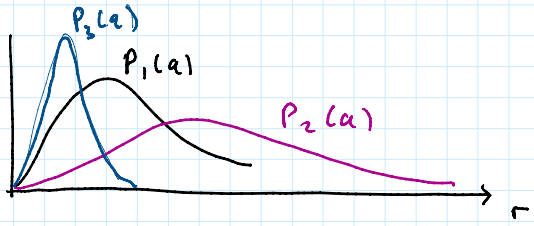
Monday, June 6, 2022 10:36 AM

One-armed bandits

choose $a \in \{1, \dots A\} =: \mathcal{A}$



goal: what policy $\pi(a)$ over actions maximises reward?

$$Q(a) := \mathbb{E}[R \mid A = a] = \mathbb{E}^{\mathbb{P}_a}[R]$$

- greedy approach:

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(a)$$

$R^{(1)}, R^{(2)}, R^{(3)}, \dots, R^{(n)}$

$a^{(1)}, a^{(2)}, a^{(3)}, \dots, a^{(n)}$

1.0	1.1	-5	+2	+1	10	-5	-1
1	1	3	2	2	3	3	2

$Q(1) \sim \frac{1}{2}(1.0 + 1.1) = 1.05 \leftarrow$

$Q(2) \sim \frac{1}{2}(2 + 1) = 1.5$

↳ $Q(2) = \frac{1}{3}(2 + 1 - 1) = \frac{2}{3} \leftarrow$

$Q(3) \sim \frac{1}{2}(-5 + 10) = 2.5$

↳ $Q(3) = \frac{1}{3}(-5 + 10 - 5) = 0 \leftarrow$

$$Q(a) = \frac{1}{\sum_{k=1}^{N} \mathbb{1}_{a^{(k)}=a}} \sum_{k=1}^{N} R^{(k)} \mathbb{1}_{a^{(k)}=a}$$

$$Q(a) = \frac{1}{N(a)} \sum_{k=1}^{N} R^{(k)} \mathbb{1}_{a^{(k)}=a}$$

$\hookrightarrow := \sum_{k=1}^{N} \mathbb{1}_{a^{(k)}=a}$

$a^* = \arg\max_a Q(a)$ — apply action — update

- ϵ-greedy.

$H^{(k)}$ - Bernoulli r.v. success prob $\epsilon_k \in (0,1)$

$\epsilon_1 > \epsilon_2 > \epsilon_3 > \cdots$

$\epsilon_k \xrightarrow[k\to\infty]{} 0$

e.g. $\epsilon_k = 1/k$

$\epsilon_k = \frac{c}{D+k}$

↳ r.v. uniform on $\{1, \ldots, A\}$

$$a^* = \mathbb{1}(H^{(1)} = 1)\, A^{(1)} + \mathbb{1}(H^{(1)} = 0) \arg\max_a Q^{(0)}(a)$$

suppose $Q^{(0)}(1) = Q^{(0)}(2)$

$a^{*(1)} \to R^{(1)} \to$ update $Q(a^{*(1)})$

$$\mu^{(n+1)} = \frac{1}{n+1} \sum_{k=1}^{n+1} Y^{(k)} = \frac{1}{n+1}\left(\sum_{k=1}^{n} Y^{(k)} + Y^{(n+1)}\right)$$

Y, $\underbrace{Y^{(1)}, \ldots, Y^{(n)}}$, $Y^{(n+1)}$, $\ldots$

$$= \frac{1}{n+1}\left(\mu^{(n)} n + Y^{(n+1)}\right)$$

$$\mu^{(n)} = \frac{1}{n} \sum_{k=1}^{n} Y^{(k)}$$

$$= \mu^{(n)}\left(\frac{n}{n+1}\right) + \frac{Y^{(n+1)}}{n+1}$$

↳ $1 - \frac{1}{n+1}$

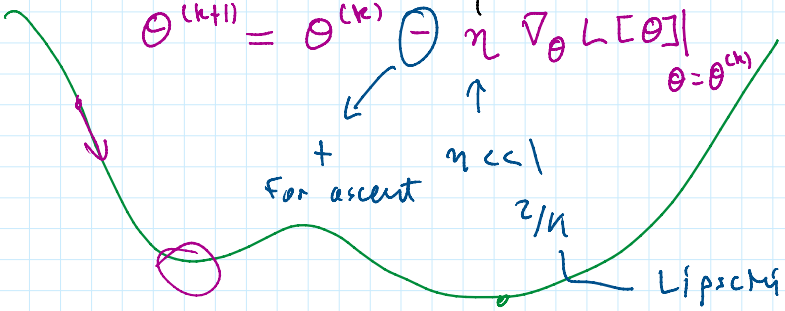
$$\mu^{(n+1)} = \mu^{(n)} + \frac{1}{n+1}\left(Y^{(n+1)} - \mu^{(n)}\right)$$

$\left(1 - \frac{1}{n+1}\right)^{n+1}$

$\pi_\theta(a)$ - distribution over actions parameterised by θ

$$Q(a) = \mathbb{E}^{\mathbb{P}_a}[R]$$

$$Q[\theta] = \int Q(a)\,\pi_\theta(a)\,da =: \mathbb{E}^{\mathbb{P}_\pi}[R]$$



$$Q[\theta] = \int Q(a)\,\pi_\theta(a)\,da = \mathbb{E}^{\mathbb{P}_\pi}[R]$$

e.g. $\pi_\theta(a) = p\,\phi\left(\frac{a-a_0}{\sigma_0}\right) + (1-p)\,\phi\left(\frac{a-a_1}{\sigma_1}\right)$

$a^{(1)} \sim \pi_\theta \rightarrow r^{(1)}$

$a^{(2)} \sim \pi_\theta \rightarrow r^{(2)}$

$\vdots \quad \vdots \quad \vdots$

$\frac{1}{N}\sum_{n=1}^{N} r^{(n)} \sim Q[\theta]$ unbiased estimator of

---

REINFORCE

$$\nabla_\theta Q[\theta] = \int Q(a)\,\nabla_\theta \pi_\theta(a)\,da$$

$= \mathbb{E}^{\mathbb{P}_\pi}[\ldots]$

$$\nabla_\theta Q[\theta] = \int Q(a)\, \nabla_\theta \pi_\theta(a)\, da$$

$$= \int Q(a) \left(\frac{\nabla_\theta \pi_\theta(a)}{\pi_\theta(a)} \right) \cdot \pi_\theta(a)\, da \qquad \text{likelihood ratio "trick"}$$

$$\hookrightarrow \nabla_\theta \log \pi_\theta(a)$$

$$= \int\int r\, \nabla_\theta \log \pi_\theta(a)\, \underline{p_a(r)\, \pi_\theta(a)}\, dr\, da$$

$$\mathbb{P}(r|a)\, \mathbb{P}_\theta(a) = \mathbb{P}_\theta(r)$$

$$= \mathbb{E}^{\mathbb{P}_\pi}\left[R\, \nabla_\theta \log \pi_\theta(a)\big|_{a=A} \right]$$

$$\boxed{\nabla_\theta Q[\theta] = \mathbb{E}^{\mathbb{P}_\pi}\left[R\, \nabla_\theta \log \pi_\theta(A) \right]}$$

$a^{(k)} \sim \pi_\theta(a) \rightarrow r^{(k)}$ ←

$$\nabla_\theta Q[\theta] \sim \frac{1}{N} \sum_{k=1}^{N} r^{(k)}\, \nabla_\theta \log \pi_\theta(a^{(k)})$$

$$\pi_\theta(a) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{-\frac{1}{2}\frac{(a-\bar{a})^2}{\sigma^2}} \qquad a|_\pi \sim \mathcal{N}(\bar{a}; \sigma^2)$$

$$\log \pi_\theta(a) = -\frac{1}{2}\left[\left(\frac{a-\bar{a}}{\sigma}\right)^2 + \log(2\pi\sigma) \right] \qquad \eta := \sigma^{-2}$$

$$(a-\bar{a})^2 \eta - \frac{1}{2}\log\eta + \log(2\pi)$$

$$\partial_{\bar{a}} \log \pi_\theta(a) = \left(\frac{a-\bar{a}}{\sigma}\right)$$

$$\partial_\eta \log \pi_\theta(a) = -\frac{1}{2}\left[(a-\bar{a})^2 - \frac{1}{2\eta} \right]$$

---

recall logistic $\mathbb{P}(Y=1 \mid X=x) = \rho^{\beta' x}$

recall logistic regression model

$$\mathbb{P}(Y=1 \mid X=x) = \frac{e^{\beta' x}}{1+e^{\beta' x}}$$

multi-class logistic regression

$$\mathbb{P}(Y=y \mid X=x) = \frac{e^{\beta_c' x}}{\sum_{c=0}^{K-1} e^{\beta_c' x}}$$

$y, Y \in \{0, \dots, K-1\}$

a canonical choice $\beta_0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

---

so take $\pi_\theta(a) = \frac{e^{H_\theta(a)}}{\sum_{a'} e^{H_\theta(a')}}$ (soft-max)

$$\nabla_\theta \log \pi_\theta(a) = \nabla_\theta H_\theta(a) - \nabla_\theta \log \sum_{a'} e^{H_\theta(a')}$$

$$= \nabla_\theta H_\theta(a) - \frac{\nabla_\theta \sum_{a'} e^{H_\theta(a')}}{\sum_{a'} e^{H_\theta(a')}}$$

$$= \nabla_\theta H_\theta(a) - \frac{\sum_{a''} e^{H_\theta(a'')} \nabla_\theta H_\theta(a'')}{\sum_{a'} e^{H_\theta(a')}}$$

$H_\theta : \mathcal{A} \to \mathbb{R}$

$H_\theta(a)$ =

$0 \longrightarrow H_\theta[0] = \gamma_0$

$1 \longrightarrow H_\theta(1) = \gamma_1$

$2 \longrightarrow H_\theta(2) = \gamma_2$

$$\partial_{\gamma_k} \log \pi_\theta(a) = \mathbb{1}_{a=k} - \frac{\sum_{a''} e^{H_\theta(a'')} \mathbb{1}_{a''=k}}{\sum_{a'} e^{H_\theta(a')}}$$

$$= \mathbb{1}_{a=k} - \left(\frac{e^{H_\theta(k)}}{\sum_{a'} e^{H_\theta(a')}} \right)$$

$$\boxed{\partial_{\gamma_k} \log \pi_\theta(a) = \mathbb{1}_{a=k} - \pi_\theta(k)}$$

$$\boxed{\partial_{\gamma_k} Q[\theta] = \mathbb{E}^{\mathbb{P}_\pi}\left[R \left(\mathbb{1}_{A=k} - \pi_\theta(k) \right) \right]}$$

$$\gamma_k^{(n)} \longrightarrow \gamma_k^{(n)} + \eta \underbrace{\frac{1}{M} \sum_{m=1}^{M} r^{(m)} \left(\mathbb{1}_{a^{(m)}=k} - \pi_\theta(k) \right)}_{Y_k^{(M)}}$$

$$Y_k^{(1)} = r^{(1)} \left(\mathbb{1}_{a^{(1)}=k} - \pi_\theta^{(1)}(k) \right)$$

$$Y_k^{(n+1)} = Y_k^{(n)} + \frac{1}{n+1} \left(\underbrace{r^{(n+1)} \left(\mathbb{1}_{a^{(n+1)}=k} - \pi_\theta^{(n)}(k) \right)}_{} - Y_k^{(n)} \right)$$