Week 10 – RL3 – Blackboard 1

statistical weight $P(y,\vec{x}) =$

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

policy depends on w

$y=1$ $y=0$

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

take derivative and update (batch)

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

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

use $\Pi_\omega(y|\vec{x}) = g(\vec{\omega}\cdot\vec{x})$ for $y=1$ and $\Pi_\omega(y|\vec{x}) = (1-g(\vec{\omega}\cdot\vec{x}))$ for $y=0$

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

if condition

(3) $$\Delta w_j = \text{"} = \alpha \sum_{\vec{x}} \sum_{y\in\{0,1\}} \underbrace{P(\vec{x}) \cdot \Pi(y|\vec{x})}_{\text{statistical weight}} R(y,\vec{x}) \cdot g'(\vec{\omega}\cdot\vec{x}) \left[\frac{y}{g(\vec{\omega}\cdot\vec{x})} - \frac{1-y}{1-g(\vec{\omega}\cdot\vec{x})} \right] \cdot x_j$$

online: cut statistical weight ⇒ self-averaging over samples

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

Week 10 — 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) $$\left\| \Delta w_j = \alpha \cdot \underset{\uparrow \text{ reward}}{R(y,\vec{x})} \frac{\partial}{\partial w_j} \ln \underset{\uparrow \text{ policy}}{\pi_\omega(y|\vec{x})} \right\|$$

"log-likelihood trick"

evaluate for our case: (Blackboard 2b/continued)

"if - condition"

$$\left.\begin{aligned} &\text{if } y=1 \quad \Pi_\omega(y=1|\vec{x}) = g(\vec{\omega}\cdot\vec{x}) \\ &\text{if } y=0 \quad \Pi_\omega(y=0|\vec{x}) = \left(1-g(\vec{\omega}\cdot\vec{x})\right) \end{aligned}\right\} \quad \Pi_\omega(y|\vec{x}) = g^{y}\cdot(1-g)^{(1-y)}$$

$$\Rightarrow \quad \ln \Pi_\omega(y|\vec{x}) = y\cdot \ln g + (1-y)\cdot \ln(1-g)$$

compare: log-likelihood

$$\Rightarrow \quad \frac{\partial}{\partial \omega_j} \ln_\omega \Pi(y|\vec{x}) = \frac{y}{g}\cdot g'\cdot x_j - \frac{(1-y)}{(1-g)}\cdot g'\cdot x_j$$

with (5)

$$\Delta \omega_j = \alpha\cdot R(y,\vec{x})\cdot g'\left[\frac{y}{g} - \frac{(1-y)}{1-g}\right]\cdot x_j \qquad \text{compare (4)}$$

evaluate further:

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

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

$$g = \langle y\rangle = 1\cdot \mathrm{Prob}(y=1) + 0\cdot \mathrm{Prob}(y=0)$$

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

Week 10 – Blackboard 3 : multi-step policy gradient

estimated return (total discounted future reward)

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

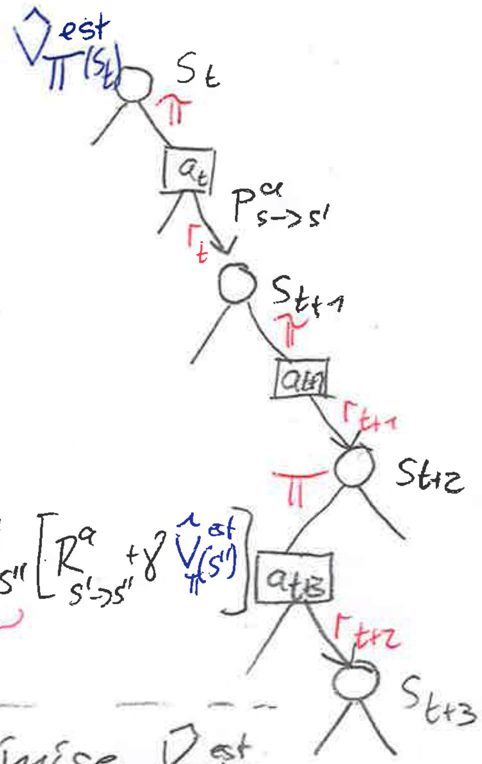
depends on policy

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

(product rule)

also depends on π

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

expand iteratively

online rule, drop statistical weight

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

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

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