Lecture 04

March 27, 2024

Revision so far, TRPO, PPO

$P(x_1), P(x_2) \ldots \ldots P(x_n)$

$H = -\sum_{x_i} P \log P$

Heads / tails

SAC

Usual use of entropy as regulariser

↓

Why not include in the reward directly?

$$J(\pi) = \sum_t^T E\left[r(s_t, a_t) + \alpha H[\pi(\cdot|s_t)]\right]$$

"soft value"

$$V(s_t) = E\left[Q(s_t, a_t) - \log \pi(a_t|s_t)\right]$$

Then update becomes, ($-\sum p \log p/q$)

$$\pi_{new} = \arg\min_{\pi' \in \Pi} D_{KL}\left[\pi'(\cdot|s_t) \,\|\, \frac{\exp(Q^{\pi_{old}}(s_t, \cdot))}{Z^{\pi_{old}}(s_t)}\right]$$

→ Q^π gets updated arbitrarily

→ π is parameterised, so we find a π_{new} within the family of distributions that minimises KL

Expand as $\pi_{new} = \arg\min E\left[\log \pi' - Q^{\pi_{old}} + \underbrace{\log Z^{\pi_{old}}}_{\text{constant}}\right]$

in the sense of policy evaluation

Parameterise by three networks,

$J_V(\psi)$, $J_Q(\theta)$, $J_\pi(\phi)$

all interdependent

Summary, motivation for trust region methods

Value based methods → Approximate $Q(s_t, a_t)$
- Easy to use
- Difficult to scale

Policy gradient → Direct optimisation of $V_\theta(s)$ (using $\nabla(\log \pi)(G_t)$)
- Gradients may be difficult to distinguish

Actor critics → Policy gradient with baseline "Advantage"
- Baseline provided by value function
  - $V_\theta(s)$
  - $Q_\theta(s, a)$

DDPG → Continuous control
- Removes need to compute expection over a_{t+1} in $r_t + E_{a_{t+1}}[Q(s_{t+1}, a_{t+1})]$
- Requires addition of noise at the end for exploration

SAC → Include entropy in the reward itself

Summary of challenges → A lot of hyperparameters to tune
- Can we guarantee monotonic improvement for a certain step size?

TRPO

→ Defines cost rather than reward
⇒ positive advantage is bad

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}\left[\sum \gamma^t A_\pi(s_t, a_t)\right]$$ (just rewriting)

$\eta(\tilde{\pi})$: cost function

$$= \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

this is hard to do, so replace by ρ_π

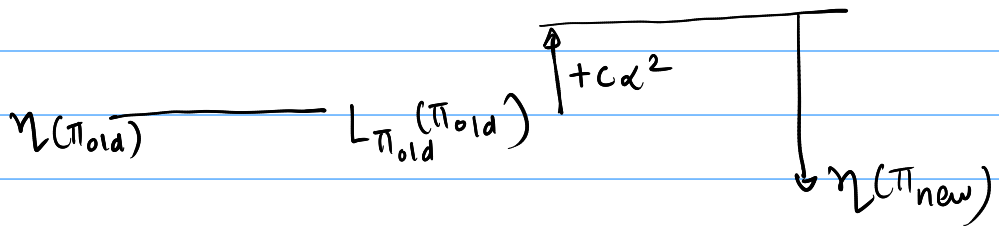
$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

Note: $L_\pi(\pi) = \eta(\pi) + 0$

Result in paper:

$$\eta(\pi_{new}) \leq L_{\pi_{old}}(\pi_{new}) + C\alpha^2$$

C: constant; α = mixture ratio



$$\eta(\tilde{\pi}) \leq L_\pi(\tilde{\pi}) + C D_{KL}^{max}(\pi, \tilde{\pi})$$

Define $M_i(\pi) = L_{\pi_i}(\pi) + C D_{KL}^{max}(\pi_i, \pi)$,

then $\eta(\pi_{i+1}) - \eta(\pi_i) \leq M_i(\pi_{i+1}) - M(\pi_i)$

$M_i(\pi_{i+1})$: minimise; $M(\pi_i)$: equal to $\eta(\pi_i)$

$\min[M_i(\pi)]$ is an unconstrained problem, but D_{KL}^{max} is difficult

↳ make it a constrained version

$$\min_\theta L_{\theta_{old}}(\theta) \quad \text{subject to} \quad D_{KL}^{max}(\theta_{old}, \theta) \leq \delta$$

still hard, so use sample average

After much ado:

replacement for advantage

$$\min_\theta \mathbb{E}\left[\frac{\pi_\theta(a|s)}{q(a|s)} Q_{\theta_{old}}(s,a)\right]$$

sampling

$\nabla L^T(\theta_{new} - \theta)$:

$\frac{1}{2}(\theta_{new}-\theta)^T F(\theta_{new}-\theta) \leq \delta$ subject to $\mathbb{E}_{s\theta_{old}}[D_{KL}(\pi_{\theta_{old}} \| \pi_\theta)] \leq \delta$

quad approx

$\partial\theta\partial\theta_i$

How to solve?

- first order approximation, no easy step size
- second order
  - need covariance matrix of gradients $\|\theta\| \times \|\theta\|$ sized
  - instead, use conjugate gradients

Conjugate gradients

$$\Delta\theta = \sqrt{\frac{2\delta}{\nabla L^T F(\theta)^{-1} \nabla L}} F^{-1} \nabla L$$

$\Delta\theta$ ↳ $(\theta_{new} - \theta)$

guess product directly

$x = f(x)$ $\quad x_1 = f(x_0)$ $\quad x_2 = f(x_1)$ $\quad x_{n+1} \approx f(x_n)$

PPO

How can we avoid conjugate gradients and all the other drama?

Recall $L - D_{KL}^{max}$ is the original objective, converted to $\min L$ subject to $D_{KL} \leq \delta$

$$L(\theta) = \mathbb{E}\left[\frac{\pi_\theta(a|s)}{q(a|s)} Q_\theta(s,a) \right]$$

$q(a|s)$: use $\pi_{\theta_{old}}$ as sampler

$Q_\theta(s,a)$: revert to A_t

conservative policy iteration

$$L^{CPI}(\theta) = \mathbb{E}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t \right]$$

$\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} = r_t$

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) \right]$$

We can even reintroduce D_{KL} in the objective (see eqn 8 in paper), so long as its coefficient is adaptively updated