Back to J

- Objective:

$$J(\theta) = \sum_s \mu_{\pi_\theta}(s) v_{\pi_\theta}(s)$$
$$= \sum_s \mu_{\pi_\theta}(s) \sum_a \pi_\theta(a|s) q_{\pi_\theta}(s,a)$$

- True gradient:

$$\nabla J(\theta) \propto \sum_s \mu_{\pi_\theta}(s) \sum_a \nabla \pi_\theta(a|s) q_{\pi_\theta}(s,a)$$
$$\propto \sum_s \mu_{\pi_\theta}(s) \sum_a \nabla \pi_\theta(a|s)(q_{\pi_\theta}(s,a) - v_{\pi_\theta}(s))$$

- Stochastic gradient:

$$\widehat{\nabla} J(\theta) \propto \nabla \log \pi_\theta(A_t|S_t)(q_{\pi_\theta}(S_T, A_T) - v_{\pi_\theta}(S_t))$$
$$\propto \nabla \log \pi_\theta(A_t|S_t) a_{\pi_\theta}(S_T, A_T)$$

- On policy algorithm if we can estimate $a_{\pi_\theta}(S_T, A_T) = q_{\pi_\theta}(S_T, A_T) - v_{\pi_\theta}(S_t)$. (Critic)

Off-Policy J

- Objective:

$$J_b(\theta) = \sum_s \mu_b(s) v_{\pi_\theta}(s) = \sum_s \mu_b(s) \sum_a \pi_\theta(a|s) q_{\pi_\theta}(s,a)$$

- True gradient:

$$\nabla J_b(\theta) = \sum_s \mu_b(s) \sum_a \left(\nabla \pi_\theta(a|s) q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \boxed{\nabla q_{\pi_\theta}(s,a)}\right)$$

- ∇q_{π_θ} term hard to compute!
- Descent direction:

$$\tilde{\nabla} J_b(\theta) = \sum_s \mu_b(s) \sum_a \nabla \pi_\theta(a|s) q_{\pi_\theta}(s,a) = \sum_s \mu_b(s) \sum_a \nabla \pi_\theta(a|s) a_{\pi_\theta}(s,a)$$

- Stochastic descent direction

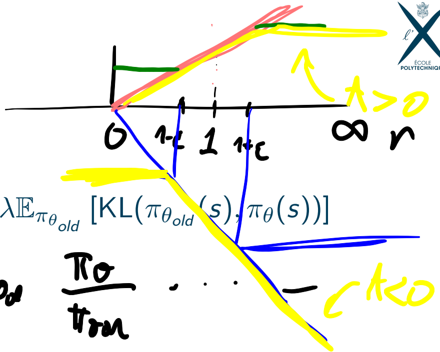
$$\hat{\tilde{\nabla}} J_b(\theta) = \frac{\pi_\theta(A_t)|S_t}{b(A_t|S_t)} \nabla \log \pi_\theta(a|s) a_{\pi_\theta}(S_t, A_t)$$

- Off-policy algorithm if we can estimate $a_{\pi_\theta}(S_t, A_t)$ (Critic)

Trust Region Policy Optimization

- Local objective:
$$J_{\pi_{\theta_{old}}}(\theta) = \sum_s \mu_{\pi_{old}}(s) \sum_a \pi_\theta(a|s) a_{\pi_{\theta_{old}}}(s,a)$$
- If convergence we recover the on-policy goal.
- True gradient:
$$\nabla J_{\pi_{\theta_{old}}}(\theta) = \sum_s \mu_{\pi_{\theta_{old}}}(s) \sum_a \nabla \pi_\theta(a|s) a_{\pi_{\theta_{old}}}(s,a)$$
- Identical to the descent direction of the off-policy algorithm at θ_{old}.
- Optimization of the local objective only in a neighborhood of π_{old}:
$$\sup_s \mathrm{KL}(\pi_{\theta_{old}}(s), \pi_\theta(s)) \leq \epsilon$$
- Strong link with a backtracking algorithm of the off-line version.
- Need to replace the trust region by an approximate one based on $\mathbb{E}_{\pi_{\theta_{old}}}[\mathrm{KL}(\pi_{\theta_{old}}(s), \pi_\theta(s))]$ (quadratic constraints in θ)

Proximal Policy Optimization



- State Of The Art actor-critic algorithm.
- First version:

$$J_{\pi_{\theta_{old}}}(\theta) = \sum_s \mu_{\pi_{old}}(s) \sum_a \pi_\theta(a|s) a_{\pi_{\theta_{old}}}(s,a) + \lambda \mathbb{E}_{\pi_{\theta_{old}}} [\mathrm{KL}(\pi_{\theta_{old}}(s), \pi_\theta(s))]$$

- Very similar to TRPO.

$$= \sum_s \mu_{\pi_{old}} \sum_a \pi_{\theta_{old}} \frac{\pi_\theta}{\pi_{old}} \cdots$$

- Second version:

$$J_{\pi_{\theta_{old}}}(\theta) = \sum_s \mu_{\pi_{old}}(s) \sum_a \pi_{\theta_{old}}(a|s)$$
$$\times \min\left(r_{\theta_{old},\theta}(a,s) a_{\pi_{\theta_{old}}}(s,a), \mathrm{clip}(r_{\theta_{old},\theta}(a,s), 1-\epsilon, 1+\epsilon) a_{\pi_{\theta_{old}}}(s,a) \right)$$

with $r_{\theta_{old}}(\theta) = \pi_\theta(a|s)/\pi_{\theta_{old}}(a|s)$.

- Worst case scenario for the advantage as soon as the ratio is away from 1.
- No theoretical justification but simple and efficient algorithm.

Soft Actor Critic

$\pi^* = \arg\max q(s,a)$

$\pi \propto e^{-\frac{1}{\lambda} q(s,a)}$

- Modification of the reward to favor high entropy policy:
$$R_t \to R_t + \lambda \mathcal{H}(\pi(S_t))$$
- Classical state value fonction:
$$v(s) = \mathbb{E}\left[\sum_{i=0} \gamma^i (R_{t+i+1} + \lambda \mathcal{H}(\pi(\cdot|S_{t+i}))\middle| S_t = s\right]$$
$+1$
- Modified state action value fonction defined by
$$v(s) = \sum_a \pi(a|s)\left(q(s,a) - \lambda \log(\pi(a|s))\right)$$
- Fixed point operator:
$$\begin{aligned}\mathcal{T}^\pi q(s,a) &= r(s,a) + \mathbb{E}\left[\gamma v(s', a)\right] \\ &= r(s,a) + \mathbb{E}_\pi\left[q(s', a') - \lambda \log(\pi(a'|s'))\right]\end{aligned}$$

Soft Actor Critic

- Policy improvement rule:
$$\pi^+(\cdot|s) = \underset{\pi(\cdot|s)}{\operatorname{argmax}} \sum_a \pi(a|s)\left(q(s,a) - \lambda \log(\pi(a|s))\right)$$
$$\pi^+(a|s) \propto \exp(-\frac{1}{\lambda} q(s,a))$$
implies $v_{\pi^+}(s) \geq v_\pi(s)$.
- Stronger link between the critic and the actor with an ideal update following the policy improvement rule.

Soft Actor Critic

- Parametric π_θ, Q_ϕ and $Q_{\phi'}\ldots$
- Optimization in ϕ:

$$V_{\phi'}(s,a) = \mathbb{E}_{\pi_\theta}\left[Q_{\phi'}(S_t,A_t) - \lambda \log \pi_\theta(A_t|S_t)\right]$$

$$J(\phi) = \mathbb{E}_{\pi_\theta}\left[\left(Q_\phi(S_t,A_t) - (r(S_t,A_t) + \gamma \mathbb{E}\left[V_{\phi'}(S_{t+1})\right])\right)^2\right]$$

$$\widehat{\nabla} J(\phi) = 2\nabla Q(S_t,A_t) \times \left(Q_\phi(S_t,A_t) - (r(S_t,A_t) + \gamma\left(Q_{\phi'}(S_{t+1},A_{t+1}) - \lambda \log \pi_\theta(A_{t+1}|S_{t+1})\right)\right)\right)$$

- ϕ': slow version of ϕ (exponential average) $\phi' \leftarrow (1-\rho)\,\phi' + \rho\,\phi$
- Two-scales trick!

Soft Actor Critic

- Optimization in θ:

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\mathrm{KL}(\pi_\theta(\cdot|S_t), e^{\frac{1}{\lambda}Q_\phi(S_t,\cdot)}/Z(S_t, \phi))\right]$$

$$= \mathbb{E}_{\pi_\theta}\left[\sum_a \pi_\theta(a|S_t)\left(-\log \pi_\theta(a|S_t) + \frac{1}{\lambda}Q_\phi(S_t, a)\right)\right] + Cst$$

- Easy optimization when neglecting the effect of π_θ in S_t...

$$\widehat{\tilde{\nabla}} J(\theta) = \sum_a \left(-log \pi_\theta(a|S_t) + \frac{1}{\lambda}Q_\phi(S_t, a) - 1\right)\nabla_\theta \pi_\theta(a|S_t)$$

- θ is updated at a much slower pace than ϕ.
- Two-scales algorithm trick again!

Soft Actor Critic

$a = \mu + \sigma \varepsilon$

- Adaptation possible to continuous action using the reparametrization trick:
$$A_t = \Phi_\theta(S_t, \epsilon_t)$$
with ϵ_t a known density $p(\epsilon)$ and Φ an invertible transform ($\epsilon_t = \Phi_\theta^{-1}(S_t, A_t)$)
- Implicit parametrization of π_θ:
$$\pi_\theta(a|s) = |J_{\Phi_\theta^{-1}(s,a)}| p(\Phi_\theta^{-1}(s,a))$$
- Rewriting of the objective:
$$\begin{aligned} J(\theta) &= \mathbb{E}_{\pi_\theta}\left[\mathrm{KL}(\pi_\theta(\cdot|S_t), e^{\frac{1}{\lambda}Q_\phi(S_t,\cdot)}/Z(S_t,\theta))\right] \\ &= \mathbb{E}_{\pi_\theta,\epsilon}\left[-\log \pi_\theta(\Phi_\theta(S_t,\epsilon)|S_t) + \frac{1}{\lambda}Q_\phi(S_t, \Phi_\theta(S_t,\epsilon))\right] + Cst \end{aligned}$$
- Easy optimization when neglecting the effect of π_θ in S_t...
$$\begin{aligned} \hat{\tilde{\nabla}} J(\theta) = &-\nabla_\theta \log \pi_\theta(\Phi_\theta(S_t,\epsilon)|S_t) \\ &+ \nabla_\theta \Phi_\theta(S_t, \Phi_\theta(S_t,\epsilon))\left(-\nabla_a \log \pi_\theta(\Phi_\theta(S_t,\epsilon)|S_t) + \frac{1}{\lambda}\nabla_a Q_\phi(S_t, \Phi_\theta(S_t,\epsilon))\right) \end{aligned}$$

Double Q learning and extension

- Classical Q learning:
  - Target: $R_{s,a} + \max_{a'} Q_\phi(s', a')$
  - Approximation issue: $Q_\phi(s', a') \sim Q(s, a) + \epsilon(s, a)$
  - Consequence: $\mathbb{E}\left[\max_a Q_\phi(S_t, a)\right] \geq \max\left(Q(s, a) + \mathbb{E}\left[\epsilon(s, a)\right]\right)$
- Double Q learning:
  - Two Q learning function: $Q_{\phi_i}(s, a)$
  - Used in a crossed way for the target of Q_{ϕ_i}:
  $$R_{s,a} + Q_{\phi_{i'}}(s', \operatorname*{argmax}_{a'} Q_{\phi_i}(s', a'))$$
  - Mitigate the bias.
- Similar overstimation bias issue in actor critic approach.
- Clipped double Q learning:
  - Two Q learning function: $Q_{\phi_i}(s, a)$
  - Used in a pessimistic way for the target of Q_{ϕ_i}:
  $$R_{s,a} + \min_i Q_{\phi'_i}(s', a') - \lambda \log \pi_\theta(a'|s')$$
  as well as in the optimization on θ.

Replay Buffer

- In most of the algorithm, the expectation along trajectory is replaced by an empirical average over past short pieces of trajectories stored in a replay buffer.
- Corresponds exactly to the idea of an empirical simulator when the policy is fixed.
- If the policy is changing across time, we should use a importance sampling correction to be faithful with the theory...
- Not necessary for one-step Q learning but necessary for the actor-critic approach as the stationnary law on the state is used.
- Not an issue in practice: *neglecting the effect of* π_θ *in* S_t of the previous slides.
- To even reduce the issue: use only *recent* trajectories in the buffer.