Another broad class of policies for MDPs are index policies

- Suppose state and action spaces decompose as $S = S_1 \times S_2 \times \dots \times S_k$ and $A = \{a_1, a_2, \dots, a_k\}$. Then an index policy comprises a set of fns $\Phi_1, \Phi_2, \dots, \Phi_m$ with $\Phi_i : S_i \to \mathbb{R}$ (indices) s.t. $A(s) = \arg\max_{k \in [m]} \{\Phi_i(s_i)\}$ for $s = (s_1, s_2, \dots, s_m)$
- In other words, for each 'part' of the state, we compute a fn, and then 'act on the part' with the highest fn value

Eg - (single machine scheduling with discounting)

- There are m jobs, with each job i having known processing time t_i and reward r_i upon completion. Want to schedule them on a single machine to maximize discounted sum of rewards (discount factor β)
- If $S_k \subseteq [m] \equiv$ set of remaining jobs, then
$$V(S_k) = \max_{j \in S_k} \left[r_j \beta^{t_j} + \beta^{t_j} V(S_k \setminus j) \right]$$
- If $m = 2$, $V(\{1,2\}) = \max\left[r_1 \beta^{t_1} + r_2 \beta^{t_1+t_2},\ r_2 \beta^{t_2} + r_1 \beta^{t_1+t_2} \right]$
  $\Rightarrow$ we first serve $\arg\max_{i \in \{1,2\}} \{ r_i \beta^{t_i} / 1-\beta^{t_i} \}$
- This problem has an easy soln via an interchange argument
  Suppose order is $i_1 \to i_2 \to \dots \to j \to k \to \dots \to i_m$
  $\Rightarrow$ Reward is of form $R_1 + \beta^{T+t_j} r_j + \beta^{T+t_j+t_k} r_k + R_2$
  by previous argument, comparing $i_1 \to i_2 \to \dots \to j \to k \to \dots \to i_m$ and $i_1 \to i_2 \to \dots \to k \to j \to \dots \to i_m$ is same as comparing $r_j \beta^{t_j} / 1-\beta^{t_j}$ and $r_k \beta^{t_k} / 1-\beta^{t_k}$
  $\Rightarrow$ OPT policy $\equiv$ serve jobs in decreasing order of $r_j \beta^{t_j} / 1-\beta^{t_j}$

Bayesian Multi-Armed Bandits

- Now we consider a vast generalization of the above
  - There are m 'arms', where each arm i is a Markov Chain $X_i[t]$ on state S_i.
  - In each round, we can play a single arm, ie, $A = [m]$
  - If $A[t] = i$, then $R(t) = R_i(X_i[t])$ and $X_i[t] \to X_i[t+1]$
  - (Non-restlessness) The state of arm i changes only when it is 'played' (ie, $X_i[t+1] = X_i[t]\ \forall i$ s.t $A[t] \neq i$)
  - (Discounted infinite-horizon objective) $\max \sum_{t=0}^{\infty} \beta^t \left(\sum_{i=1}^{n} \mathbb{1}_{\{A[t]=i\}} R_i(X_i[t])\right)$

- Eg (MAB with Beta-Bernoulli priors)-
  - m actions, where action i gives reward $Ber(P_i)$
  - P_i unknown, but assume $P_i \sim Beta(N_i, S_i)$ (prior)
  - If we play action k, then posterior $= \begin{cases} Beta(N_i+1, S_i+1) & \text{if } R_i = 1 \\ Beta(N_i+1, S_i) & \text{if } R_i = 0 \end{cases}$
  - Aim: $\max \sum_{t=0}^{\infty} \beta^t \left(\sum_{i=1}^{m} \mathbb{1}_{\{A[t]=i\}} R_i[t]\right)$

The 1.5 arm problem - Suppose we just have 2 arms
1) MC $X[t]$ with reward $R_i(X_i[t])$
2) Constant reward arm with reward γ

Now opt policy is a stopping problem: play arm 1 till some time τ (stopping time), then play 2 forever (since no new info)

$$\Rightarrow \quad R = \sup_{\tau \geq 0} \mathbb{E}\left[\sum_{t=0}^{\tau-1} R_i(X_i[t]) \beta^t + \beta^\tau \gamma / 1-\beta\right]$$

The Gittins Index of arm 1 in state $X_1[0] = 2$ is the smallest constant reward γ s.t. you are indifferent between playing arm 1 in state $X_1[0]$ and arm 2

• Formally for any $x \in S_1$

$$\boxed{\gamma_1(x) = \sup\left\{\gamma : \frac{\gamma}{1-\beta} \le \sup_{\tau>0} \mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t R_1(X_1[t]) + \frac{\beta^\tau \gamma}{1-\beta} \,\middle|\, X_1[0]=x\right]\right\}}$$

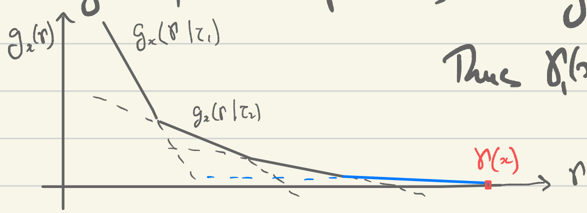
This can also be viewed as the minimum per-round charge for pulling arm 1 s.t. you are indifferent between pulling once or not when $X_1[0] = x$

$$\gamma_1(x) = \sup\left\{\gamma : 0 \le \underbrace{\sup_{\tau>0}\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t [R_1(X_1[t]) - \gamma] \,\middle|\, X_1[0]=x\right]}_{g_x(\gamma)}\right\}$$

Note that $g_x(\gamma)$ is decreasing and convex in γ

- To see this, note that for a fixed sample path $x_1 x_2 \ldots$ and fixed τ, $\sum_{t=0}^{\tau-1}\beta^t[\mathbb{E}[R_1(x_t)] - \gamma]$ is linear decreasing
- Taking expectation preserves linearity, and sup over τ makes it convex

Thus $\gamma_1(x)$ has a unique soln



Also for the optimal τ, we have $\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t R_1(X_1[t]) - \gamma_1(x)\sum_{t=0}^{\tau-1}\beta^t \,\middle|\, X_1[0]=x\right] = 0$

Thus

$$\boxed{\gamma_1(x) = \sup_{\tau>0,\ \text{stopping time}} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t R_1(X_1[t]) \,\middle|\, X_1[0]=x\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t \,\middle|\, X_1[0]=x\right]}}$$

Gittins Index for arm 1 in state x

Eg - Suppose $\underbrace{X_1[0]}_{\text{unknown}} = X_1[1] = X_1[2] = \dots = \begin{cases} M & \text{wp } p \\ 0 & \text{ow} \end{cases}$ 'Collapsing' arm (play once to learn)

$$R_1(x) = x$$

Then $\gamma(\text{'unknown'}) = \sup\left[\gamma \;\middle|\; \gamma/1-\beta \leq pM/1-\beta + (1-p)\beta\gamma/1-\beta\right]$

$$= \frac{pM}{1-\beta(1-p)}$$

Eg - Single job with reward r_i, processing time t_i

"amount of job processed"

$$R_i(0) = \sup_{\tau>0} \frac{r_i \beta^{t_i} \mathbb{1}\{\tau \geq t_i\}}{\sum_{t=1}^{\tau-1} \beta^t} = \underbrace{\left(\frac{r_i \beta^{t_i}}{1-\beta^{t_i}}\right)}_{\text{the index we got via interchange}}(1-\beta)$$

Thm (Gittins '79) - For finite arms $[m]$, and bounded rewards $R_i(x) \in [-c, c]$ $\forall i \in [m]$, $x \in S_i$:
A policy is optimal if and only if it always selects arm i at time t with highest Gittins index $\gamma_i(X_i[t])$.

There are actually more general conditions for when an index policy is optimal. When is it not, though?

- Independence of irrelevant alternatives (IIA): A policy π satisfies IIA if for any set of arms $[m]$ and $i \in [m]$, if $\pi([m]) = i$, then $\pi([n]) \neq i$ for any $[n] \supseteq [m]$.

An index policy is optimal if and only if opt policy is IIA.