use Center blackboard (1)

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 $\vec{\omega}$

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

leave space!

derivative / batch update

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

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

Blackboard rightside:

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 4: — Blackboard 2 log-likelihood trick

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

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

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

online

(5) $$\Delta\omega_j = \alpha \cdot \underset{\uparrow \text{reward}}{R(y,\vec{x})}\, \frac{\partial}{\partial \omega_j} \ln \underset{\uparrow \text{policy}}{\pi_\omega(y|\vec{x})}$$

"log-likelihood trick"

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

"if - condition"

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

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

compare: log-likelihood

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

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:

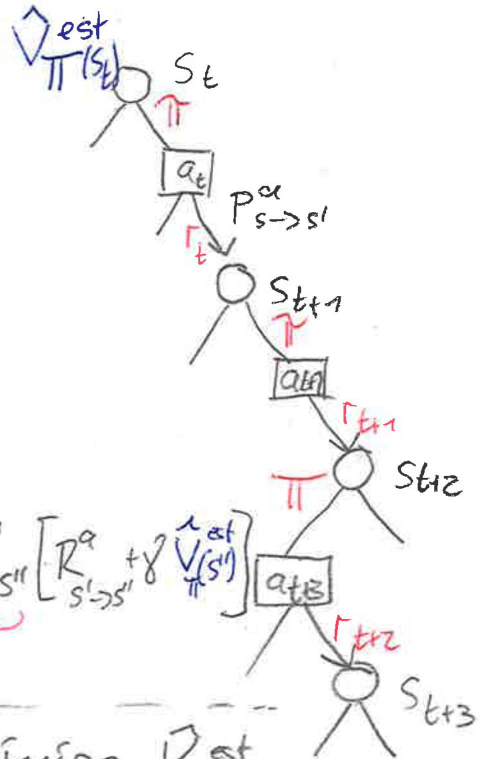
$$\Delta\omega_j = \alpha\cdot R(y,\vec{x})\,\frac{g'}{g\cdot(1-g)}\left[\cancel{(1-g)\cdot y} - \cancel{g\cdot(1-y)}\right]\cdot x_j$$

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

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

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

Week 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 $\hat V^{est}_{\pi}(s_t)$

contains natural statistical weight

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

also depends on π

(product rule)

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

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"

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

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