policy = solve(<u>solver</u>, <u>m</u>) offline calculation

a = action(policy, s) online calculation

k-armed Bernoulli Bandits

θ $p(w)$

$$p(\theta_i \mid o_1, o_2, \dots o_{t-1}) = \text{Beta}$$

loop

pull arm i based on policy $\leftarrow$

observe o_t, get r_t

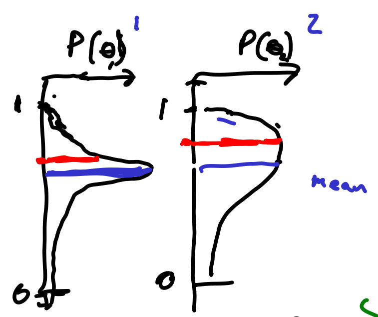
$$P(\theta_i \mid o_1, o_2, \dots o_{t-1}) = \text{Beta}(w+1, l+1)$$

$$\hat{Q}_i = \frac{w}{w+l} = \hat{\theta} \qquad \operatorname*{argmax}_i \hat{Q}_i$$

ad hoc

- ε-greedy: w.p. $1-\varepsilon$ choose $\operatorname*{argmax}_i \hat{Q}_i$
  w.p. ε choose random
- softmax

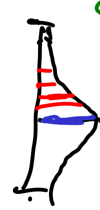
$$P(i) = \frac{e^{\alpha \hat{Q}_i}}{\sum_i e^{\alpha \hat{Q}_i}}$$



$\hat{\theta} + 1$ standard dev

- Interval
  - UCB

$$\operatorname*{argmax}_i \hat{Q}_i + c\sqrt{\frac{\ln \sum_j N_j}{N_i}}$$



(not expensive)

Bayesian

Expensive

- Thompson Sampling
  sample θ_j from $P(\theta_j \mid o_1 \dots o_t)$
  choose $\operatorname*{argmax}_i \theta_i$

Optimal Policies for Bandits Finite Horizon h

MDP $\quad P(\theta \mid \ldots) = \mathrm{Beta}(w+1, l+1)$

$S = \{(w_1, l_1, \ldots w_n, l_n)\}_{w, l \in 1..h}$

exponential in h

$R(s, i, s') = \begin{cases} +1 & \text{if } w_i \text{ was incremented} \\ 0 & \text{o.w.} \end{cases}$

$A = \{1 \ldots k\}$

T = increment win or loss according to belief probabilities

$P(o_t = 1 \mid s) = \int p_{\mathrm{Beta}(\ )}(\theta)\, \theta \, d\theta$

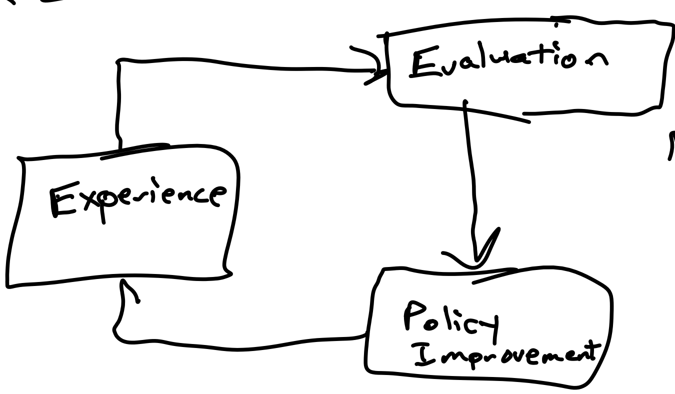
Gittins Index

Exploration + Exploitation ← Bandits
Credit Assignment
Generalization

RL

Offline or Online

often used in offline fashion with simulator



Model Based: 1. Learn MDP model from experience
2. Solve MDP

Model Free: Learn value or policy directly from experience

On-Policy

Off-Policy: Able to improve policy without new experience from that policy

Model-Based RL

Max Likelihood

$N(s,a,s')$, $\rho(s,a)$

$N(s,a) = \sum_{s'} N(s,a,s')$

$T(s'|s,a) = \dfrac{N(s,a,s')}{N(s,a)}$

$R(s,a) = \dfrac{\rho(s,a)}{N(s,a)}$

loop

Choose a based on exp. strat.

Observe s', r

$N(s,a,s')$ ++

$\rho(s,a)$ += r

update T, R (Value Iteration)

update Q $\Leftarrow$ expensive

$s \leftarrow s'$