Bayesian Bandit:

- Bayesian measure: using the observation data to establish the belief of the hyperparameters of a distribution.

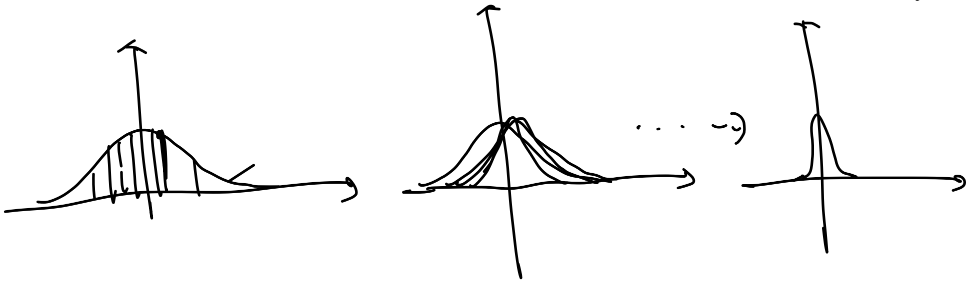
e.g. $X \sim N(\mu, \sigma)$

μ ← unknown.

observation: $x_1, x_2, \cdots, x_n$

Goal: Based on the observation, derive a belief for μ.

$\mu \sim p(\mu)$



the prob of observing x for μ

prior measure

$$p(\mu|x) = \frac{P(x|\mu)}{\underbrace{P(x)}_{\text{normalizing constant}}} p(\mu) = \underbrace{\frac{1}{Z} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(\mu-\mu_i)^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{\text{a new prob measure for } \mu}$$

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(\mu-\mu_i)^2}{2\sigma_i^2}}$$

At the beginning of round i,

we have prior measure

$\mu \sim N(\mu_{i-1}, \sigma_{i-1})$

After one observation x_i,

we update the posterior measure $\mu \sim N(\mu_i, \sigma_i)$

$$\mu_i = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_{i-1}^2}\right)^{-1} \left[\frac{x_i}{\sigma^2} + \frac{\mu_{i-1}}{\sigma_{i-1}^2}\right]$$

$$\sigma_i^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_{i-1}^2}\right)^{-1}$$

or equivalently, before any observations, with prior measure $N(\mu_0, \sigma_0)$

& after n observations $X_1, \cdots, X_n$, we have empirical mean $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

Then the posterior measure of $u \sim N(\mu_n, \sigma_n)$ will be

$$\mu_n = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \left[\frac{n}{\sigma^2}\bar{u} + \frac{1}{\sigma_0^2}\mu_0\right]$$

$$\sigma_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

- Bayesian for Binomial distribution $\begin{cases} 1 & w.p.\ p \\ 0 & w.p.\ 1-p. \end{cases}$

$p \sim \underbrace{Beta(a_0, b_0)}_{\text{prior}}$, $\rightarrow \mathbb{E}[p] = \dfrac{a_0}{a_0+b_0}$

After n observations, we can update the posterior measure for p:

$p \sim \underbrace{Beta(a_n, b_n)}$ ← posterior.

$$\begin{cases} a = a_0 + \sum_{j=1}^{n} x_j \\ b = n - \sum_{j=1}^{n} x_j + b_0 \end{cases} \Rightarrow \mathbb{E}[p] = \frac{a_0 + \sum_{j=1}^{n} x_j}{n + a_0 + b_0}$$

or equivalently

→ with prior $p \sim Beta(a_{i-1}, b_{i-1})$ & observation X_i

posterior $p \sim Beta(a_{i-1} + X_i,\ 1 - X_i + b_{i-1})$

· Thomson Sampling

e.g. $X_0 = \frac{1}{2}$ 1-armed reward.

$X_1 \sim$ Binomial (p) unknow   start with $p \sim$ Beta $(1, 1)$

At the beginning of round t, we have prior measure of $p \sim$ Beta $(a_{t-1}, b_{t-1}) \rightarrow$ sample from Beta $(a_{t-1}, b_{t-1}) \rightarrow p_t$

- If $p_t > \frac{1}{2} \Rightarrow$ pull unknown arm

  $\Rightarrow$ update the posterior measure of p based on the observation X_t.
  $$a_t = a_{t-1} + X_t$$
  $$b_t = 1 - X_t + b_{t-1}$$

- If $p_t < \frac{1}{2} \Rightarrow$ pull known arm

  $$\Rightarrow \begin{cases} a_t = a_{t-1} \\ b_t = b_{t-1} \end{cases}$$

HW $\rightarrow$ Try TS for $\underbrace{\text{Normal}}_{\text{2-armed}}$ reward & Bionomial reward with different Δ & differe prior.

Can we do better than TS based on the posterior measure if we know how many rounds we left ?

- Bayesian optimal policy

$X_0 = \frac{1}{2}$

$X_1 \sim \text{Binomial}(p)$ ← unknow

Q: At the beginning of round n, I have prior measure for $p \sim \text{Beta}(a_{n-1}, b_{n-1})$, what is the best policy according to this posterior ?

$$\mathbb{E}[X_n] = \mathbb{E}[p] = \frac{a_{n-1}}{b_{n-1}+a_{n-1}} \quad \bigcirc \quad \frac{1}{2}$$

" > " arm 1

" < " arm 0 .

e.g. $a_{n-1} = b_{n-1} = 1$, then it is the same to pull unknown arm or known arm

Define $V(a,b,l)$ as the optimal expected cumulative reward with n rounds left.

$$V(a,b,1) = \max\left\{ \frac{a_{n-1}}{b_{n-1}+a_{n-1}} , \frac{1}{2} \right\}.$$

Q: At the beginning of round $n-1$, I have prior measure for $p \sim \text{Beta}(a_{n-2}, b_{n-2})$, what is the best policy according to this posterior ?

If arm 0 is pulled

$$V^2(a_{n-2}, b_{n-2}, 2) = \frac{1}{2} + V(a_{n-2}, b_{n-2}, 1)$$

If arm 1 is pulled:

$$V^1(a_{n-2}, b_{n-2}, 2) = \frac{a_{n-2}}{b_{n-2}+a_{n-2}} + \mathbb{E}[\, V(a_{n-1}, b_{n-1}, 1)\,]$$

The posterior measure of p after pulling arm 1 at round $n-1$

If arm 1 is pulled at round $n-1$:

$$\begin{cases} 1 & \text{w.p. } p \rightarrow \text{posterior } p \sim Beta(a_{n-2}+1, b_{n-2}) \\ 0 & \text{w.p. } 1-p \rightarrow \text{posterior } p \sim Beta(a_{n-2}, b_{n-2}+1) \end{cases}$$

$$V^1(a_{n-2}, b_{n-2}, 2) = \frac{a_{n-2}}{b_{n-2}+a_{n-2}} + \frac{a_{n-2}}{b_{n-2}+a_{n-2}} \cdot V(a_{n-2}+1, b_{n-2}, 1) + \frac{b_{n-2}}{b_{n-2}+a_{n-2}} V(a_{n-2}, b_{n-2}+1, 1)$$

$$V(a, b, 2) = \max\{V^1, V^2\}$$

e.g. If $a_{n-2} = b_{n-2} = 1$

$$V^1(a,b,2) = \frac{1}{2} + \frac{1}{2} V(2,1,1) + \frac{1}{2} V(1,2,1)$$

$$= \frac{1}{2} + \frac{1}{2}\left(\max\left\{\frac{2}{3}, \frac{1}{2}\right\} + \max\left\{\frac{1}{3}, \frac{1}{2}\right\}\right)$$

$$= \frac{1}{2} + \frac{1}{2}\left(\frac{2}{3} + \frac{1}{2}\right) = \frac{1}{2} + \frac{1}{2}\cdot\frac{4+3}{6} = \frac{1}{2} + \frac{7}{12} = \frac{13}{12}$$

$$V^2(a,b,2) = \frac{1}{2} + V(1,1,1) = \frac{1}{2} + \max\left\{\frac{1}{2}, \frac{1}{2}\right\} = 1$$

If 2 rounds left, even if the expectation is the same for 2 arms, it will automatically prefer to explore the unknown arm.

In general: $V(a,b,n)$

$$= \max\left\{ \frac{1}{2} + V(a,b,n-1),\ \frac{a}{a+b} + \frac{a}{a+b} V(a+1,b,n-1) + \frac{b}{a+b} V(a,b+1,n-1) \right\}$$

with $V(a,b,0) = 0$

Q: what's the complexity to obtain the optimal Bayesian policy for MAB with horizon n?

$O(n^3)$

If it is k-armed bandit prob, $\Rightarrow O(n^{3k})$.

Q How to reduce the computational cost

1). Derive a limiting HJB eqn $\Rightarrow$ independ of n

" Continuous-in-time limit for Bayesian Bandits. Z - Izzo - Ying JMLR 23 "

$t = \frac{i}{n}$, $\hat{a} = \frac{1}{n}a$, $\hat{b} = \frac{1}{n}b$, $\hat{V}(\hat{a},\hat{b},t) = \frac{1}{n}V(a,b,i)$, $a = n\hat{a}$, $b = n\hat{b}$

$a+1 = n\hat{a}+1 = n(\hat{a}+\frac{1}{n})$.

$$\hat{V}(\hat{a},\hat{b},t) = \max\left\{\frac{1}{2}\frac{1}{n} + \hat{V}(\hat{a},\hat{b},t-\frac{1}{n}),\ \frac{1}{n}\frac{\hat{a}}{\hat{a}+\hat{b}} + \frac{\hat{a}}{\hat{a}+\hat{b}}\hat{V}(\hat{a}+\frac{1}{n},\hat{b},t-\frac{1}{n}) + \frac{\hat{b}}{\hat{a}+\hat{b}}\hat{V}(\hat{a},\hat{b}+\frac{1}{n},t-\frac{1}{n})\right\}$$

$$\frac{\hat{V}(\hat{a},\hat{b},t) - \hat{V}(\hat{a},\hat{b},t-\frac{1}{n})}{\frac{1}{n}} = \max\left\{\frac{1}{2},\ \frac{\hat{a}}{\hat{a}+\hat{b}} + \frac{1}{\frac{1}{n}}\frac{\hat{a}}{\hat{a}+\hat{b}}\left[\hat{V}(\hat{a}+\frac{1}{n},\hat{b},t-\frac{1}{n}) - \hat{V}(\hat{a},\hat{b},t-\frac{1}{n})\right] + \frac{1}{\frac{1}{n}}\frac{\hat{b}}{\hat{a}+\hat{b}}\left[\hat{V}(\hat{a},\hat{b}+\frac{1}{n},t-\frac{1}{n}) - \hat{V}(\hat{a},\hat{b},t-\frac{1}{n})\right]\right\}$$

$$\frac{\hat{V}(\hat{a},\hat{b},t) - \hat{V}(\hat{a},\hat{b},t-\frac{1}{n})}{\delta_t} = \frac{1}{2} + \max\left\{0,\ \frac{\hat{a}}{\hat{a}+\hat{b}} - \frac{1}{2} + \frac{\hat{a}}{\hat{a}+\hat{b}}\frac{\hat{V}(\hat{a}+\frac{1}{n},\hat{b},t-\frac{1}{n}) - \hat{V}(\hat{a},\hat{b},t-\frac{1}{n})}{\delta_a} + \frac{\hat{b}}{\hat{a}+\hat{b}}\frac{\hat{V}(\hat{a},\hat{b}+\frac{1}{n},t-\frac{1}{n}) - \hat{V}(\hat{a},\hat{b},t-\frac{1}{n})}{\delta_b}\right\}$$

Define function $\hat{V}(a,b,t)$ with $t \in (0,1)$, then

as $\delta_t, \delta_a, \delta_b \to 0$

$$\partial_t \hat{V} = \frac{1}{2} + \max\left\{0,\ \underbrace{p(\hat{a},\hat{b})}_{=\frac{\hat{a}}{\hat{a}+\hat{b}}} - \frac{1}{2} + p(\hat{a},\hat{b})\,\partial_{\hat{a}} V + (1-p(\hat{a},\hat{b}))\,\partial_{\hat{b}} V\right\}$$

$$\partial_t \hat{V} = \frac{1}{2} + \max_{\pi\in[0,1]}\left(p(\hat{a},\hat{b}) - \frac{1}{2} + \hat{p}(\hat{a},\hat{b})\,\partial_{\hat{a}} V + (1-p(\hat{a},\hat{b}))\,\partial_{\hat{b}} V\right)\pi(\hat{a},\hat{b},t)$$

exact solution or numerical solution (indep of n).

open prob: ① what theoretical guarantee can we have for this approximated solu?

② for learned bandit with no exact solution for HJB eqn, any method to find the solu efficiently?

$p = \frac{a+a_0}{b+b_0}$

approximated Bayesian-optimal policy. (obtain HJB solution $\hat{V}(x,y,t)$).

at round i, you arrive at state a, b

- $t = \frac{n-i}{n}$, $x = \frac{a}{n}$, $y = \frac{b}{n}$
- $\hat{V}(x,y,t) \Rightarrow \frac{x}{x+y} - \frac{1}{2} + \frac{x}{x+y} \partial_x \hat{V} + \frac{y}{x+y} \partial_y \hat{V} = f$
- If $f > 0 \Rightarrow$ pull unknown arm
  - observe reward 1, $a \leftarrow a+1$, $b \leftarrow b$
  - observe reward 0, $a \leftarrow a$, $b \leftarrow b+1$

  If $f < 0 \Rightarrow$ pull known arm

  $i \leftarrow i-1$

Infinite horizon with discount factor γ

$$V^{\lambda}(a,b) = \max\left\{ \lambda + \gamma V(a,b),\ \frac{a}{a+b} + \gamma\left[\frac{a}{a+b}V(a+1,b) + \frac{b}{a+b}V(a,b+1)\right]\right\} \quad (*)$$

$s \in \mathbb{Z}_+^2 = \binom{\alpha}{\beta}$; $A = \{1,0\}$; $S_1^{A=1} = \begin{cases} \binom{\alpha+1}{\beta} & \text{w.p. } \frac{\alpha}{\alpha+\beta} = P(s_1 \mid s_0, A=1) \\ \binom{\alpha}{\beta+1} & \text{w.p. } \frac{\beta}{\alpha+\beta} = P(s_1=() \mid \cdots) \end{cases}$; $r(1,s) = \frac{\alpha}{\alpha+\beta}$; $r(0,s) = \lambda$

$S_1^{A=0} = S_0$

$$V(s) = \max_a \left\{ r(0,s) + \gamma V(s),\ r(1,s) + \gamma\left[P(s_1=()\mid\cdots)V(\cdots) + P(s_1=()\mid\cdots)V(\cdots)\right]\right\}$$

$$V(s) = \max_a \left\{ r(s,a) + \gamma \mathbb{E}\left[V(s_1) \mid s_0 = s, A_0 = a\right]\right\}$$

$$= \max_{\sum_a \pi_a \in [0,1]} \left[\sum_{a\in A} \left(r(s,a) + \gamma\mathbb{E}\left[V(s_1) \mid s_0 = s, A_0 = a\right]\right)\pi_a\right]$$

$$= \max_{\pi} \mathop{\mathbb{E}}_{\substack{a_0 \sim \pi(a|s) \\ s_1 \sim P(s|a_0,s_0)}} \left[r(s,a) + \gamma V(s_1) \mid s_0 = s\right]$$

for fixed s, a prob distribution for a.

Naively, one can cut the infinite horizon to finite horizon & then derive the soln backwards, but as $\gamma \to 1$, the horizon will be larger.

open prob: ? Can we derive a similar limiting PDE by rescaling the parameters

K-armed prob:

$$V(\alpha_1, \beta_1, \cdots, \alpha_k, \beta_k) = \max_{j} \left\{ \frac{\alpha_j}{\alpha_j+\beta_j} + V(\cdots \alpha_j+1, \beta_j, \cdots) \frac{\alpha_j}{\beta_j+\alpha_j} + V(\cdots \alpha_j, \beta_j+1, \cdots) \frac{\beta_j}{\alpha_j+\beta} \right\}$$

$$|r|<1 \Rightarrow \sum_{t=T}^{\infty} \gamma^t r(s_t, a_t) \le \frac{\gamma^T}{1-\gamma} \le \varepsilon \Rightarrow T \sim O\left(\frac{1}{1-\gamma}\right)$$

$$O\left(\left(\frac{1}{1-\gamma}\right)^{2k}\right)$$

* Gittins index:

- when $s \in \mathbb{Z}_+^{2k}$, Instead of viewing it as a coupled system, one decouple it by giving each arm $\underbrace{\text{an equivalent value}}_{\text{Gittins index}}$ at their current state

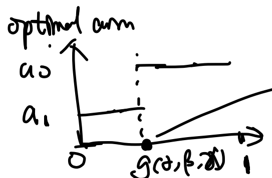
  Let $g(\alpha, \beta, \gamma)$ be the Gittins index for Binomial reward with prior measure $\text{Beta}(\alpha, \beta)$ & discount factor γ,

  $$\text{optimal policy} = \arg\max_{k \in [K]} \{ g(\alpha_k, \beta_k, \gamma) \}$$

- View K-armed Bandit as K one-armed bandit prob.

  For one-arm bandit prob with prior measure $\text{Beta}(\alpha, \beta)$

  v.s. known arm with deterministic reward λ.



$\rightarrow g(\alpha, \beta, \gamma)$ measures the value of this arm

λ.

The Gittins index is the deterministic arm reward such that it is equivalent to pull the known arm & unknown arm

One algo to calculate $g(s)$

Initialization: set $\bar{g}$ & $\underline{g}$.

while $\bar{g} - \underline{g} > \varepsilon$ $\Rightarrow$ set $g = \frac{\bar{g} + \underline{g}}{2}$

open prob ③ may reduce the cost of this step.

solve the optimal policy for one-armed bandit with an unknown arm with state s & known arm with $r = g$

If the optimal policy is to pull the unknown arm

$\underline{g} = g$

→ known arm

$\bar{g} = g$.

(*) open prob ④ For a couple HJB, can we decouple it with similar idea?