Policy Gradient Theorem:

problems: order a product more or less every day.

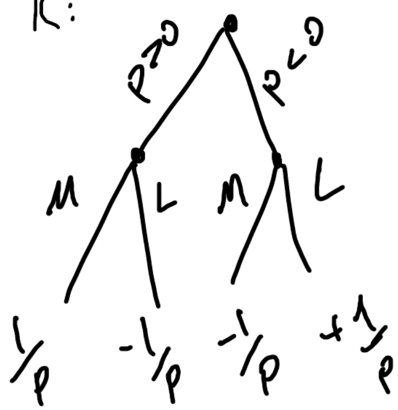
reward depends on p: Δ demand

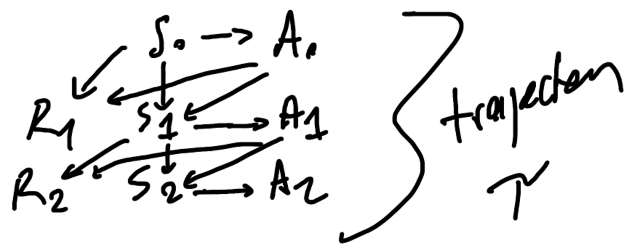
$a = \{ \text{order more}, \text{order less} \}$

s: { change in demand (or p) } (p) (if p → + more demand

if p → − less demand)

R:



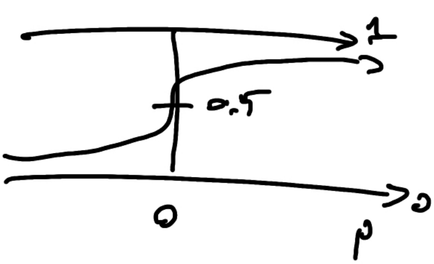


→ sigmoid

→ learn this one!

$P(\text{more}) = \sigma(\theta p)$

→ demand change



Mapping state of the world (p) to a the probability of taking a certain action.

↳ Policy (π)

find ideal θ in

$P(\text{More}) : \sigma(\theta_P)$

$J(\theta) = E(\theta)$ [total reward which we want to maximize over all τ]

$\frac{\partial J}{\partial \theta}$ → gradient ascent

$\frac{\partial J}{\partial \theta} = \frac{\partial}{\partial \theta} E_\theta(R)$

$\frac{\partial}{\partial \theta} \sum_\tau \pi_\theta(\tau) R(\tau)$

$= \sum_\tau \pi_\theta(\tau) \frac{\partial}{\partial \theta} [\log \pi_\theta(\tau)] R(\tau)$

log bc $\frac{\partial}{\partial \theta} \pi_\theta(\tau)$ is long.

$$\frac{\partial}{\partial \theta} \log \pi_\theta(\tau) : \cancel{\frac{\partial}{\partial \theta} \log P(S_0)}^{\,0} + \sum_{t=0}^{T} \frac{\partial}{\partial \theta} \log \underbrace{P(A_t | S_t)}_{\text{policy basically}}$$

$$+ \cancel{\sum_{t=0}^{\tau} \frac{\partial}{\partial \theta} \log P(R_{t+1}, S_{t+1} | S_t, A_t)}^{\,0}$$

no need to model the environment. model-free

$$E_\theta \left(R(\tau) \cdot \sum_{t=0}^{\tau} \frac{\partial}{\partial \theta} \log P(A_t \mid S_t) \right) \Rightarrow$$

$$\frac{\partial J}{\partial \theta} = p\, P(\text{loss})$$