Week 5 – RL 4 Blackboard 1

$$\langle R(\vec{\omega})\rangle = \underbrace{\sum_{\vec{x}} \sum_{y\in\{0,1\}}}_{\text{sum all possibilities}} \overset{\text{reward}}{R(y,\vec{x})} \cdot \underbrace{\Pi_{\vec{\omega}}(y|\vec{x}) \cdot P(\vec{x})}_{\text{statistical weight } P(y,\vec{x})=}$$

input output

policy depends on $\vec{\omega}$

$$= \sum_{\vec{x}} P(\vec{x}) \left[R(y=1,\vec{x}) \cdot g(\vec{\omega}\cdot\vec{x}) + R(y=0,\vec{x}) \cdot (1-g(\vec{\omega}\cdot\vec{x})\right]$$

derivative / batch update

$$\Delta\omega_{ji} = \alpha \cdot \frac{\partial}{\partial \omega_j}\langle R(\vec{\omega})\rangle = \sum_{\vec{x}} P(\vec{x}) \cdot g'(\vec{\omega}\cdot\vec{x})\left[R(y=1,\vec{x}) - R(y=0,\vec{x})\right] \cdot x_j$$

This is the correct batch update!
But it is not easy to go from here to online, because we do not have the correct statistical weight: we miss $\sum_y P_{\vec{\omega}}(y|\vec{x})$

aim: make statistical weight

$$\sum_{\vec{x}} \sum_{y} P_{\vec{\omega}}(y|\vec{x}) \cdot P(\vec{x})$$

visible! Must show up explicitly!

approach ① "pedestrian"

approach ② "log-likelihood trick"

Week 5 - RL4 — Blackboard 1B ("pedestrian")

$$(*)\quad \langle R\rangle = \sum_{\substack{\vec{x}\\ \text{input}}} \sum_{\substack{y\in\{0,1\}\\ \text{output}}} \overset{\text{reward}}{R(y,\vec{x})} \cdot \overbrace{\pi_w(y|\vec{x})\cdot P(\vec{x})}^{\text{statistical weight } P(y,\vec{x}) =}$$

policy depends on w

make "if-condition" explicit

$y=1$ if $y=0$

$$= \sum_{\vec{x}} \sum_{y} P(\vec{x}) \left[R(y=1,\vec{x})\cdot g(\vec{w}\cdot\vec{x})\cdot y + R(y=0,\vec{x})\cdot\left(1-g(\vec{w}\cdot\vec{x})\right)\cdot(1-y) \right]$$

if I add the sum over y, I need to add the "if-condition"!

take derivative and update (batch)

$$(2)\quad \Delta w_j = \alpha\cdot\frac{\partial}{\partial w_j}\langle R\rangle = \alpha \sum_{\vec{x}} \sum_{y} P(\vec{x}) \Big[\underbrace{R(1,\vec{x})\cdot g'(\vec{w}\cdot\vec{x})\cdot y}_{y=1 \text{ "if condition"}} - \underbrace{R(0,\vec{x})\cdot g'(\vec{w}\cdot\vec{x})\cdot(1-y)}_{y=0 \text{ condition}} \Big]\cdot x_j$$

online rule? need to make statistical weight explicit! Need $\pi(y|\vec{x})\cdot P(\vec{x})$!

---

use (1) $\pi_w(y|\vec{x}) = g(\vec{w}\cdot\vec{x})$ for $y=1$ and (2) $\pi_w(y|\vec{x}) = \left(1-g(\vec{w}\cdot\vec{x})\right)$ for $y=0$

$$\Delta w_j = \alpha\cdot\frac{\partial}{\partial w_j}\langle R\rangle = \alpha \sum_{\vec{x}} \sum_{y\in\{0,1\}} P(\vec{x}) \Big[\underbrace{\frac{\pi_w(y=1|\vec{x})}{g(\vec{w}\cdot\vec{x})}\cdot R(1,\vec{x})\cdot g'(\vec{w}\cdot\vec{x})\, y}_{y=1} - \underbrace{\frac{\pi_w(y=0|\vec{x})}{1-g(\vec{w}\cdot\vec{x})}\, R(0,\vec{x})\cdot g'(\vec{w}\cdot\vec{x})\cdot(1-y)}_{y=0} \Big]\cdot x_j$$

if condition

$$(3)\quad \Delta w_j = \text{"} = \alpha \overbrace{\sum_{\vec{x}} \sum_{y\in\{0,1\}}}^{\text{sum all possibilities}} \underbrace{P(\vec{x})\cdot\pi(y|\vec{x})}_{\text{statistical weight}}\, R(y,\vec{x})\cdot g'(\vec{w}\cdot\vec{x}) \left[\frac{y}{g(\vec{w}\cdot\vec{x})} - \frac{1-y}{1-g(\vec{w}\cdot\vec{x})} \right]\cdot x_j$$

online: cut statistical weight $\Rightarrow$ self-averaging over samples

$$\boxed{(4)\quad \Delta w_j = \alpha\, R(y,\vec{x})\cdot g'(\vec{w}\cdot\vec{x})\cdot\left[\frac{y}{g} - \frac{1-y}{1-g} \right]\cdot x_j}$$

Week 4: — Blackboard 2 log - likelihood trick

copy (1) $$\langle R \rangle = \sum_{\vec{x}} \sum_{y} R(y,\vec{x})\, \pi_\omega(y|\vec{x}) \cdot P(\vec{x})$$

$$\Delta w_j = \alpha \frac{\partial}{\partial w_j} \langle R \rangle = \alpha \sum_{\vec{x}} \sum_{y} R(y,\vec{x})\, P(\vec{x})\, \frac{\pi_\omega(y|\vec{x})}{\pi_\omega(y|\vec{x})} \frac{\partial}{\partial w_j} \pi_\omega(y|\vec{x})$$

$$= \alpha \sum_{\vec{x}} \sum_{y} \underbrace{P(\vec{x})\, \pi_\omega(y|\vec{x})}_{\text{statistical weight}} \cdot R(y,\vec{x}) \frac{\partial}{\partial w_j} \ln \pi_\omega(y|\vec{x})$$

online

(5) $$\boxed{\Delta w_j = \alpha \cdot \underbrace{R(y,\vec{x})}_{\text{reward}} \frac{\partial}{\partial w_j} \ln \underbrace{\pi_\omega(y|\vec{x})}_{\text{policy}}}$$ "log-likelihood trick"

Week 4 – Blackboard 3 : multi-step policy gradient

estimated return (total discounted future reward)

$$\hat{V}^{est}_{\pi}(s_t) = \langle r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots \rangle_{\text{all paths from } s_t}$$

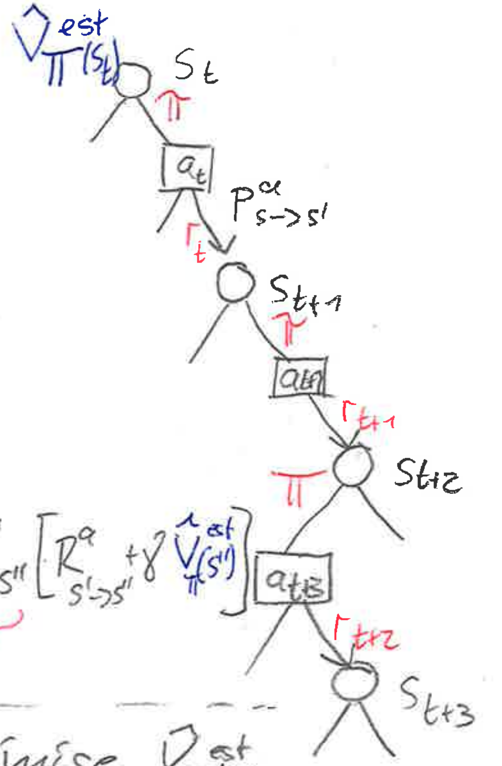
depends on policy

Bellman

$$= \underbrace{\sum_{a_t} \pi_\theta(a_t|s_t) \cdot \sum_{s'} P^{a_t}_{s_t \to s'}}_{\text{natural statistical weight}} \left[R^a_{s_t \to s'} + \gamma \cdot \hat{V}^{est}_{\pi}(s') \right]$$

expand

$$\underbrace{\sum_{a_{t+1}} \pi_\theta(a_{t+1}|s') \sum_{s''} P^{a_{t+1}}_{s' \to s''}} \left[R^a_{s' \to s''} + \gamma \hat{V}^{est}_{\pi}(s'') \right]$$



Change parameters θ of policy $\pi_\theta(a|s)$ so as to maximise $\hat{V}^{est}_{\pi}(s_t)$

$$\Delta\theta = \alpha \frac{\partial}{\partial\theta} \hat{V}^{est}_{\pi}(s_t) = \alpha \sum_{a_t} \pi_\theta(a_t|s_t) \cdot \frac{\partial}{\partial\theta} \ln \pi(a_t|s_t) \sum_{s'} P^{a_t}_{s_t \to s'} \left[R^a_{s_t \to s'} + \gamma \hat{V}^{est}_{\pi}(s') \right]$$

contains natural statistical weight

also depends on π

(product rule)

$$+ \alpha \sum_{a_t} \pi_\theta(a_t|s_t) \cdot \sum_{s'} P^{a_t}_{s_t \to s'} \gamma \cdot \frac{\partial}{\partial\theta} \hat{V}^{est}_{\pi}(s')$$

expand iteratively

online rule, drop statistical weight

$$\Delta\theta = \alpha \cdot \frac{\partial}{\partial\theta} \ln \pi(a_t|s_t) \underbrace{\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots \right]}_{R^{a_t}_{s_t \to s_{end}}}$$

"Return"

$$+ \alpha\gamma \cdot \frac{\partial}{\partial\theta} \ln \pi(a_{t+1}|s_{t+1}) \underbrace{\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \right]}_{R^{a_{t+1}}_{s_{t+1} \to s_{end}}}$$

$+ \alpha \cdot \gamma^2 \dots$

(4)