(3) Policy gradients methods

So far: local view using Bellman equations

$Q^*(s,a) = \dots \quad Q^*(s',a')$

Different approach: global view

$$\max_\theta \mathop{\mathbb{E}}_{\pi \sim \sigma_\theta} \left[\underbrace{r_0 + \gamma r_1 + \gamma^2 r_2 + \dots}_{R(\pi|\theta)} \right]$$

actual objective

using (stochastic) gradient ascent

$$\max_\theta \mathop{\mathbb{E}}_{\pi \sim \sigma_\theta} \left[R(\pi|\theta) \right]$$

$$R(\pi|\theta) = \prod_t \underbrace{\Delta(r_t, s_{t+1} | s_t, a_t)}_{\text{independent of } \theta} \underbrace{\sigma_\theta(a_t|s_t)}_{\text{only dependence in } \theta}$$

fact: $\nabla_\theta f(x|\theta) = f(x|\theta) \nabla_\theta \log f(x|\theta)$

$\nabla_\theta R(\pi|\theta) = R(\pi|\theta) \nabla_\theta \log R(\pi|\theta)$

$\nabla_\theta R(\pi|\theta) = \sum_t \nabla_\theta \overbrace{\log}^{\text{log-probability ...}} \underbrace{\pi_\theta(a_t|s_t)}_{\text{probability of playing } a_t \text{ from } s_t}$

Lemma:

$$\nabla_\theta \mathop{\mathbb{E}}_{\pi\sim\pi_\theta}[R(\pi|\theta)] = \mathop{\mathbb{E}}_{\pi\sim\pi_\theta}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\pi|\theta)\right]$$

empirical loss: $\mathbb{E} \rightsquigarrow \hat{\mathbb{E}}$

Policy gradient algorithms

initialise model θ

iterate:

- batch sample
- estimate $\nabla_\theta \mathcal{L}(\theta)$ " $\nabla_\theta \mathbb{E}[R(\pi|\theta)]$
- update using ∇_θ: $\theta \leftarrow \theta + \alpha \nabla_\theta \mathcal{L}(\theta)$

Issue and improvement:

① $\nabla_\theta \mathbb{E}[R(\pi|\theta)] = \mathop{\mathbb{E}}_{\pi\sim\pi_\theta}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\pi|\theta)\right]$

given by PyTorch

[the whole episode
or
$r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots$

→ only the future !

replace $R(\pi|\theta) = \sum_t r^t r_t$ by $\sum_{t' \geq t} \gamma^{t'-t} r_{t'}$

"reward to go"

(2) Using "baselines"

$$b : S \to \mathbb{R}$$

lemma : $$\mathop{\mathbb{E}}_{a_t \sim \sigma_\theta} \left[\nabla_\theta \log \sigma_\theta(a_t|s_t) \; b(s_t) \right] = 0$$

So: we are free to choose b:

$$\nabla_\theta \mathbb{E}[R(\pi|\theta)] = \mathop{\mathbb{E}}_{\pi \sim \sigma_\theta} \left[\sum_t \nabla_\theta \log \sigma_\theta(a_t|s_t) \left(\sum_{t' \geq t} r_{t'} - b(s_t) \right) \right]$$

examples : $val(s)$

$q(s,a)$

$q(s,a) - val(s,a)$: advantage function

↳ REDUCE VARIANCE