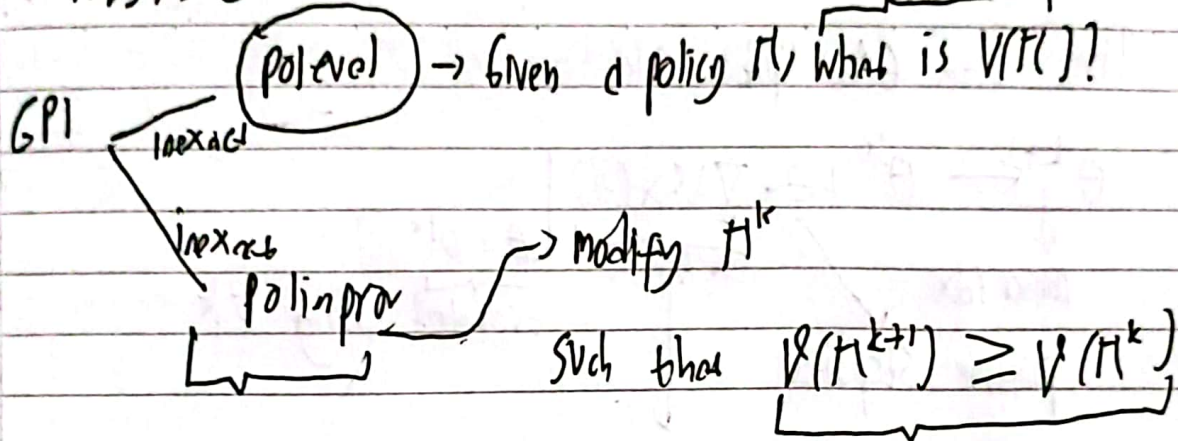
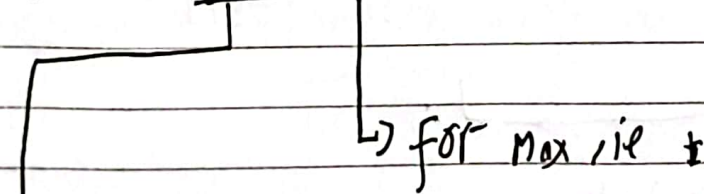


MRL 25/17/3/2025

determine its value
DATE



eg V_{π} grad ascend method



$$\pi_x^* = \operatorname{argmax}_{\pi \in \Pi_S} V_x(\pi)$$

$x = \delta, 1, 0, 1$

$$\nabla V_x(\pi) = \left(\frac{d}{d\pi} \right) V_x(\pi) \rightarrow \nabla$$

pol grad in direct policy parametrization

hard to use (not beneficial) $\xrightarrow{\text{as remedy}}$ "explicit" pl param

Example

Let $\pi \in \Pi_S$

Set $\pi \sim \text{Gaussian}(\mu=0, \sigma^2=1)$

normal

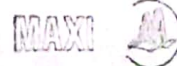
\rightarrow where we introduce a param vector $\theta \in \Theta = \mathbb{R}^{\dim(\Theta)}$

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

Why explicit: - utilize ~~high~~ hi-dim representation for policy

eg det \rightarrow learning

SOTA - encode levels of exploration "Share the similarity id & greed"



Then the GA update rule:

$$\theta^{k+1} \leftarrow \theta^k + \alpha \nabla V_{\pi}(\theta) \quad \left| \begin{array}{l} \theta = \theta^k \\ \text{evaluated using } \theta^k \end{array} \right.$$

\downarrow parameter
 \swarrow positive step size
 \downarrow

Assume: $V_{\pi}(\theta)$ is differentiable

θ : init eg randomly
 L : init eg

$\nabla V_{\pi}(\theta)$? Big question \rightarrow sub's policy grad

$$\begin{aligned} \nabla V_{\pi}(\theta) &= \nabla \left\{ \sum_{s \in S} \sum_{a \in A} r(s|a) p^*(s) \pi(a|s; \theta) \right\} \quad \left. \begin{array}{l} \text{by def of} \\ \text{gain} \end{array} \right\} \\ &= \sum_s \sum_a r(s|a) \nabla \left\{ p_{\pi}^*(s), \pi(a|s; \theta) \right\} \\ &= \sum_s \sum_a r(s|a) \left\{ \nabla p^* \cdot \pi + p^* \cdot \nabla \pi \right\} \end{aligned}$$

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

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

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

Sampling friendly
but in KL,
 p^* is unknown
to learn

$$= \sum_s \sum_0 p^*(s) H(A|s) \underbrace{q_b^n(s|A)}_{\text{exclude } \nabla \log p^* \text{ and } I(s|A)} \nabla \log H(A|s; \theta)$$

exclude $\nabla \log p^*$ and $I(s|A)$

$$= E_{\substack{S \sim p^*(\cdot) \\ A \sim H(A|S)}} [q_b^n(S|A) \nabla \log H(A|S; \theta)]$$

Ok, then $\bar{M} \approx \text{Sample Mean}$ which unbiased $E[\bar{M}] = M$

here stochastic grad ascent

p^* is unknown but appears under E "hence in practice we sample from $p^* H(\theta)$ "

Run unbiased approx of $\nabla V(x)$ $\leftarrow p^*$ exists at $t = t_{mix}$
 \hookrightarrow hence SGA

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

hence biased approx of $\nabla V(x)$
 \hookrightarrow Semi-SGA

DATE :

① BCV \rightarrow goal reduce var w/o introducing bias-error

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

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

$\phi(s, a)$: state-action feature
 $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$

action

prob ~~distribution~~ distribution
 over action

both are correct

Recall:

$$\pi(a|s) \rightarrow \Delta^{|A|}$$

eg $s^1 \rightarrow$

a^0	a^1	a^2
0.3	0.6	0.1

$s^0 \rightarrow$

1	0	0
---	---	---

Discounted Reward Policy Gradient

Recall

$$V_{\pi}(\pi) = \sum_{t=0}^{\infty} \gamma^t r_t \quad \bigg| \quad V_{\pi}(\pi) = \sum_{t=0}^{\infty} \gamma^t r_t$$

Recall in gen: $V_{\pi}(s_0)$ constant over π all states

$$\nabla V_{\pi}(\pi, s_0) = \sum_{\pi} p_{\pi}^{\pi}(s_1 | s_0) \sum_{a \in A} q_0^{\pi}(s_1, a) \nabla \pi(a | s_1; \theta)$$

Not sampling friendly
 bcos p_{π}^{π} is not a distribution

$$\nabla V_{\pi}(\pi, s_0) = \frac{1}{(1-\gamma)} \sum_{\pi} (1-\gamma) p_{\pi}^{\pi}(s_1 | s_0) \sum_a$$

becomes a distrib of state
 i.e. geometric distrib

How to sample from

$$V_g^* = \max_{\pi \in \Pi} V_g(x)$$

$$\pi_g^* = \arg \max_{\pi \in \Pi} E[V_g(\pi, S_0)]$$

$S_0 \sim p \rightarrow$ initial state distrib

π_g^* depends on p^0