Lec 07 a - Policy Optimization.

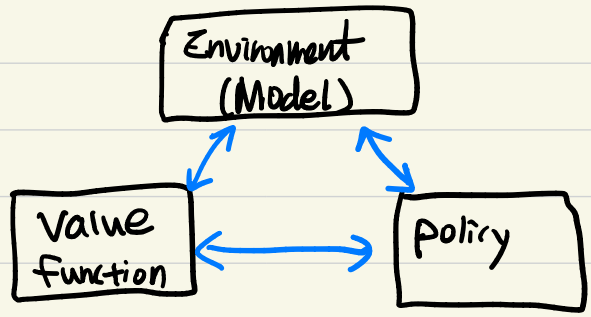
* where we are.

  ⓞ Markov Decision Process

  ⓞ Dynamic Programming
  - Learn value function
  - Implicit policy

  ⓞ Policy Optimization
  - No value function
  - Learn policy

    ex) Policy gradient



Formulation :

$$\max_{\theta} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \,\middle|\, \pi_\theta \right] \quad -(1)$$

(1) is finite-horizon discounted problem

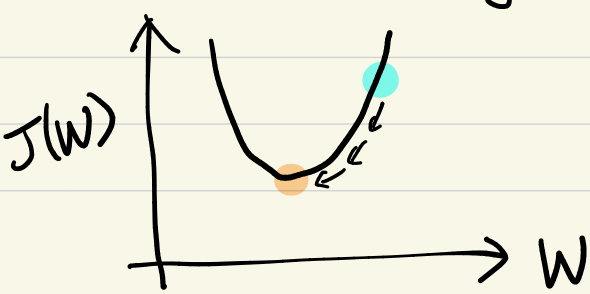
The goal is to maximize the return, R_t in an episodic setting.

parameterized policy $\pi_\theta(a|s)$ is used for stochastic policy.

Q: How to solve this optimization problem?

A: Gradient-based technique.

Gradient Desent Algorithm.



$$\text{"} \; \underbrace{W_{k+1} = W_k - \alpha \nabla_w J(\omega)}_{\text{update law.}} \; \text{"}$$

- gradient of objective fcn.

$$g = \nabla_\theta \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \,\Big|\, \pi_\theta\right]$$

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

- example, Direct Policy grad.

$$\mathbb{E}\left[\sum_{t=0}^{1} r(s_t, a_t) \,\Big|\, \pi_\theta\right] = \mathbb{E}\Big[\underbrace{r(s_0, a_0)}_{①} + \underbrace{r(s_1, a_1)}_{②} \,\Big|\, \pi_\theta\Big]$$

$$①: \nabla_\theta \mathbb{E}[r(s_0, a_0)] = \nabla_\theta \int \underbrace{r(s_0, a_0)}_{\text{RV}} \underbrace{\mu(s_0) \pi_\theta(a_0|s_0)}_{\text{prob.}} \, \underline{\underline{ds_0}}$$

$$②: \nabla_\theta \mathbb{E}[r(s_1, a_1)] = \nabla_\theta \int r(s_1, a_1) \pi_\theta(a_1|s_1) P(s_1|s_0, a_0) \mu(s_0) \pi_\theta(a_0|s_0) \, \underline{\underline{ds_0}} \, \underline{\underline{ds_1}}$$

Every subsequent term adds additional dimension of integration.
$\Leftrightarrow$ It's computationally intractable to compute gradient analytically

We are going to "<u>estimate</u>" the gradient "g".
$\Leftrightarrow$ policy gradient.