# Policy Gradient - II

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_\tau P(\tau;\theta) R(\tau)$$

$$\approx \underline{\frac{1}{N} \sum_{i=1}^{N} \boxed{R(\tau_i)} \boxed{\nabla_\theta \log P(\tau_i;\theta)}} \qquad -(1)$$

where

$$R(\tau_i) = \sum_{t=0}^{T-1} r(s_t, a_t)$$

$$\nabla_\theta \log P(\tau_i;\theta) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | s_t)$$

There is Causality Violation issue

"Action in the future can not affect rewards in the past"

* Causality

For example, let's assume we have single reward, at j-th time step.

$$\nabla_\theta \mathbb{E}_{s_0, a_0, \ldots s_j \ldots s_{T-1}} [r_j] = \nabla_\theta \mathbb{E}_{s_0, a_0 \ldots s_j} [r_j]$$

future

the expectation stops at j-th term, all other terms cancel out.

Using the linearity of expectation, such as

$$\nabla_\theta \mathbb{E}_\tau \left[ \sum_{t=0}^{T-1} r(s_t, a_t) \right] = \nabla_\theta \sum_{t=0}^{T-1} \mathbb{E}_\tau [r(s_t, a_t)]$$

so, only causal terms matter,

$$\nabla_\theta \mathbb{E}_\tau [r(s_j, a_j)] = \mathbb{E}_{s_0, a_0 \ldots s_j, a_j}\left[ r(s_j, a_j) \sum_{t=0}^{j} \nabla_\theta \log \pi_\theta (a_t|s_t) \right]$$

the reward is only affected by actions that came before,

$$g_i = \sum_{t'=0}^{T-1} \underbrace{r(s_{t'}, a_{t'})}_{r_{t'}} \sum_{t=0}^{t'} \nabla_\theta \log \pi_\theta (a_t|s_t) \quad \text{— (2)}$$

$$g_i = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta (a_t|s_t) \left( \sum_{t'=t}^{T-1} r_{t'} \right) \quad \text{— (3)}$$

$r_{t'} \rightarrow r(s_{t'}, a_{t'})$

Simplified version

(2) → (3)

$$g_i = \sum_{t'=0}^{T-1} \underbrace{r(s_{t'}, a_{t'})}_{r_{t'}} \underbrace{\sum_{t=0}^{t'} \nabla_\theta \log \pi_\theta (a_t|s_t)}_{h_t}$$

$$\begin{array}{lllll}
 & t'=0 & r_0 h_0 & & \\
 & t'=1 & r_1 h_0 & + r_1 h_1 & \\
 & t'=2 & r_2 h_0 & + r_2 h_1 & + r_2 h_2 \\
 & \vdots & & & \\
+ & t'=T-1 & r_{T-1} h_0 & + r_{T-1} h_1 & \cdots \\
\hline
\end{array}$$

$$\left(\sum_{t=0}^{T-1} r_t\right) h_0 + \left(\sum_{t=1}^{T-1} r_t\right) h_1 + \cdots \left(\sum \ldots\right) h_{T-1}$$

$$\Leftrightarrow \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta (a_t|s_t) \left( \sum_{t'=t}^{T-1} r(s_{t'}, a_{t'}) \right)$$

## Quick Review.

$\boxed{\text{Causal}}$

$$g_i = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \Big( \sum_{t'=t}^{T-1} \underbrace{r_{t'}}_{r(s_{t'},a_{t'})} \Big)$$

$\boxed{\text{Non-Causal}}$

$$g_i = \Big( \sum_{t=0}^{T-1} r(s_t,a_t) \Big) \Big( \sum_{t'=0}^{T-1} \nabla_\theta \log \pi_\theta(a_{t'},s_{t'}) \Big)$$

$\boxed{\text{Gradient Update}}$

$$g \approx \frac{1}{N} \sum_{i=1}^{N} g_i$$

$$\theta_{k+1} \leftarrow \theta_k + \alpha g.$$

---

## REINFORCE algorithm (1992, Ronald Williams)

- Initialize policy $\theta_0$, and learning rate $\alpha$
- For $i=1$ : num_Iter :

    For $j=1$ : num_rollouts :

        Compute the grad. estimate $g_i = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \Big( \sum_{t'=t}^{T-1} r_{t'} \Big)$

    Estimate $g \approx \frac{1}{N} \sum_{i=1}^{N} g_i$

    Gradient update: $\theta_{k+1} \leftarrow \theta_k + \alpha g$.

What's "$\nabla_\theta \log \pi_\theta (a_t | s_t)$"?

$\pi_\theta$ : policy — Discrete action space / Continuous action space.

$\downarrow$

$\pi_\theta(a|s)$

* Diagonal Gaussian Policy.

↳ p-dimensional Gaussian Policy, $|A| = p$

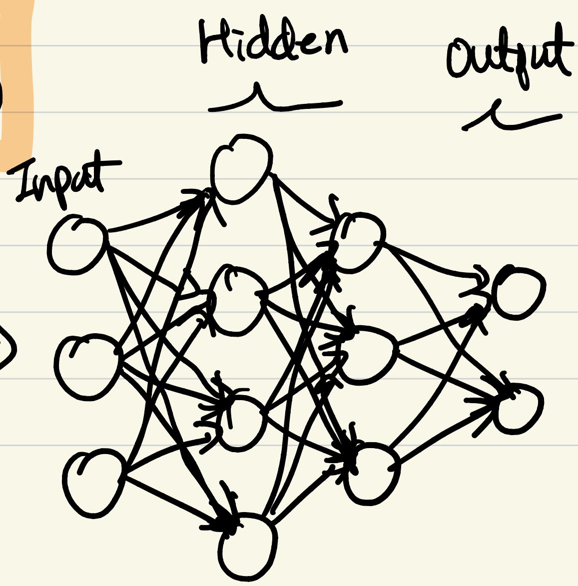
NN takes in states and outputs mean, $\mu \in \mathbb{R}^{|A|}$

and covariance, $\Sigma \in \mathbb{R}^{|A|}$

$$\pi(a|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp^{-\frac{1}{2}(a-\mu)^T \Sigma^{-1} (a-\mu)}$$

Where $\Sigma$ is covariance matrix, $\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \ \sigma_p^2 \end{bmatrix}$

$\boxed{\pi(\underline{a}|\underline{s})}$




States → (NN) → $\mu, \Sigma$ → Diagonal Gaussian policy → sample

$\nabla_\theta \log \pi_\theta(a_t | s_t)$:

- $$\pi(a | \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp^{-\frac{1}{2}(a-\mu)^T \Sigma^{-1} (a-\mu)}$$

- $$\log \pi_\theta(a | \mu, \Sigma) = -\frac{1}{2}\left(p \log(2\pi) + \log\left(\prod_i \sigma_i^2\right)\right) - \frac{1}{2}(a-\mu)^T \Sigma^{-1} (a-\mu)$$

- $$\nabla_\theta \log \pi_\theta(a | \mu, \Sigma) = -\frac{1}{2}\left(\nabla_\theta \log\left(\prod_i \sigma_i^2\right)\right) - \nabla_\theta \frac{1}{2}(a-\mu)^T \Sigma^{-1} (a-\mu)$$

↳ Automatic differentiation tool, i.e., tensorflow

Computes automatically.