Another broad class of policies for MDPs are index policies.

- Suppose state and action spaces decompose as $S = S_1 \times S_2 \times \ldots \times S_k$ and $A = \{a_1, a_2, \ldots, a_k\}$. Then an index policy comprises a set of functions $Q_1, Q_2, \ldots, Q_m$ with $Q_i : S_i \rightarrow \mathbb{R}$ (indices)
  
  s.t. $A(s) = \arg \max_{k \in [m]} \{Q_i(s_i)\}$ for $s = (s_1, s_2, \ldots, s_m)$

- 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 $\beta$)

- If $S_m = [m] = \text{set of remaining jobs}$, then
  
  $V(S_m) = \max_{j \in S_m} [r_j \beta t_j + \beta t_j \max_j V(S_\setminus j)]$

- If $m = 2$ 
  
  $V(1,2) = \max [r_1 \beta t_1 + r_2 \beta t_2, r_2 \beta t_1 + r_1 \beta t_2]$ 
  
  $\Rightarrow$ we first serve any $i \in \{1,2\}$, $\Rightarrow$ we first serve any $\arg \max_{i \in \{1,2\}} \{ r_i \beta t_i / (1 - \beta t_i) \}$

- This problem has an easy soln via an interchange argument.

Suppose order is $i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow j \rightarrow k \rightarrow \ldots \rightarrow i_m$

$\Rightarrow$ Reward is of form $R_1 + \beta t_j + \beta t_i + \beta t_{j+1} \beta t_{k+1} + R_2$

by previous argument, comparing $i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow j \rightarrow k \rightarrow \ldots \rightarrow i_m$ and $i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow j \rightarrow i_m \rightarrow k \rightarrow \ldots \rightarrow i_1$, $\Rightarrow$ OPT policy = 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 $X_i \in \{\mathcal{S}_i\}$. Each state is
  - In each round, we can play a single arm, i.e., $d = \{m\}$
  - If $A[t] = i$, then $R(t) = R_i(X_i[t])$ and $X_i[t] \rightarrow X_i[t+1]$
  - (Non-restlessness) The state of arm $i$ changes only when it is ‘played’ (i.e., $X_i[t+1] = X_i[t]$ \forall i \neq A[t]$
  - (Discounted infinite-horizon objective) $\max \sum_{t=0}^\infty \beta^t (\mathbf{1}_{A[t] = i} R_i(X_i[t]))$

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

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

No optimal policy is a stopping problem: play arm 1 till some time $\tau$ (stopping time), then play 2 forever (since no new info)
$\Rightarrow R = \sup_{\tau \geq 0} \mathbb{E} \left[ \sum_{t=0}^\tau R_i(X_i[t]) \beta^t + \beta^\tau \frac{R}{1-\beta} \right]$

The Gittins Index of arm 1 in state $X_i[0] = 2$ is the smallest constant reward $\gamma$ s.t. you are indifferent between playing arm 1 in state $X_i[0]$ and arm 2
Formally, for any $x \in S$,

$$\gamma(x) = \sup \left\{ \eta : \frac{\Lambda(x)}{1-\beta} \leq \sup_{t > 0} \mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t R_t(X_{\{t\}}) + \frac{\beta^t \tau}{1-\beta} | X_{\{0\}} = x \right] \right\}$$

This can also be viewed as the minimum per-bound charge for pulling arm 1 s.t. you are indifferent between pulling once or not when $X_{\{0\}} = x$.

$$\gamma(x) = \sup \left\{ \eta : 0 \leq \sup_{t > 0} \mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t [R_t(X_{\{t\}}) - \eta] | X_{\{0\}} = x \right] \right\}$$

Note that $g_\eta(\theta)$ is decreasing and convex in $\theta$.

To see this, note that for a fixed sample path $x_1, x_2, \ldots$ and fixed $\tau$, $\sum_{t=0}^{\infty} \beta^t [\mathbb{E}[R_t(x_t)] - \eta]$ is linearly decreasing.

Taking expectation preserves linearity, and sup over $\tau$ increases it.

Thus, $\gamma(x)$ has a unique solution.

Also, for the optimal $\tau$, we have $\mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t R_t(X_{\{t\}}) - \gamma(x) \sum_{t=0}^{\infty} \beta^t | X_{\{0\}} = x \right] = 0$.

Thus,

$$\gamma(x) = \sup_{\tau > 0, \text{stopping time}} \frac{\mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t R_t(X_{\{t\}}) | X_{\{0\}} = x \right]}{\mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t | X_{\{0\}} = x \right]}$$

Gittins index for arm 1 in state $x$. 

\[\blacksquare\]
Eg - Suppose $X[0]=X, [1]=X, [2]=\ldots = \{M \text{ w.p.} p \text{ 'collapsing'} \}
\text{known}
R_i(x) = x
\text{(play once to learn)}

Then $g$ ('unknown') = \sup \left\{ g \mid g/1-\beta \leq pM/1-\beta + (1-p)B/1-\beta \right\}
= \frac{pM}{1-\beta (1-p)}

\text{Eg - Single job with reward } r_i, \text{ processing time } t_i
\text{amount of job processed}
R_i(0) = \sup_{\tau > 0} \frac{r_i \beta^\tau}{\sum_{k=1}^{t_i} \beta^k} = \left( \frac{r_i \beta^{t_i}}{1-\beta^{t_i}} \right) (1-\beta)
\text{The index we get via interchange}

\text{Thm (Gittins '79) - For finite arms } [m], \text{ and bounded rewards } R_i(x) \in [-c,c] \forall i \in [m], \forall x \in S_i
\text{A policy is optimal if and only if it always selects arm } i \text{ at time } t \text{ with highest Gittins index } R_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 $\Pi$ satisfies IIA if for any set of arms $[m]$ and $i \in [m]$, if $\Pi([m]) = i$, then $\Pi([n]) = i$ for any $[n] \supset [m]$. An index policy is optimal if and only if optimal is IIA.