$y_t = r + \gamma \mathbb{E}_{\mathbf{a}_{t+1} \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_{t+1})} \left[\min_{i=1,2} \hat{Q}_i^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}) \right]$$

And the losses:

$$\begin{cases} J(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) \sim \mathcal{D}} \left[\left(\hat{Q}_1^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}_t, \mathbf{a}_t) - y_t \right)^2 + \left(\hat{Q}_2^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}_t, \mathbf{a}_t) - y_t \right)^2 \right] \\ J(\boldsymbol{\theta}) = \mathbb{E}_{s \sim \mathcal{D}} \left[\mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[\alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) - \min_{i=1,2} \hat{Q}_i^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] \end{cases}$$

Fujimoto, S., van Hoof, H., & Meger, D. (2018) Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*

Automatic Entropy Adjustment

- The temperature α needs to be tuned for each task
- Finding a good α is non trivial
- Instead of tuning α, tune a lower bound $\mathcal{H}_0$ for the policy entropy
- And change the optimization problem into a constrained one

$$\begin{cases} \pi^* = \underset{\pi}{\operatorname{argmax}} \sum_t \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_{\pi_{\boldsymbol{\theta}}}} \left[r(\mathbf{s}_t,\mathbf{a}_t)\right] \\ \text{s.t. } \forall t\ \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_{\pi_{\boldsymbol{\theta}}}} \left[-\log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)\right] \geq \mathcal{H}_0, \end{cases}$$

- Use heuristic to compute $\mathcal{H}_0$ from the action space size

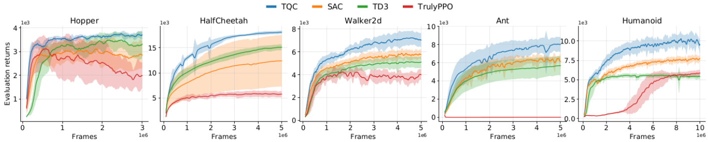
α can be learned to satisfy this constraint by minimizing:

$$J(\alpha) = \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D}} \left[\mathbb{E}_{\mathbf{a}_t\sim\pi_{\boldsymbol{\theta}}(.|\mathbf{s}_t)} \left[-\alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) - \alpha\mathcal{H}_0\right]\right]$$

Practical algorithm

- Initialize neural networks $\pi_{\boldsymbol{\theta}}$ and $\hat{Q}^{\pi_{\boldsymbol{\theta}}}$ weights
- Play k steps in the environment by sampling actions with $\pi_{\boldsymbol{\theta}}$
- Store the collected transitions in a replay buffer
- Sample k batches of transitions in the replay buffer
- Update the temperature α, the actor and the critic using SGD
- Repeat this cycle until convergence

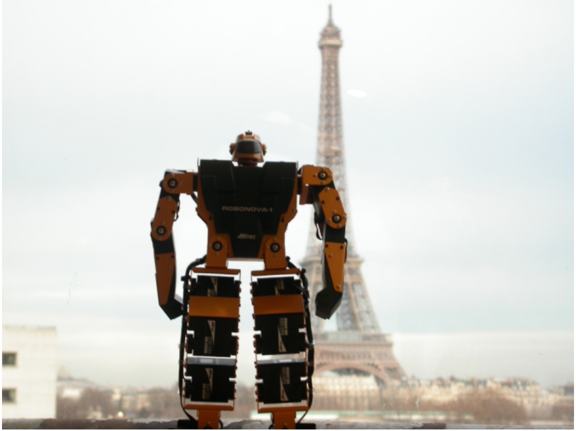
Truncated Quantile Critics



- Using a distribution of estimates is more stable than a single estimate
- To fight overestimation bias, TD3 and SAC take the min over two critics
- Truncating the higher quantiles is another option
- No need for two critics
- Better performance than SAC

Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, pp. 5556–5566. PMLR, 2020

Any question?



Send mail to: `Olivier.Sigaud@upmc.fr`

Scott Fujimoto, Herke van Hoof, and Dave Meger.
Addressing function approximation error in actor-critic methods.
arXiv preprint arXiv:1802.09477, 2018.

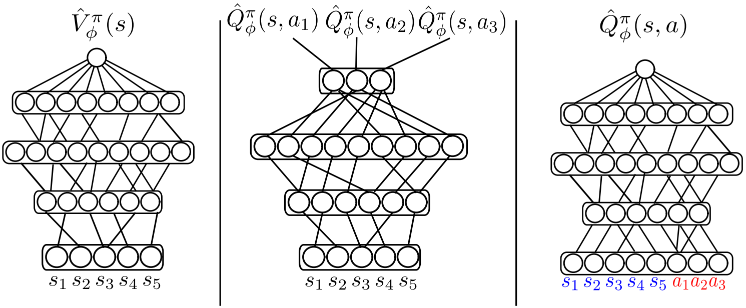
Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine.
Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.
arXiv preprint arXiv:1801.01290, 2018a.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al.
Soft actor-critic algorithms and applications.
arXiv preprint arXiv:1812.05905, 2018b.

Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov.
Controlling overestimation bias with truncated mixture of continuous distributional quantile critics.
In *International Conference on Machine Learning*, pp. 5556–5566. PMLR, 2020.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu.
Asynchronous methods for deep reinforcement learning.
arXiv preprint arXiv:1602.01783, 2016.

Practical implementation of neural critics



- $\hat{V}^{\pi_{\boldsymbol{\theta}}}$ is smaller, but not necessarily easier to estimate
- Given the implicit max in $\hat{V}^{\pi}_{(}s)$, approx. may be less stable than $\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{(}s)$ (?)
- Note: a critic network provides a value even in unseen states