Week 7 - Reinforcement Learning

GPI — Monday

grad ascent method

↳ for max

↳ of what?

$\nabla v_\gamma(\pi) \stackrel{def}{=} \frac{\partial v_\gamma(\pi)}{\partial \pi}$ ~→ ∇ / partial derivative

difficult to use / [illegible]

↳ gantinya / as a remedy → "explicit" policy parameterization

↳ we introduce a param vector

↳ $\theta \in \mathbb{R}^{\dim(\theta)}$, so $V(\pi(\theta)) \stackrel{def}{=} V(\theta)$

DATE / / NOTES

Why explicit? → Best Practice

① can utilize high dimension Repr. for policy

② can encode levels of exploration "share the same idea with ϵ-greedy"

③ cari yang lain for "exceeding esp".

The $\underbrace{\text{Grad Ascend update}}_{\text{Policy Gradient}}$ rule:

$$\theta^{K+1} \leftarrow \theta^{K} + \alpha \nabla V_{\pi}(\theta) \Big|_{\theta = \theta^{K}}$$

evaluated using θ^{K}

$\underbrace{K = \text{policy iteration index}}$

θ = initialized / choose randomly

α = choose random small number

$\nabla V_{\pi}(\theta) = ?$ → Sutton's policy grad

DATE / / OM OT OW OT OF OS OS NOTES

set $x \rightarrow$ gain

$$\nabla \vartheta_g(\theta) = \nabla \left\{ \sum_{s \in S} \sum_{a \in A} r(s,a)\, \rho^*_{\pi(\theta)}(s) \ldots \right.$$

$$\left. \ldots \pi(a|s;\theta) \ldots \right\}$$ by def of gain

$$= \sum_s \sum_a r(s,a) \left\{ \nabla \rho^*_\pi(s) . \pi + \rho^*_\pi(s) . \nabla \pi \right\}$$

$\rightarrow$ if we stop here, still its not sampling friendly ~~in~~ in RL, thus we use "score function".

$$= \sum_s \sum_a r(s,a) \left\{ \left(\rho^*_\pi . \nabla \log \rho^*_\pi . \pi \right) + \rho^* . \pi . \nabla_{\log} \pi \right\}$$

$$= \sum_s \sum_a \rho^*(s) . \pi(a|s) . r(s,a) \left\{ \nabla \log \rho^* + \nabla \log \pi \right\}$$

$\rightarrow$ sampling-friendly but in RL, ρ^* is still unknown to the agent.

DATE / / OM OT OW OT OF OS OS NOTES

Score function :

$$\nabla \log \pi(a|s;\theta) = \frac{\nabla \pi(a|s;\theta)}{\pi(a|s;\theta)}$$

$$= \sum_s \sum_a p^*(s)\, \pi(a|s)\, q_b^\pi(s,a)\, \nabla \log \pi(a|s;\theta)$$

$$= \mathbb{E}_{\substack{S \sim p^* \\ A \sim \pi(\cdot|s)}} \left[q_b^\pi(S|A)\ \nabla \log \pi(A|s;\theta) \right]$$

↳ exceeding exp

find proof!

p^* is unknown but appears under $\mathbb{E}$, hence in practice we simply sample from $p^*_\pi(\theta)$

↳ p^* exist, at $t \geq t_{mix}$

run unbiased approx of $\nabla V(\alpha)$.

but, at $t < t_{mix}$, $S \sim p^t \neq p^*$

hence biased approx. of $\nabla v(x)$

1. BCV : Reduce var w/o introducing bias error

$$\nabla v_g(\theta) = \mathbb{E}\Big[\underbrace{\{q_b^\pi(S,A) - \overbrace{v_b(S)}^{\text{baseline}}\}}_{\text{advantage function}} \nabla \log \ldots$$

$$\ldots \pi(A|S;\theta)\Big]$$

$= a_b^\pi(S,A)$

$$= \mathbb{E}\left[\delta_v(S, A, S') \, \nabla \log \pi(A|S;\theta)\right]$$

DATE / / OM OT OW OT OF OS OS NOTES

Week 7 – Reinforcement Learning

Wednesday.

Policy param for Discrete Action (focus: finite MDP)

using: $\underbrace{\text{Gibbs / Boltzman / ~~softmax~~ categorical}}_{\text{softmax parameterization}}$ dist

$$\pi(a|s;\theta) = \frac{\exp(\theta^T \phi(s,a))}{\sum \exp(\theta^T \phi(s,a))} \quad \forall (s,a) \in (S \times A)$$

Discounted Reward Policy Gradient

symbol for discounted

Recall:

$$v_\gamma(\pi) = P_\pi^\gamma \, r_\pi \qquad \Big| \qquad v_g(\pi) = P_\pi^{*} \, r_\pi$$

Where

$$P_\pi^\gamma = \lim_{t_{max}\to\infty} \sum_{t=0}^{t_{max}-1} (\gamma \cdot P_\pi)^t \qquad \Big| \qquad P_\pi^{*} = \text{(see Prev)}$$

↳ is not a proper distribution

because $\sum_{s=0} P_\pi^\gamma(s|s_0) \neq 1$

$$\underbrace{\nabla v_\gamma(\pi, s_0) = \sum_s P_\pi^\gamma(s|s_0) \sum_{a\in A} q_\gamma^\pi(s,a) \ldots \nabla \log \pi(a|s; \theta)}_{\text{not sampling friendly, not a distribution}}$$

times with $(1-\gamma)$

$$= \frac{1}{(1-\gamma)} \underbrace{\sum (1-\gamma) P_{\pi}^{\gamma}(s|s_0)}_{\text{becomes a geometric Distribution}} \sum_a \ldots$$

becomes a geometric Distribution

to make it back to equality.

$$\pi_g^* = \arg\max_{\pi \in \Pi} V_g(\pi, s_0)$$

$$\pi_\gamma^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{s_0 \sim \rho} \left[V_\gamma(\pi, s_0) \right]$$

ρ → initial state Distribution

π^{*}_{γ} depends on $\mathring{p}$

call non-uniform optimal