MRL 25/, 17/3/2025

GPI — inexact — poleval → given a policy π, what is $V(\pi)$? (determine its value)

GPI — inexact — polinpro → modify π^k such that $V(\pi^{k+1}) \geq V(\pi^k)$ ("better")

eg via grad ascend method

- ascend → for max, ie $\pi^*_x = \arg\max_{\pi \in \Pi_S} V_x(\pi)$, $x = \gamma, g, tot$
- grad → $\nabla V_x(\pi) = \frac{d}{d\pi} V_x(\pi)$, $\frac{d}{d\pi} \to \nabla$

pol grad in direct policy parametization

hard to use (not beneficial) —as remedy→ "explicit" pol param

→ where we introduce a param vector $\theta \in \Theta = \mathbb{R}^{\dim(\theta)}$

so, $V(\pi(\theta)) \stackrel{\text{def}}{=} V(\theta)$

Example
Let $\pi \in \Pi_{SR}$
Set $\pi \sim$ Gaussian ($\mu = 0$, $\sigma = \theta$)
(Gaussian = normal)

why explicit: - utilize hi-dim representation for policy
eg net → learning θ
SOTA
- encode levels of xploitation "share the similarity ide & greedy"

DATE :

Then the GA update rule:

$$\theta^{k+1} \leftarrow \theta^k + \alpha \nabla V_x(\theta)\Big|_{\theta=\theta^k}$$

- k: power idx
- α: positive step size
- $\theta = \theta^k$: evaluated using θ^k

Assume: $V_x(\theta)$ is differentiable

θ: init eg randomly

α: init eg

$\nabla V_x(\theta)$ = ? Big question → Sutton's policy grad

$$\nabla V_g(\theta) = \nabla \left\{ \sum_{s \in S} \sum_{a \in A} r(s|a) p^{\star}(s) \pi(a|s;\theta) \right\} \quad \text{by def of gain}$$

$$= \sum_s \sum_a r(s|a) \nabla \left\{ p^{\star}_{\pi(\theta)}(s) \cdot \pi(a|s;\theta) \right\}$$

$$= \sum_s \sum_a r(s|a) \left\{ \nabla p^{\star} \cdot \pi + p^{\star} \cdot \nabla \pi \right\}$$

Recall: Score fun: $\nabla \log \pi(a|s;\theta) = \frac{\nabla \pi(a|s;\theta)}{\pi(a|s;\theta)}$ } diff of log

$$= \sum_s \sum_a r(s|a) \cdot \left\{ p^{\star}_{\pi} \cdot \nabla \log p^{\star}_{\pi} \cdot \pi + p^{\star}_{\pi} \nabla \log \pi \right)$$

$$= \sum_s \sum_a p^{\star}(s) \pi(a|s) r(s|a) \left\{ \nabla \log p^{\star} + \nabla \log \pi \right\}$$

Simplify finally bcoz KL / $p^{\star}$ is unknown so again

DATE :

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

encode $\nabla \log p^*$ and $\Gamma(s,a)$

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

aka mean $M \approx$ Sample mean $\hat{M}$ (RV) which Unbiased $E[\hat{M}] = M$

here Stochastic Grad Asc $\uparrow$

$\rightarrow p^*$ is unknown but appears under E "hence in practice we simply Sample from $p^*\pi(\theta)$

Run unbiased approx of $\nabla V(x)$ ← p^* exist at $t \geq t_{max}$

- hence SGA

In contrast at $t < t_{max}$, $S \sim p^t \neq p^*$

hence biased approx of $\nabla V x$

↳ Semi-SGA

DATE :

① BCV → goal : reduce var w/o introducing bias-error

DATE :

Policy param for discrete Action (focus: finite MDP)
eg. using Gibbs / Boltzmann / Categorical distribution

$\pi(a|s;\theta) = \frac{\exp(\theta^T \phi(s,a))}{\sum_{a' \in A} \exp(\theta^T \phi(s,a'))}$

- policy param
- $\phi(s,a)$: state-action feature
- $\forall (s,a) \in (S \times A)$
- s, A: finite

$\pi(a|s;\theta)$: conditional prob ~~distribution~~ distribution over actions

both are correct

Recall:

$\pi: S \to \Delta^{|A|}$

eg $S^1 \to$ | a^0 | a^1 | a^2 |
|---|---|---|
| 0.3 | 0.6 | 0.1 |

$S^0 \to$ | 1 | 0 | 0 |
|---|---|---|

DATE :

Discanted Reward Policy Gradient

Recall

$V^{\pi}_{\gamma}(\pi) = P^{\gamma}_{\pi} r_{\pi}$ | $V_g(\pi) = P^{*}_{\pi} r_{\pi}$

Recall in gen : $V_g(\pi, s_0)$ constant over π all states

$$\nabla V^{\pi}_{\gamma}(\pi, s_0) = \sum_{s} \underbrace{P^{\gamma}_{\pi}(s|s_0)}_{\text{Gin chain}} \sum_{a \in A} q^{\pi}_{0}(s,a) \nabla \pi(a|s;\theta)$$

not sampling friendly because P^{γ} is not a distribution

$$\nabla V_{\gamma}(\pi, s_0) = \frac{1}{(1-\gamma)} \sum \underbrace{(1-\gamma) P^{\gamma}_{\pi}(s|s_0)}_{\text{becomes a distrib of state}} \sum_{a}$$

i.e. geometric distrib

How to sample from

DATE :

$\pi_g^* = \underset{\pi \in \Pi}{\operatorname{argmax}} V_g(x)$

$\pi_\gamma^* = \underset{\pi \in \Pi}{\operatorname{argmax}} E[V_\gamma(\pi, S_0)]$

$S_0 \sim \beta \rightarrow$ initial state distrib

π_γ^* depends on β