Dr. David A. Goldberg

**Multi-arm bandits**
December 8, 2017

# 1 Bayesian bandits and the Gittins index

## 1.1 Motivation

Many fundamental trade-offs that arise in a society can be framed as the question of whether to stick with an option that is known to perform well, or explore a more risky option that may perform better. This is known as the exploration-exploitation trade-off, i.e. selecting between exploring new options which may be better, or continuing to exploit perhaps good enough options already known to perform well. To make it concerete in Ph.D. student-land: suppose your advisor has you working on a research project, which is steadily yielding decent results. A paper seems within reach. However, a very interesting new research problem came across your advisor's desk, which she has now passed on to you. Should you spend your time continuing on the project you already know yields a steady stream of decent results, or devote valuable time to exploring the new research question which may yield even more interesting results, but for which progress is much more uncertain. Suppose you spend a day on the new problem and get nowhere. What should you do the next day? How many days on the project is too many? If you switch back to the original safe project, will it ever again make sense to give another go at the new and riskier problem? What if you are juggling many such projects? What is the optimal strategy to use in conducting your own research? Similar scenarios arise in many settings, including exploring the benefit of new drugs, the potential revenue of new products, the potential reward of unexplored alternatives for an AI-driven robot, etc.

In this class we will see several different ways to formalize the above problem mathematically. The solution methodologies and ideas we use for the different formulations will be related, but also have some important differences. We begin with the so-called Bayesian formulation for the multi-arm bandit (MAB) problem. There are a fixed number, say N, of different alternatives. Time is discrete, and in every period you must select one of the alternatives. For any alternative $i$, there is a reward distribution $\mathcal{R}^i$, where we let $R^i$ denote a r.v. distributed as $\mathcal{R}^i$. Suppose that every time you select alternative $i$, you get a reward drawn independent and identically distributed (i.i.d.) from $\mathcal{R}^i$. Of course, if one knew the reward distributions, and was trying to maximize the expected reward earned over time, one would simply in each period always select the alternative with the highest expected payout (breaking ties arbitrarily).

Let us think back on our Ph.D. research example, and make it a little more precise. For simplicity, we will only concern ourselves with number of new (sufficiently interesting) results proven each day. Suppose that for the project which is steadily yielding decent results, you have collected lots of data on your own progress, and derived the following simple model for your progress on that problem. Any day on which you work on the steady project, with probability (w.p.) $\frac{1}{6}$ you prove 0 new results, w.p. $\frac{2}{3}$ you prove 1 new result, and w.p. $\frac{1}{6}$ you prove 2 new results, and you are very confident in the validity of this model for your progress on that problem. Alternatively, you do not yet have any data on how you perform on the new research problem. Maybe it will turn out that you are very good at that type of problem, but then again maybe you will not be good at that type of problem. Perhaps you have some experience working on related problems. For example, suppose the new problem is very combinatorial in nature, in particular it is about polyhedral combinatorics. Suppose you have worked on several different problems in combinatorics, e.g. some problems in graph theory and other areas, but never any problems in polyhedral combinatorics. Suppose you have data on your performance on past combinatorial problems, and reason as follows. In about $\frac{1}{2}$ of the combinatorial research problems you have ever worked on, you were able to really internalize the problem and make great progress steadily. In about $\frac{1}{4}$ of the combinatorial research problems you have ever worked on, you were able to really internalize the problem, but due to unforeseen fundamental difficulties in the problem progress is not steady, instead most days you make no progress while every once in a while you have a Eureka day and prove many results. Alternatively, in about $\frac{1}{4}$ of the combinatorial research problems

you have ever worked on, the problem was essentially hopeless and you steadily made no progress at all. More precisely, suppose your data on your own ability to solve problems revevaled the following. For any given combinatorial problem (given no other useful data on your ability to solve the problem), w.p. $\frac{1}{2}$ each day you work on the problem you prove 1 result w.p. 1, independently across days (situation I); w.p. $\frac{1}{4}$ each day you work on the problem you prove 0 results w.p. $\frac{2}{3}$, and 5 results w.p. $\frac{1}{3}$ (situation II); and w.p. $\frac{1}{4}$ you never prove any results on any of the days (situation III). Note that what we have here is a **distribution on distributions**. In particular, you know (under our modeling assumptions) that each day you work on this new problem the number of results you prove is drawn i.i.d. from some distribution which is innately fixed at time 0 by some inherent notion of your innate compatability with the given problem. Namely, the actual distribution of the number of results you prove each day you work on this new problem is fixed and not changing over time. However, this fixed distribution is itself drawn randomly - and you don't know which distribution is fixed and governing your research progress.

## 1.2 Bayesian probability: an example

So what (operationally) does Bayesian mean here? Suppose you decide to "roll the dice", and spend a day on the new research problem. Suppose you prove zero results. The fact that you now have the additional knowledge that you proved zero results will change your estimates for the different probabilities governing which "type" of combinatorial problem this is (for you). Your initial beliefs here are generally referred to as your prior, and your beliefs after a new observation(s) are generally called your posterior. Here, the prior probabilities were $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. Using Bayes rule, we compute the posterior probabilities as

$$\begin{aligned}\Big(P(I|0), P(II|0), P(III|0)\Big) &= \left(\frac{P(I,0)}{P(0)}, \frac{P(II,0)}{P(0)}, \frac{P(III,0)}{P(0)}\right) \\ &= \left(\frac{0 \times \frac{1}{2}}{0 \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{4} + 1 \times \frac{1}{4}}, \frac{\frac{2}{3} \times \frac{1}{4}}{\frac{5}{12}}, \frac{1 \times \frac{1}{4}}{\frac{5}{12}}\right) \\ &= (0, \frac{2}{5}, \frac{3}{5}).\end{aligned}$$

What if we see another 0?

$$\begin{aligned}\Big(P(I|0,0), P(II|0,0), P(III|0,0)\Big) &= \left(\frac{P(I,0,0)}{P(0,0)}, \frac{P(II,0,0)}{P(0,0)}, \frac{P(III,0,0)}{P(0,0)}\right) \\ &= \left(\frac{0 \times \frac{1}{2}}{0 \times \frac{1}{2} + (\frac{2}{3})^2 \times \frac{1}{4} + 1 \times \frac{1}{4}}, \frac{(\frac{2}{3})^2 \times \frac{1}{4}}{\frac{1}{9} + \frac{1}{4}}, \frac{1 \times \frac{1}{4}}{\frac{1}{9} + \frac{1}{4}}\right) \\ &= (0, \frac{4}{13}, \frac{9}{13}).\end{aligned}$$

Alternatively, note that if we take as our prior $(0, \frac{2}{5}, \frac{3}{5})$, and observe a zero, our posterior is

$$\begin{aligned}\Big(P(I|0), P(II|0), P(III|0)\Big) &= \left(\frac{P(I,0)}{P(0)}, \frac{P(II,0)}{P(0)}, \frac{P(III,0)}{P(0)}\right) \\ &= \left(\frac{0 \times 0}{0 \times 0 + \frac{2}{3} \times \frac{2}{5} + 1 \times \frac{3}{5}}, \frac{\frac{2}{3} \times \frac{2}{5}}{\frac{13}{15}}, \frac{1 \times \frac{3}{5}}{\frac{13}{15}}\right) \\ &= (0, \frac{4}{13}, \frac{9}{13}).\end{aligned}$$

In particular, in a Bayesian setting, there are two conceptually different ways to compute your posterior distribution when making multiple observations sequentially. One can compute the posterior "statically" using the initial prior, or compute the posterior "dynamically" by continuing to update one's prior as one sees subsequent observations. These two conceptually different approaches will yield the same posterior distribution, and in general we will in any given situation use the approach which is most convenient. We

also note that in many of the examples we consider, all that will matter are the observation counts / values, as opposed to the order in which one has seen these given observations, and this will further simplify certain calculations.

### 1.3 Bayesian probability: notation cheatsheet

As we shall be working with a fairly general notion of distribution (on distributions), let us set some additional notations. In general we will be working with probability measures (typically denoted $\mu$), which will allow us a single notation for discrete and continuous distributions.

- $\{\mu_\theta, \theta \in \Theta\}$ : The collection of possible probability measures for the reward distribution
  - $\Theta$ will be a general index set, and may be a finite set (e.g. $\{I, II, III\}$) or uncountably infinite (e.g. the closed interval $[0, 1]$) or even more exotic
  - For a subset $S \subseteq \mathcal{R}$, $\mu_\theta(S)$ is the probability assigned to the set $S$ under the measure $\mu_\theta$
  - Given an unordered set of real numbers $\mathcal{H}$, $\mu_\theta(\mathcal{H})$ is the probability that the first $|\mathcal{H}|$ rewards, as an unordered set, equals $\mathcal{H}$
  - Given a collection $H$ of unordered sets of real numbers, all of the same cardinality, $\mu_\theta(H)$ is the probability that the first $|\mathcal{H}|$ rewards, as an unordered set, belongs to $H$.
  - On our previous example:
    - $\Theta = \{1, 2, 3\}$
    - $\mu_1(1) = 1; \mu_2(0) = \frac{2}{3}, \mu_2(5) = \frac{1}{3}; \mu_3(0) = 1$
- $\hat{\mu}$: The prior distribution on the set of possible distributions $\{\mu_\theta, \theta \in \Theta\}$
  - The initial measure on measures
  - For simplicity, we take $\hat{\mu}$ to be a measure on $\Theta$, which (as each $\theta$ is associated with a particular distribution) induces a distribution on the set of all distributions.
  - For $T \subseteq \Theta$, $\hat{\mu}(T)$ is the probability the $\theta$ of the true distribution belongs to that subset of $\Theta$. We will sometimes be informal and simply say that $\hat{\mu}(T)$ is the probability that the true distribution belongs to $T$ (i.e. not make reference to the index of that distribution).
  - On our previous example:
    - $\hat{\mu}(1) = \frac{1}{2}, \hat{\mu}(2) = \hat{\mu}(3) = \frac{1}{4}$
- $\mathcal{H}$: A history of observations
  - Will generally be taken as an unordered set of observed rewards
  - On our previous example:
    - $\mathcal{H} = \{0, 0\}$
- $\hat{\mu}_{\mathcal{H}}$: The posterior distribution on $\Theta$ given the observation history $\mathcal{H}$
  - For $T \subseteq \Theta$,
$$\hat{\mu}_{\mathcal{H}}(T) \triangleq \int_{\theta \in T} \frac{\mu_\theta(\mathcal{H}) d\hat{\mu}(\theta)}{\int_{\theta' \in \Theta} \mu_{\theta'}(\mathcal{H}) d\hat{\mu}(\theta')}$$
  - Also generalizes naturally to a set $H$ of histories, all of the same cardinality:
$$\hat{\mu}_H(T) \triangleq \int_{\theta \in T} \frac{\mu_\theta(H) d\hat{\mu}(\theta)}{\int_{\theta' \in \Theta} \mu_{\theta'}(H) d\hat{\mu}(\theta')}$$
  - On our previous example:

* $\hat{\mu}_{\{0,0\}}(1) = 0, \hat{\mu}_{\{0,0\}}(2) = \frac{4}{13}, \hat{\mu}_{\{0,0\}}(3) = \frac{9}{13}$

- $\overline{\mu}$: Measure on rewards under the prior (on distributions)
  - For $S \subseteq \mathcal{R}$,
$$\overline{\mu}(S) \triangleq \int_{\theta \in \Theta} \mu_\theta(S) d\hat{\mu}(\theta)$$
  - On our previous example:
    * $\overline{\mu}(0) = \sum_{i=1}^{3} \mu_i(0)\hat{\mu}(i) = 0 \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{4} + 1 \times \frac{1}{4} = \frac{5}{12}$
    * $\overline{\mu}(1) = \sum_{i=1}^{3} \mu_i(1)\hat{\mu}(i) = 1 \times \frac{1}{2} + 0 \times \frac{1}{4} + 0 \times \frac{1}{4} = \frac{1}{2}$
    * $\overline{\mu}(5) = \sum_{i=1}^{3} \mu_i(5)\hat{\mu}(i) = 0 \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{4} + 0 \times \frac{1}{4} = \frac{1}{12}$
- $\overline{\mu}_{\mathcal{H}}$: Measure on rewards under the posterior (given the history $\mathcal{H}$)
  - For $S \subseteq \mathcal{R}$,
$$\overline{\mu}_{\mathcal{H}}(S) \triangleq \int_{\theta \in \Theta} \mu_\theta(S) d\hat{\mu}_{\mathcal{H}}(\theta)$$
  - Also generalizes naturally to a set $H$ of histories, all of the same cardinality:
$$\overline{\mu}_{H}(S) \triangleq \int_{\theta \in \Theta} \mu_\theta(S) d\hat{\mu}_{H}(\theta)$$
  - On our previous example:
    * $\overline{\mu}_{\{0,0\}}(0) = \sum_{i=1}^{3} \mu_i(0)\hat{\mu}_{\{0,0\}}(i) = 0 \times 0 + \frac{2}{3} \times \frac{4}{13} + 1 \times \frac{9}{13} = \frac{35}{39}$
    * $\overline{\mu}_{\{0,0\}}(1) = \sum_{i=1}^{3} \mu_i(1)\hat{\mu}_{\{0,0\}}(i) = 1 \times 0 + 0 \times \frac{4}{13} + 0 \times \frac{9}{13} = 0$
    * $\overline{\mu}_{\{0,0\}}(5) = \sum_{i=1}^{3} \mu_i(5)\hat{\mu}_{\{0,0\}}(i) = 0 \times 0 + \frac{1}{3} \times \frac{4}{13} + 0 \times \frac{9}{13} = \frac{4}{39}$
- $\hat{\theta}$: the r.v. representing the random $\theta$ governing the true distribution
- $E[R|\hat{\theta}]$: Conditional expected per-period reward given the index of the true distribution equals $\theta$
  -
$$E[R|\hat{\theta}] \triangleq \int_{-\infty}^{\infty} x d\mu_{\hat{\theta}}(x)$$
  - Sometimes, if it clarifies a concept, we will write $E[R|\hat{\theta} = \theta]$ (i.e. the less formal use of conditional expectation)
  - On our previous example:
    * $E[R|\hat{\theta} = 1] = \int_{-\infty}^{\infty} x d\mu_1(x) = 1 \times 1 = 1$
    * $E[R|\hat{\theta} = 2] = \int_{-\infty}^{\infty} x d\mu_2(x) = 0 \times \frac{2}{3} + 5 \times \frac{1}{3} = \frac{5}{3}$
    * $E[R|\hat{\theta} = 3] = \int_{-\infty}^{\infty} x d\mu_3(x) = 0 \times 1 = 0$
- $E[R|\mathcal{H}]$: Conditional expected per-period reward given the observed history equals $\mathcal{H}$
  - $E[R|\mathcal{H}] \triangleq \int_{-\infty}^{\infty} x d\overline{\mu}_{\mathcal{H}}(x)$
  - Also generalizes naturally to a set $H$ of histories, all of the same cardinality:
$$E[R|H] \triangleq \int_{-\infty}^{\infty} x d\overline{\mu}_{H}(x)$$
  - On our previous example:

* $E[R|\mathcal{H} = \{0,0\}] = \int_{-\infty}^{\infty} x d\overline{\mu}_{\{0,0\}}(x) = 0 \times \frac{35}{39} + 1 \times 0 + 5 \times \frac{4}{39} = \frac{20}{39}$

- $E[R]$: Unconditional expected per-period reward
  - $E[R] \triangleq \int_{-\infty}^{\infty} x d\overline{\mu}(x) = \int_{\theta \in \Theta} E[R|\hat{\theta} = \theta] d\hat{\mu}(\theta)$
  - On our previous example:
    * $E[R] = \int_{-\infty}^{\infty} x d\overline{\mu}(x) = 0 \times \frac{5}{12} + 1 \times \frac{1}{2} + 5 \times \frac{1}{12} = \frac{11}{12}$

## 1.4 Another Bayesian example: Beta-Bernoulli model

We now further illustrate the above ideas through a canonical Bayesian setting called the Beta-Bernoulli model, in which your prior is uniform over all Bernoulli distributions.

- $\{\mu_\theta, \theta \in \Theta\}$ :
  - $\Theta = [0,1]$
  - $\mu_\theta(1) = \theta, \mu_\theta(0) = 1 - \theta$. Namely $\mu_\theta$ is a Bernoulli($\theta$) distribution, abbreviated Be($\theta$)
- $\hat{\mu}$:
  - For every closed interval $[a,b] \subseteq [0,1]$, $\hat{\mu}([a,b]) = b - a$, i.e. Lebesgue measure
- $\mathcal{H}$: A history of observations
  - An unordered set of 0's and 1's
  - $n_{\mathcal{H}}(1)$: number of 1's in the history; $n_{\mathcal{H}}(0)$: number of 0's in the history
- $\hat{\mu}_{\mathcal{H}}$: The posterior distribution on $\Theta$ given the observation history $\mathcal{H}$
  - For $T \subseteq [0,1]$,

$$\begin{aligned}
\hat{\mu}_{\mathcal{H}}(T) &= \int_{\theta \in T} \frac{\binom{n_{\mathcal{H}}(1)+n_{\mathcal{H}}(0)}{n_{\mathcal{H}}(1)} \theta^{n_{\mathcal{H}}(1)} (1-\theta)^{n_{\mathcal{H}}(0)} d\theta}{\int_0^1 \binom{n_{\mathcal{H}}(1)+n_{\mathcal{H}}(0)}{n_{\mathcal{H}}(1)} \theta'^{n_{\mathcal{H}}(1)} (1-\theta')^{n_{\mathcal{H}}(0)} d\theta'} \\
&= \int_{\theta \in T} \frac{\theta^{n_{\mathcal{H}}(1)} (1-\theta)^{n_{\mathcal{H}}(0)} d\theta}{\int_0^1 \theta'^{n_{\mathcal{H}}(1)} (1-\theta')^{n_{\mathcal{H}}(0)} d\theta'} \\
&= \frac{(n_{\mathcal{H}}(1) + n_{\mathcal{H}}(0) + 1)!}{n_{\mathcal{H}}(1)! \times n_{\mathcal{H}}(0)!} \int_{\theta \in T} \theta^{n_{\mathcal{H}}(1)} (1-\theta)^{n_{\mathcal{H}}(0)} d\theta.
\end{aligned}$$

- $\overline{\mu}$:
  - $\overline{\mu}(1) = \int_0^1 \mu_\theta(1) d\hat{\mu}(\theta) = \int_0^1 \theta d\theta = \frac{1}{2}$
  - $\overline{\mu}(0) = \frac{1}{2}$
- $\overline{\mu}_{\mathcal{H}}$:

–

$$\begin{aligned}
\overline{\mu}_{\mathcal{H}}(1) &= \int_0^1 \frac{\binom{n_{\mathcal{H}}(1)+n_{\mathcal{H}}(0)}{n_{\mathcal{H}}(1)}\theta^{n_{\mathcal{H}}(1)}(1-\theta)^{n_{\mathcal{H}}(0)}\theta d\theta}{\int_0^1 \binom{n_{\mathcal{H}}(1)+n_{\mathcal{H}}(0)}{n_{\mathcal{H}}(1)}\theta'^{n_{\mathcal{H}}(1)}(1-\theta')^{n_{\mathcal{H}}(0)}d\theta'} \\
&= \int_0^1 \frac{\theta^{n_{\mathcal{H}}(1)+1}(1-\theta)^{n_{\mathcal{H}}(0)}d\theta}{\int_0^1 \theta'^{n_{\mathcal{H}}(1)}(1-\theta')^{n_{\mathcal{H}}(0)}d\theta'} \\
&= \frac{\frac{(n_{\mathcal{H}}(1)+1)!n_{\mathcal{H}}(0)!}{(n_{\mathcal{H}}(1)+n_{\mathcal{H}}(0)+2)!}}{\frac{n_{\mathcal{H}}(1)!n_{\mathcal{H}}(0)!}{(n_{\mathcal{H}}(1)+n_{\mathcal{H}}(0)+1)!}} \\
&= \frac{n_{\mathcal{H}}(1)+1}{n_{\mathcal{H}}(1)+n_{\mathcal{H}}(0)+2} \\
&= \frac{n_{\mathcal{H}}(1)+1}{(n_{\mathcal{H}}(1)+1)+(n_{\mathcal{H}}(0)+1)}
\end{aligned}$$

  - $\overline{\mu}_{\mathcal{H}}(0) = \frac{n_{\mathcal{H}}(0)+1}{(n_{\mathcal{H}}(1)+1)+(n_{\mathcal{H}}(0)+1)}$

- $E[R|\hat{\theta}=\theta]$:
  - $E[R|\hat{\theta}=\theta] = \theta$
- $E[R|\mathcal{H}]$:
  - $E[R|\mathcal{H}] = \int_{-\infty}^{\infty} x d\overline{\mu}_{\mathcal{H}}(x) = 0 \times d\overline{\mu}_{\mathcal{H}}(0) + 1 \times d\overline{\mu}_{\mathcal{H}}(1) = \frac{n_{\mathcal{H}}(1)+1}{(n_{\mathcal{H}}(1)+1)+(n_{\mathcal{H}}(0)+1)}$
- $E[R]$:
  - $E[R] = 0 \times \overline{\mu}(0) + 1 \times \overline{\mu}(1) = \frac{1}{2}$

## 1.5 Sequences of histories and the history Markov chain

Let $\{\mathcal{H}_t, t \geq 1\}$ denote the random nested sequence of unordered sets of realized rewards observed over time. Thus $\mathcal{H}_1$ is the reward you see after your first observation, $\mathcal{H}_2$ is the unordered set consisting of the first two rewards, etc. Then $\mathcal{H}_t \subseteq \mathcal{H}_{t+1}$ for all $t \geq 1$. Note that $\mathcal{H}_{t+1} \setminus \mathcal{H}_t$ is the unique observation one must see in period $t+1$ so that, starting with history $\mathcal{H}_t$ in period $t$, one has history $\mathcal{H}_{t+1}$ in period $t+1$. For histories $\mathcal{H}_t \subseteq \mathcal{H}_{t+1}$, let $\mu(\mathcal{H}_{t+1}|\mathcal{H}_t)$ denote the probability that, starting with history $\mathcal{H}_t$ in period $t$, one has history $\mathcal{H}_{t+1}$ in period $t+1$. Then in our previous notation, supposing that for any given t the set of possible histories is countable (if not one would operate with corresponding sets of histories),

$$\mu(\mathcal{H}_{t+1}|\mathcal{H}_t) = \overline{\mu}_{\mathcal{H}_t}\big(\mathcal{H}_{t+1} \setminus \mathcal{H}_t\big).$$

Similarly, let $\mu(\mathcal{H}_{t+1}|\{\mathcal{H}_i, i=1,\ldots,t\})$ denote the probability that, having observed histories $\mathcal{H}_1,\ldots,\mathcal{H}_t$ in periods $1,\ldots,t$, one has history $\mathcal{H}_{t+1}$ in period $t+1$. Note that specifying the sequence of histories is equivalent to specifying the unordered set $\mathcal{H}_t$ with the elements ordered by the time at which they were observed. However, as specifying the history as an ordered and unordered set yield the same posterior distribution $\overline{\mu}_{\mathcal{H}_t}$, it follows from a straightforward argument that $\mu(\mathcal{H}_{t+1}|\{\mathcal{H}_i, i=1,\ldots,t\}) = \mu(\mathcal{H}_{t+1}|\mathcal{H}_t)$. Combining the above, we find that $\{\mathcal{H}_t, t \geq 1\}$ is a Markov chain with Markov kernel $\mu(\mathcal{H}_{t+1}|\mathcal{H}_t)$. We will in general take the initial state $\mathcal{H}_0$ to be a default empty history $\emptyset$, and note that for the case of uncountable histories one also gets a Markov chain, albeit on an uncountably infinite state-space.

## 1.6 Bayesian formulation of the MAB problem

We now make completely formal an initial rigorous formulation for the Bayesian MAB problem. This formulation is standard/classical, although later in the class we wil see other variants with additional features. We suppose there are $N$ alternatives (i.e. arms, we note that a MAB is British slang for a slot machine), and time is discrete. For each arm $i$, there is a prior distribution $\hat{\mu}^i$ on reward distributions, and the reward distribution of each arm is drawn independently from its prior. Namely, letting $\hat{\theta}^i$ denote the true $\theta$ realized for arm $i$ (governing its reward distribution), $\{\hat{\theta}^i, i = 1, \ldots, N\}$ are independent. For arm $i$, there is a set $\{\mu^i_\theta, \theta \in \Theta^i\}$ of possible measures

Our goal will be to maximize the net expected reward that we earn over the entire time horizon. As a mathematical convenience, we consider an infinite-horizon discounted model, standard in the Markov decision process (MDP) literature. If $R_t$ is the reward we earn in period $t$, then we wish to maximize $E[\sum_{t=1}^{\infty} \beta^t R_t] = \sum_{t=1}^{\infty} \beta^t E[R_t]$, where $\beta$ is the discount factor $\in (0, 1)$, and $R_t$ is the random reward we earn in period $t$. Of course, the term $R_t$ is hiding a lot, including randomness associated with the distribution associated with each arm (for which you have a prior), randomness associated with the rewards you see on the given arm, and the actions of your policy for switching between the arms. Let $\{R^i_k, k \geq 1\}$ denote the sequence of rewards one would realize on arm $i$ if arm $i$ was played consecutively, an infinite number of times. Note that $\{R^i_k, k \geq 1\}$ is NOT i.i.d., but is conditionally i.i.d., conditional on $\hat{\theta}^i$. Indeed, based on our discussion of the history Markov chain, it is easy to see that the conditional distribution of $R^i_k$, given $\{R^i_j, j = 1, \ldots, k-1\}$, is the same as $\overline{\mu}_{\{R^i_j, j=1,\ldots,k-1\}}$. As a shorthand, we will let $R^i$ denote $R^i_1$.

Before further formalizing all of these things, let's warm up a bit. First, suppose that before the first pull, you had to select an arm and play that arm forever. With no ability to learn anything, it will be optimal to simply select the arm with the highest expected reward. Of course, this means the highest expected reward with regards to (w.r.t.) two sources of randomness: the randomness in which distribution that arm has, and the randomness in the reward associated with that given distribution. In this case, the optimal policy (subject to these "no-learning" constraints) would simply select the arm $i$ maximizing $E[R^i]$, and in expectation (over the entire horizon) earn $\frac{\beta}{1-\beta} \max_{i \in [1,N]} E[R^i]$.

As a second warmup, suppose that before your first pull, an omniscient being reveals to you the true distribution associated with each arm (the problem as stated only endows you with a prior over distributions for each arm). In this case, what is the expected reward earned by an optimal policy (given this impossible side information)? Clearly the optimal policy is to select the arm whose realized distribution has the highest expectation. Then the optimal policy can earn

$$\frac{\beta}{1-\beta} E[\max_{i \in [1,N]} E[R^i | \hat{\theta}_i]].$$

Namely, here we can obtain the expected value (over the priors) of the max of the expectations (of the realized distributions), which will always be larger than the maximum of the associated expectations (acheived under the previous assumption, by Jensen's inequality).

To make this even more concrete, suppose that each arm has the Beta-Bernoulli prior defined earlier. Then if we had to select an arm with no information at all, all arms are equally good, and an optimal policy would simply select any of the arms, and earn $\frac{\beta}{1-\beta} \times \frac{1}{2}$. If we see the true distributions before selecting an arm, which here would mean seeing the true success probability of each arm, we would select the maximum, and (as the expected value of a Bernoulli is simply its probability of equalling 1) earn in expectation the expected value of the maximum of N independent uniforms, equal to (by the tail-integral formula for expected

value, letting $\{U_i, i \in [1,N]\}$ be i.i.d. U[0,1])

$$\begin{aligned}
\frac{\beta}{1-\beta} E[\max_{i\in[1,N]} U_i] &= \frac{\beta}{1-\beta}\int_0^\infty P\big(\max_{i\in[1,N]} U_i > x\big)dx \\
&= \frac{\beta}{1-\beta}\int_0^1 \Big(1 - P\big(\max_{i\in[1,N]} U_i \le x\big)\Big)dx \\
&= \frac{\beta}{1-\beta}\int_0^1 \Big(1 - \prod_{i=1}^N P\big(U_i \le x\big)\Big)dx \\
&= \frac{\beta}{1-\beta}\int_0^1 \Big(1 - x^N\Big)dx \quad = \quad \frac{\beta}{1-\beta}\big(1 - \frac{1}{N+1}\big).
\end{aligned}$$

Note that for $N \ge 2$, $\frac{\beta}{1-\beta}\big(1 - \frac{1}{N+1}\big) > \frac{\beta}{1-\beta} \times \frac{1}{2}$.

We have thus analyzed the "two extremes" - one in which you must base all decisions on no information (beyond the prior), and one in which you can base all decisions on perfect information. In the actual problem, one bases one's decisions on a policy which gains information over time, albeit in general never having perfect information about the true distribution on any of the arms. In general, a policy uses past rewards, combined with the initial prior, to inform future decisions. More formally, a policy must be adapted (in the sense of filtration) to the $\sigma$-field generated by the past observed rewards. Of course, which arm's rewards you have observed will itself depend on the policy choices, so this is a bit subtle. We will define a history vector, $\mathcal{H}_t^i$, for each arm $i$. Namely, $\mathcal{H}_t^i$ is the unordered set (possibly empty) consisting of the set of realized rewards observed on arm $i$ in periods $1, \ldots, t$. Note that in general the histories of different arms will have different cardinalities, as your policy may have pulled certain arms more times than other arms.

In that case, a policy is a mapping which, in period $t$, takes as input any possible vector of histories up to time $t-1$, $\big(\mathcal{H}_{t-1}^i, i = 1, \ldots, N\big)$, and outputs a distribution (possibly degenerate) on the index set $[1, N]$ (corresponding to your next move being to select an arm according to that distribution). Here we take $\mathcal{H}_0^i = \emptyset$. For a given policy $\pi$, let $t^{\pi,i}(0) = 0$; $N^{\pi,i}(t)$ denote the (random) number of times that $\pi$ pulls arm $i$ during periods $1, \ldots, t$; $N^{\pi,i}(\infty)$ denote the total number of such pulls over the entire time horizon (may be finite or infinite); and $t^{\pi,i}(k)$ denote the (random) time that $\pi$ pulls arm $i$ for the $k$th time, $k = 1, \ldots, N^{\pi,i}$. Note that although for any given policy $\pi$, it will in general NOT be the case that $\{\mathcal{H}_t^i, t \ge 0\}$ is a Markov chain (since the times at which arm $i$ is pulled may depend in a complex way on the realized rewards on all arms), it WILL be the case that $\{\mathcal{H}^i_{t^{\pi,i}(k)}, k = 0, \ldots, N^{\pi,i}\}$ evolves as a Markov chain, independent of the particular policy $\pi$ or any information about the realized rewards of the other arms. Note that if the policy $\pi$ pulls arm $i$ at time $t$, this will correspond to the $N^{\pi,i}(t)$ time that the policy has pulled arm $i$, equivalently (if arm i is pulled at time t) the $N^{\pi,i}(t-1)+1$ time that the policy has pulled arm $i$. Hence (by construction and our definitions), if policy $\pi$ pulls arm $i$ at time $t$, it earns (random) reward $R^i_{N^{\pi,i}(t-1)+1}$. We also note that $N^{\pi,i}(t) = |\mathcal{H}_t^i|$.

We note that really the policy can in principle depend on the sequence of histories, i.e. $\Big(\big(\mathcal{H}_k^i, k = 1, \ldots, t-1\big), i = 1, \ldots, N\Big)$, but the fact that each history evolves as a Markov chain (as already discussed) implies by standard arguments from the theory of Markov decision processes (MDP) that one may w.l.o.g. (without loss of generality) restrict to the family of so-called Markov policies whose decision in period $t$ depends only on $\big(\mathcal{H}_{t-1}^i, i = 1, \ldots, N\big)$. We let $\mathcal{H}_{t-1}$ denote the (ordered) vector of unordered sets $\big(\mathcal{H}_{t-1}^i, i = 1, \ldots, N\big)$. We note that the policy also knows the initial prior, but this will generally be implicit (i.e. notationally supressed). It also follows from the general theory of MDP that there exists an optimal policy which is deterministic (i.e. maps every possible $\big(\mathcal{H}_{t-1}^i, i = 1, \ldots, N\big)$ to a deterministic single index), and stationary (i.e. makes the same action whenever faced with the same set of histories, although here this is not really meaningful as no collection of histories can be visited twice as the total number of observations grows with time). We let $\Pi$ denote the family of policies satisfying these requirements.

For a given policy $\pi \in \Pi$ and $t \ge 1$, we let $\pi(\mathcal{H}_{t-1})$ denote the index of the arm the policy selects

in period $t$, conditional on the state at the start of period t being $\mathcal{H}_{t-1}$. We will also let $\pi(t)$ denote the random index of the arm the policy selects in period $t$, where under that notation the dependence on $\mathcal{H}_{t-1}$ is implicit, but it is of course true that w.p.1 $\pi(t) = \pi(\mathcal{H}_{t-1})$. We will broadly use the notation $I(\cdot)$ to denote the indicator function, evaluating to 1 if the expression inside is true, and 0 otherwise.

Now, we can formally state our problem as:

$$\max_{\pi \in \Pi} E[\sum_{t=1}^{\infty} \beta^t R^{\pi(t)}_{|\mathcal{H}^{\pi(t)}_{t-1}|+1}].$$

It turns out that the right way to think about the expectation is to condition on $\mathcal{H}_{t-1}$, as follows.

$$\begin{aligned}
\max_{\pi \in \Pi} E[\sum_{t=1}^{\infty} \beta^t R^{\pi(t)}_{|\mathcal{H}^{\pi(t)}_{t-1}|+1}] &= \max_{\pi \in \Pi} \sum_{t=1}^{\infty} \beta^t E[R^{\pi(t)}_{|\mathcal{H}^{\pi(t)}_{t-1}|+1}] \\
&= \max_{\pi \in \Pi} \sum_{t=1}^{\infty} \beta^t E\Big[E[R^{\pi(t)}_{|\mathcal{H}^{\pi(t)}_{t-1}|+1}|\mathcal{H}_{t-1}]\Big] \\
&= \max_{\pi \in \Pi} \sum_{t=1}^{\infty} \beta^t E\Big[E[\sum_{i=1}^{N} R^{i}_{|\mathcal{H}^{i}_{t-1}|+1} I\big(\pi(\mathcal{H}_{t-1}) = i\big)|\mathcal{H}_{t-1}]\Big] \\
&= \max_{\pi \in \Pi} \sum_{t=1}^{\infty} \beta^t E\Big[\sum_{i=1}^{N} E\big[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1} I\big(\pi(\mathcal{H}_{t-1}) = i\big)|\mathcal{H}_{t-1}\big]\Big] \\
&= \max_{\pi \in \Pi} \sum_{t=1}^{\infty} \beta^t E\Big[\sum_{i=1}^{N} I\big(\pi(\mathcal{H}_{t-1}) = i\big) E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}_{t-1}]\Big] \quad \text{by meas. and the pull-out property} \\
&= \max_{\pi \in \Pi} \sum_{t=1}^{\infty} \beta^t E\Big[\sum_{i=1}^{N} I\big(\pi(\mathcal{H}_{t-1}) = i\big) E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}^i_{t-1}]\Big] \quad \text{by independence of the arms.}
\end{aligned}$$

We will have more to say on this point later, but for now note that $E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}^i_{t-1}]$ does not in any way depend on the choices made by the policy, or anything else, except in-so-far as those things impact $\mathcal{H}^i_{t-1}$. Furthermore, $E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}^i_{t-1}]$ does not even really depend on $t$, again except in-so-far as $t$ impacts the unordered set $\mathcal{H}^i_{t-1}$. Namely, if under two different policies, and for two different times, it turns out that the unordered set comprising the history on arm i is the same, then these conditional expectations will w.p.1 equal one-another. Another way to understand this is that, by the measurability aspect in the definition of conditional expectation, it holds that $E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}^i_{t-1}]$ is a deterministic function of $\mathcal{H}^i_{t-1}$, the history (as an unordered set) on arm $i$ up to time $t$.

We also make another peculiar observation. If we considered the same exact MAB problem, except that instead of seeing a random reward every time you pulled arm $i$, you instead received $E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}^i_{t-1}]$ whenever you pull arm $i$ and the history on arm $i$ at that time is exactly $\mathcal{H}^i_{t-1}$, we are faced with exactly the same optimization, which will have exactly the same optimal policies. Namely, we can w.l.o.g. assume that if we ever play arm $i$ while its history is the unordered set $\mathcal{H}$, we receive a deterministic (as a function of $\mathcal{H}$) payout $E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}^i_{t-1}]$.

Lastly, we note that the dynamic evolution of the state is implicit in the above formulation. In particular, at the start of time $t$, the state of the system is $\mathcal{H}_{t-1} = \big(\mathcal{H}^i_{t-1}, i = 1, \ldots, N\big)$. Under policy $\pi$, in period $t$ you select arm $\pi(\mathcal{H}_{t-1})$, receive reward $\sum_{i=1}^{N} I\big(\pi(\mathcal{H}_{t-1}) = i\big) E[R^{i}_{|\mathcal{H}^{i}_{t-1}|+1}|\mathcal{H}^i_{t-1}]$, and then the history of arm $\pi(\mathcal{H}_{t-1})$ updates according to its Markovian dynamic by one step, while all other arm's histories remain unchanged. This updated set of histories constitutese $\mathcal{H}_t$.

## 1.7 Connection to the Markov Chain Selection Problem

When written formally as above, we make an important observation connecting the Bayesian MAB problem to a more general stochastic control problem. In particular, this Bayesian MAB problem is actually a special case of the following problem, which we will call the Markov Chain Selection Problem. In the Markov Chain Selection Problem (MCSP), one has $N$ independent Markov chains, $\{M^i, i = 1, \ldots, N\}$. Associated to each state $s$ in the (here we assume countable, but this is not essential) state-space of $M^i$ is a deterministic reward $R^i(s)$. Markov chain $M^i$ is initialized in a particular state $M^i_0$. For this problem a policy $\pi$ is a mapping which, in period $t$, takes as input any possible vector comprising the current states of all $N$ Markov chains at (the end of) time $t-1$ (equivalently at the start of time $t$), $M_{t-1} = \big(M^i_{t-1}, i = 1, \ldots, N\big)$, and outputs a deterministic index $\pi(M_{t-1}) = \pi(t)$ in $[1, N]$. We also require the policy be stationary, and let $\Pi$ denote the family of all such policies. In period $t$ you earn the (time-discounted) reward of the state of the arm you pulled, i.e. $\beta^t \sum_{i=1}^N I\big(\pi(M_{t-1}) = i\big) R^i(M^i_{t-1})$. The dynamics are then as follows: the Markov chain selected in period $t$, $\pi(M_{t-1})$, advances by one step (according to its Markovian dynamics), while all other Markov chains remain in their past states. This updated set of states constitutese $M_t$.

In that case, our objective for this Markov Chain Selection Problem is

$$\max_{\pi \in \Pi} \sum_{t=1}^{\infty} \beta^t E\Big[\sum_{i=1}^{N} I(\pi(t) = i) R^i(M^i_{t-1})\Big].$$

That the Bayesian MAB problem is a special case follows from our observations regarding the history Markov chain, our observation that one can replace the random rewards with the appropriate conditional expectations, and the analogous dynamics (i.e. only the selected alternative updates).

Let us illustrate with a concrete example of this reduction, and begin to think on the optimal policy and the trade-offs involved. Suppose there are two arms, with $\Theta^1 = \{1, 2, 3, 4\}$; $\hat{\mu}^1(\theta) = \frac{1}{4}$ for $\theta = 1, 2, 3, 4$; and $\mu^1_\theta(\theta) = 1$ for $\theta = 1, 2, 3, 4$. Namely, the reward distribution on arm 1 is either always 1, always 2, always 3, or always 4, with all possibilities equally likely. Note that in this simple example, after pulling arm one one time, Baye's rule will indeed ensure that you know the true distribution on that arm (which is a very special and simple case). Suppose $\Theta^2 = 3$, $\hat{\mu}^2(3) = 1$; and $\mu^2_3(3) = 1$. Namely, the reward distribution on arm 2 is always 3. Note that the prior on the second arm is degenerate, and you thus know the true distribution on the second arm before any pulls at all. A two-arm MAB problem in which you apriori know the true distribution for one of the arms is sometimes referred to as the 1-and-$\frac{1}{2}$-arm bandit problem. What would the reduction to the MCSP look like? For arm 1, we initialize in a state $\emptyset$, with $R^1(\emptyset) = \frac{1+2+3+4}{4} = 2.5$. From state $\emptyset$ there are 4 states to transition to, states $1, 2, 3, 4$. $P^1_{\emptyset,i} = \frac{1}{4}$, and $R^1(i) = i$, for all $i = 1, 2, 3, 4$. Furthermore, letting $i^k$ denote the unordered set consisting of exactly $k$ $i$'s (for $k \geq 1$), $P^1_{i^k, i^{k+1}} = 1, R^1(i^k) = i$, $i = 1, 2, 3, 4$. It is easy to see that this chain will have the same associated dynamics and rewards as the following simpler 5-state chain. There is an initial state $\emptyset$ with $R^1(\emptyset) = 2.5$. There are 4 additional states $1, 2, 3, 4$, with $P^1_{\emptyset,i} = \frac{1}{4}, R^1(i) = i$, and $P^1_{i,i} = 1$ for $i = 1, 2, 3, 4$.

For arm 2, we initialize in a state $\emptyset$, with $R^2(\emptyset) = 3$. Furthermore, $P^2_{\emptyset,3} = 1$, $P^2_{3^k, 3^{k+1}} = 1$, and $R^2(3^k) = 3$ for all $k \geq 1$. It is again easy to see that the dynamics of this chain are equivalent to the single-state chain with only a state 3, s.t. $P^2_{3,3} = 1, R^2(3) = 3$.

What is the optimal policy here? Let us begin with a few observations. Working with the reduced chains, note that if an optimal policy ever selects a state that loops to itself w.p.1, then by stationarity, it must be optimal to continue to advance that state forever, as no states of any of the Markov chains will have changed. In particular, in the MAB language, if one ever pulls an arm whose distribution is known with certainty, then it must be optimal to continue to play that arm forever (at least in the infinite-horizon discounted setting). Thus, considering the reduced chains, there are only a few policies we need to consider. Namely, our policy is completely determined if we decide: 1. which arm to play first, and 2. if we play the first arm and see an $i$ do we continue to play the first arm forever or switch to the second arm. A straightforward argument further shows that, having played the first arm, it is optimal to continue to play the first arm forever if the reward is at least 3, and otherwise to switch to the second arm and play it forever. Thus there are only two policies to compare. If we play the second arm at time 1, we earn $3\frac{\beta}{1-\beta}$. If we play the first arm first, we

earn

$$\begin{aligned} 2.5\beta + \sum_{i=1}^{4} \frac{\beta^2}{1-\beta}\frac{1}{4}\max(i,3) &= 2.5\beta + \frac{\beta^2}{4(1-\beta)}(3+3+3+4) \\ &= 2.5\beta + \frac{13}{4}\frac{\beta^2}{1-\beta} \\ &= \frac{\beta}{4}\big(10 + 13\frac{\beta}{1-\beta}\big). \end{aligned}$$

So when do we play the second arm at time 1? This is optimal when

$$\begin{aligned} 3\frac{\beta}{1-\beta} \geq \frac{\beta}{4}\big(10 + 13\frac{\beta}{1-\beta}\big) &\leftrightarrow \frac{12}{1-\beta} \geq 10 + 13\frac{\beta}{1-\beta} \\ &\leftrightarrow 12 \geq 10(1-\beta) + 13\beta \\ &\leftrightarrow \beta \leq \frac{2}{3}. \end{aligned}$$

Here we see the fundamental trade-off of the MAB problem. If the discount factor is close-enough to 1, a significant premium is put on ultimately learning which arm is truly better, and the penalty of settling for a sub-optimal arm is very high. Alternatively, if the discount factor is closer to 0, the future is heavily discounted and the value of learning is not that great. Note that if one was myopically selecting the arm with the highest short-term (single-period) expected payout, one would select the second arm at time 1, as $3 \geq 2.5$. Thus here we find that such a myopic policy is optimal if the future is heavily discounted, and otherwise it is optimal to spend the first period learning on the off-chance that the first arm is strictly better (if its reward is 4). This is a concrete manifestation of the so-called exploration vs. exploitation trade-off. For general reward distributions, the trade-off manifests and interacts with the discount factor in more complicated ways, but ultimately one is always one-way-or-another trading off the exploitation of arms for which one is very confident that the payout is high with arms for which one has less information and which may perform even better.

This connection between MAB and MCSP had already been identified by the 1950's. However, as the dimension of the MDP scales linearly with the number of arms, the state-space of the corresponding DP grows exponentially with the number of arms / chains. Thus for many years, it was believed that for even a relatively small number of arms, this problem was intractable.

## 1.8 Gittin's Index Theorem

### 1.8.1 Motivation

In the 1970's, John Gittins had a breakthrough which proved otherwise. Gittins proved that the optimal policy actually had a relatively simple form, which was far more tractable than previously believed. In particular, he proved that it is possible, at time 0, to assign a strictly positive real number, i.e. index, to each state of each arm, with the following property. In each period, of all those states available on the different arms, one simply plays the state which appears highest on this list. This is called an index (i.e. priority) policy, and a problem for which such a policy is optimal is said to be indexable. Apriori, it could have been the case that the inherent "desirability" of a given state depended in a very complex manner on the states of the other arms. For example, consider the setting in which $N = 2$, and both arms have the same Markov chain. In this special case, indexability is equivalent to the following condition: if in state $(i, j)$ you select state $i$, and in state $(j, k)$ you select state $j$, then in state $(i, k)$ you must similarly select state $i$ (modulo ties). Apriori, as the order in which you play the different states creates inherent trade-offs as states played later are more heavily discounted, and as the Markov chain dynamics are general, there is no reason to believe that such a condition must hold. In fact, for many very closely related problems, it does not. Note that for the setting in which every arm has the same Markov chain, the complexity of the

associated optimal policy scales almost independent of the number of arms (depending on the number of arms only through a few lookups and comparisons which can be implemented using efficient data structures and sorting algorithms). This is a vast improvement over the exponential scaling (in the number of arms) of the naive dynamic program.

### 1.8.2 Some simple reductions

We now prove Gittin's theorem. We present a proof which uses an elegant and insightful polyhedral argument, although at the sacrifice of assuming that all chains have a finite number of states (which, I note, is not the case for e.g. the Beta-Bernoulli setup). However, there are many proofs of this theorem at many levels of generality, and proofs are also known for the setting in which the chains have an infinite number of states. In fact, there is a direct generalization of our polyhedral argument to this more general setting, although it requires some additional details that we will not present here. Thus suppose each of the N chains has a finite number of states. W.l.o.g., we suppose that each chain has the same state-space, rewards (for any given state), and transition probabilities. This is w.l.o.g. since we can simply create a state of the form (s,i) for state s on arm i, and have the associated Markov chains contain non-communicating clusters, in which case being in the correct chain can be done through proper initialization.

### 1.8.3 A state-centric linear formulation

The proof begins by forcing us to think about the performance of a policy $\pi$ in a very different way. We are used to thinking about the policy in terms of which arm is pulled when. However, in some sense that is not particularly useful, as the reward you get is governed not by the arm, but by the state of the arm. Thus, we will shift to instead thinking about when your policy plays each different state (i.e. advances a chain in that given state, receiving that state's reward). Namely, let $\mathcal{S}$ denote the state-space of (any) one of the Markov chains (as they all have the same state space). Note that $\mathcal{S}$ is NOT the state-space for the entire system, which must keep track of all $N$ arms, but just the state-space for any one of the arms. For $\pi \in \Pi$, let $\overline{\pi}(t)$ denote the state of the chain which is activated at time t. In particular, $\overline{\pi}(t) = \sum_{i=1}^{N} I(\pi(t) = i) M_{t-1}^{i}$. Note that the total reward earned over the entire time horizon due to contributions from state $s \in \mathcal{S}$ equals $\sum_{t=1}^{\infty} \beta^t I\big(\overline{\pi}(t) = s\big) R(s)$, and the corresponding expectation equals $R(s) E\big[\sum_{t=1}^{\infty} \beta^t I\big(\overline{\pi}(t) = s\big)\big]$. More precisely, as the corresponding expectation depends on the initial set of states $M_0$, and hence is (making this dependence explicit) equals $R(s) E\big[\sum_{t=1}^{\infty} \beta^t I\big(\overline{\pi}(t) = s\big) | M_0\big]$. For a given policy $\pi$ and $M_0$, let $x_s^{\pi, M_0} \stackrel{\Delta}{=} E\big[\sum_{t=1}^{\infty} \beta^t I\big(\overline{\pi}(t) = s\big) | M_0\big]$, and $\overline{x}^{\pi, M_0}$ the corresponding vector (as a column), where we assume there is some fixed ordering on the states. We also make the dependence on the number of arms N explicit, using the notation $x^{\pi, N, M_0}$ and $\overline{x}^{\pi, N, M_0}$. Of course $\overline{x}^{\pi, N, M_0}$ also depends implicitly on the particular Markov chain dynamics and discount factor $\beta$, but we leave those dependencies implicit. Let $\overline{R}$ denote the corresponding vector (as a row) of rewards (both vectors being $\mathcal{S}$-dimensional). Then note that for the given policy $\pi$ and initial condition $M_0$, the expected total reward earned by $\pi$ equals $\overline{R} \cdot \overline{x}^{\pi, N, M_0}$, where we remind the reader that $\overline{R}$ depends on neither $\pi, N$, nor $M_0$ (nor $\beta$ for that matter). Thus we can frame the value of an optimal policy as follows. Let $\chi^{N, M_0}$ denote the set of all vectors $\overline{x}$ s.t. there existst $\pi \in \Pi$ for which $\overline{x} = \overline{x}^{\pi, N, M_0}$. Equivalently, $\chi^{N, M_0} = \bigcup_{\pi \in \Pi} \overline{x}^{\pi, N, M_0}$. Then the value attained by an optimal policy is the solution to the following optimization problem:

$$\max_{\overline{x} \in \chi^{N, M_0}} \overline{R} \cdot \overline{x}.$$

In particular, we are maximizing a linear function over the set $\chi^{N, M_0}$. If $\chi^{N, M_0}$ was a polytope which we could separate over, we would simply have a nice linear program! However, we have no reason to believe that the set $\chi^{N, M_0}$ is in any way "nice". Note that given that there are $N$ Markov chains, each having $|\mathcal{S}|$ states, and (as we are assuming all $N$ chains are the same) we can bound the total number of states of the system by $|\mathcal{S}|^N$, in which we can bound the total number of policies (mappings from this set of states to indices) by $N^{|\mathcal{S}|^N}$, which scales terribly in $N$ and $|\mathcal{S}|$. As each such policy generates a vector, and we are maximizing

a linear function, we know from basic linear programming (LP) that we can equivalently optimize over the convex hull of these vectors. Let $\overline{\chi}^{N,M_0}$ denote the convex hull of $\chi^{N,M_0}$, which will be a poltyope. Then our problem is equivalently

$$\max_{\overline{x}\in\overline{\chi}^{N,M_0}} \overline{R}\cdot\overline{x}.$$

Of course, this polytope could have as many as $N^{|\mathcal{S}|^N}$ vertices, and it is completely unclear how to characterize this polytope, not to mention efficiently separate or optimize. We note that as essentially all general algorithms for MDP also perform worse as the number of states of the systems grows, it is unclear how to use any of those techniques here either.

### 1.8.4 A linear programming relaxation

What to do? Well, as with any linear program for which it is unclear how to characterize the underlying polyope (e.g. many polyhedra arising in combinatorial optimization), we can attempt to write down a relaxation. We begin by introducing some further notations. We will attempt to write down an inequality for every subset $S$ of states, so let us fix some such subset $S$.

- $N_0^i$: Total number of times arm $i$ must be pulled until it enters a state in $S$ for the first time. This counts the pull which takes you to a state in S. $N_0^i = 0$ if $M_0^i \in S$. $N_0^i = \infty$ if arm $i$, pulled an infinite never of times, never enters a state in $S$.
- $t_0^i(k), k = 1, \ldots, N_0^i$: time at which arm $i$ is pulled for the $k$th time. If $N_j^i = 0$, this is not defined for any k.
- $M_0'$ : unordered set of indices of those arms not initially in S, i.e. $\{i : M_i^0 \notin S\}$.
- $t_j^i$: Time at which a state in S is played on arm i for the jth time. $t_j^i = \infty$ if this never occurs.
- $N_j^i$: Total number of times arm i must be pulled, including the pull at time $t_j^i$, if pulled consecutively from that time onward, until it again enters a state in S. Thus (for example) if the pull at time $t_j^i$ results in arm i entering a state in S, it would hold that $N_j^i = 1$. $N_j^i = 0$ if $t_j^i = \infty$. Note that $N_j^i$ may equal $\infty$.
- $t_j^i(k), k = 1, \ldots, N_j^i - 1$: time at which arm $i$ is pulled for the $k$th time, strictly after time $t_j^i$. Note that, conceptually (although we do not define for $k = N_j^i$ to prevent confusion), we would have $t_j^i\big(N_j^i\big) = t_{j+1}^i$. For example, if $N_j^i = 1$, then the first pull on arm $i$ strictly after time $t_j^i$ will activate a state in S, and hence $t_{j+1}^i = t_j^i + 1 = t_j^i(N_j^i)$.
- $\sigma(t_j^i)$: identity of the state in S which is played at time $t_j^i$, i.e. $M_{t_j^i-1}^i$. Note: the value is irrelevant if $t_j^i = \infty$, in which case we set $\sigma(t_j^i) = \emptyset$.

Note that $\sum_{i=1}^N \sum_{j=1}^\infty \big(\beta^{t_j^i} + \sum_{k=1}^{N_j^i-1} \beta^{t_j^i(k)}\big)$ is the discounted time pulling states in S and getting back to states in S. We now derive a lower bound for this quantity. By accounting for how we spend every time period,

$$\frac{\beta}{1-\beta} = \sum_{i\in M_0'}\sum_{k=1}^{N_0^i} \beta^{t_0^i(k)} + \sum_{i=1}^N\sum_{j=1}^\infty \big(\beta^{t_j^i} + \sum_{k=1}^{N_j^i-1}\beta^{t_j^i(k)}\big).$$

Thus

$$\sum_{i=1}^N\sum_{j=1}^\infty \big(\beta^{t_j^i} + \sum_{k=1}^{N_j^i-1}\beta^{t_j^i(k)}\big) = \frac{\beta}{1-\beta} - \sum_{i\in M_0'}\sum_{k=1}^{N_0^i}\beta^{t_0^i(k)}.$$

Furthermore,

$$\sum_{i\in M_0'}\sum_{k=1}^{N_0^i}\beta^{t_0^i(k)} \le \sum_{k=1}^{\sum_{i\in M_0'}N_0^i}\beta^k,$$

with equality if (at least initially) we prioritize those arms not in S, trying to get them to S. Thus defining

$$T_{M_0} \triangleq \sum_{i\in M_0'} N_0^i,$$

it follows that

$$\sum_{i=1}^{N}\sum_{j=1}^{\infty}\big(\beta^{t_j^i}+\sum_{k=1}^{N_j^i-1}\beta^{t_j^i(k)}\big) \ge \frac{\beta}{1-\beta}-\sum_{k=1}^{T_{M_0}}\beta^k. \tag{1}$$

We now derive an upper bound for the left-hand-side of (1), which will thus have to be larger than the right-hand-side of (1). Since $t_j^i(k) \ge t_j^i + k$, we conclude that

$$\sum_{i=1}^{N}\sum_{j=1}^{\infty}\big(\beta^{t_j^i}+\sum_{k=1}^{N_j^i-1}\beta^{t_j^i+k}\big) \ge \frac{\beta}{1-\beta}-\sum_{k=1}^{T_{M_0}}\beta^k, \tag{2}$$

and thus

$$\sum_{i=1}^{N}\sum_{j=1}^{\infty}\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k \ge \frac{\beta}{1-\beta}-\sum_{k=1}^{T_{M_0}}\beta^k,$$

from which it follows that

$$E\Big[\sum_{i=1}^{N}\sum_{j=1}^{\infty}\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k\Big] \ge \frac{\beta}{1-\beta}-E\big[\sum_{k=1}^{T_{M_0}}\beta^k\big]. \tag{3}$$

Next, it will be useful to further partition the left-hand-side by which state we are in. In particular,

$$\sum_{i=1}^{N}\sum_{j=1}^{\infty}\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k = \sum_{i=1}^{N}\sum_{j=1}^{\infty}\sum_{s\in S}I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k,$$

and thus

$$E\Big[\sum_{i=1}^{N}\sum_{j=1}^{\infty}\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k\Big] = E\Big[\sum_{i=1}^{N}\sum_{j=1}^{\infty}\sum_{s\in S}I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k\Big],$$

itself equal to

$$\sum_{i=1}^{N}\sum_{j=1}^{\infty}\sum_{s\in S}E\Big[I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k\Big].$$

Let $T_s^S$ denote the number of transitions it takes, if the chain starts in state $s$, to first hit a state in $S$ (or to again hit a state in S if $s \in S$). By the Markov property,

$$E\Big[I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k\Big] = E\big[\sum_{k=0}^{T_s^S-1}\beta^k\big]E\big[I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\big].$$

Combining with the above, we conclude that

$$\sum_{i=1}^{N}\sum_{j=1}^{\infty}\sum_{s\in S}E\Big[I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\sum_{k=0}^{N_j^i-1}\beta^k\Big]$$

equals

$$\sum_{i=1}^{N}\sum_{j=1}^{\infty}\sum_{s\in S}E\big[\sum_{k=0}^{T_s^S-1}\beta^k\big]E\Big[I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\Big].$$

Interchanging the order of summation, this further equals

$$\sum_{s\in S}E\big[\sum_{k=0}^{T_s^S-1}\beta^k\big]\sum_{i=1}^{N}\sum_{j=1}^{\infty}E\Big[I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\Big],$$

itself equal to

$$\sum_{s\in S}E\big[\sum_{k=0}^{T_s^S-1}\beta^k\big]E\Big[\sum_{i=1}^{N}\sum_{j=1}^{\infty}I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\Big].$$

Noting that

$$E\Big[\sum_{i=1}^{N}\sum_{j=1}^{\infty}I\big(\sigma(t_j^i)=s\big)\beta^{t_j^i}\Big]=x_s^{\pi,M_0},$$

we conclude that

$$\sum_{s\in S}E\big[\sum_{k=0}^{T_s^S-1}\beta^k\big]x_s^{\pi,M_0}\geq\frac{\beta}{1-\beta}-E\big[\sum_{k=1}^{T_{M_0}}\beta^k\big],$$

and after simplifying, and making the implicit dependence of $T_{M_0}$ on S explicit with a superscript,

$$\sum_{s\in S}E\big[\sum_{k=0}^{T_s^S-1}\beta^k\big]x_s^{\pi,M_0}\geq\frac{\beta}{1-\beta}E[\beta^{T_{M_0}^S}].$$

Letting $A_i^S\triangleq E\big[\sum_{k=0}^{T_i^S-1}\beta^k\big]$, and $b_{M_0}(S)\triangleq\frac{\beta}{1-\beta}E[\beta^{T_{M_0}^S}]$, we have that for every $S\subseteq\mathcal{S}$,

$$\sum_{i\in S}A_i^Sx_i^{\pi,M_0}\geq b_{M_0}(S).$$

Furthermore, again by a simple time accounting argument,

$$\sum_{i\in\mathcal{S}}x_i^{\pi,M_0}=\frac{\beta}{1-\beta}.$$

Adding non-negativity of $x_i^{\pi,M_0}$, this gives an LP relaxation of our original problem. Note that it is a relaxation over acheivable vectors (w.r.t. discounted time in the different states), as opposed to an LP directly giving decision variables. Also (and this is very important), the coefficients in no way depend on $\pi$. Thus this must hold for ANY $\pi$. Furthermore, we have that the inequality for set $S$ is tight if our policy always prioritizes sets in $\mathcal{S}\setminus S$ over sets in $S$. Recall, our final LP relaxation is:

$$\max_{\mathbf{x}\in\mathcal{R}^{|\mathcal{S}|}}\sum_{i\in\mathcal{S}}R(i)\mathbf{x}_i,$$

s.t.

$$\sum_{i\in S}A_i^S\mathbf{x}_i\geq b_{M_0}(S)\quad\text{for all}\quad S\subseteq\mathcal{S},S\neq\emptyset;$$

$$\sum_{i\in\mathcal{S}}x_i^{\pi,M_0}=\frac{\beta}{1-\beta}.$$

$$\mathbf{x}_i \geq 0 \quad \text{for all} \quad i \in \mathcal{S}.$$

Further noting that $b_{M_0}(\mathcal{S}) = \frac{\beta}{1-\beta} E[\beta^0] = \frac{\beta}{1-\beta}$; and that $A_i^{\mathcal{S}} = 1$ for all $i$, we may rewrite as follows:

$$\max_{\mathbf{x} \in \mathcal{R}^{|\mathcal{S}|}} \sum_{i \in \mathcal{S}} R(i)\mathbf{x}_i,$$

s.t.

$$\sum_{i \in S} A_i^S \mathbf{x}_i \geq b_{M_0}(S) \quad \text{for all} \quad S \in 2^{\mathcal{S}}, S \neq \mathcal{S}, \emptyset;$$

$$\sum_{i \in \mathcal{S}} A_i^{\mathcal{S}} \mathbf{x}_i = b_{M_0}(\mathcal{S});$$

$$\mathbf{x}_i \geq 0 \quad \text{for all} \quad i \in \mathcal{S}.$$

Now, we make an important observation and interpretation regarding the primal. Note that for any given subset $S$, to derive the corresponding inequality, we only made two approximations. First, we used the inequality

$$\sum_{i \in M_0'} \sum_{k=1}^{N_0^i} \beta^{t_0^i(k)} \leq \sum_{k=1}^{\sum_{i \in M_0'} N_0^i} \beta^k.$$

Second, we used the inequality that for all $i, j, k$, it holds w.p.1 that $t_j^i(k) \geq t_j^i + k$. We now claim that for any policy $\pi$ which, when faced with some states which belong to S and some which do not will always (in a possibly very complicated and sophisticated way) pick some state in $S^c$, both of these inequalities become equalities. Indeed, we already noted that the first inequality becomes an equality if we always prioritize states not in S over states in S, as in that case, until all arms not initially in S hit S for the first time, we simply pull those arms (in a possibly very complicated way, but never pulling an arm that has already been to S at least once). To see that under any such $S^c$-favoring policy, it will hold (w.p.1) that for all $i, j, k$, $t_j^i(k) = t_j^i + k$, note that such a policy operates in two stages. First, it plays all arms not initially in S until they are all in S (again, possibly choosing among those arms in a complicated way over time until this occurs). At that time all arms are in S (unless the first stage lasts forever, which is also possible). From that time onward (supposing the first stage is finite), it will (again in a possibly complicated way) select some arm in S, and play that arm. If that arm stays in S, it will again look at all the arms, and play some arm in S. If any of the arms ever leaves S, it will by a simple contradiction be the only arm not in S, and hence must be prioritized and pulled until it gets back to S. That is by definition equivalent to enforcing $t_j^i(k) = t_j^i + k$ for all $i, j, k$.

Thus we find that the inequality for set $S$ is tight for any $S^c$-favoring policy. Equivalently, the inequality for set $S^c$ is tight for any $S$-favoring policy. Now, we make the connection to the family of so-called priority policies. A priority policy is a policy that fixes an ordering $\tau$ of the states, where this ordering depends only on the individual Markov chain associated with each arm, the associated vector of rewards, and the discount factor (not on the number of arms or the initial conditions), and when faced with a set of states for the $N$ arms, always selects the arm whose state appears highest in this ordering. Thus $\tau(1)$ refers to the most-favorable state, with $\tau(|\mathcal{S}|)$ referring to the least-favorable state. Note that there are $|\mathcal{S}|!$ such policies, one for each ordering of the states.

Note that, by a simple induction. a $\tau$-priority policy is equivalent to a policy that favors $\{\tau(1)\}$ over $\{\tau(1)\}^c$, AND favors $\{\tau(1), \tau(2)\}$ over $\{\tau(1), \tau(2)\}^c$, etc. Namely, a $\tau$-priority policy is equivalent to a policy that is $\{\bigcup_{i=1}^j \tau(i)\}$-favoring for $j = 1, \ldots, |\mathcal{S}| - 1$. It follows from the above that if the policy $\pi$ is a $\tau$-priority policy, then for all $N$ and $M_0$, $\overline{x}^{\pi,N,M_0}$ satisfies the inequalities associated with sets $\{\mathcal{S} \setminus \bigcup_{i=1}^j \tau(i)\}$, $j = 1, \ldots, |\mathcal{S}| - 1$ with equality, as well as the equality $\sum_{i \in \mathcal{S}} A_i^{\mathcal{S}} \overline{x}_i^{\pi,N,M_0} = b_{M_0}(\mathcal{S})$. But since $\{\mathcal{S} \setminus \bigcup_{i=1}^j \tau(i)\}$, $j = 1, \ldots, |\mathcal{S}|$ are nested, it follows that we may order these equalities s.t. the set of components of $\overline{x}^{\pi,N,M_0}$ with strictly positive coefficients is strictly increasing as we move down the list of equalities. But by straightforward linear algebra, this implies the constraint matrix for this set of equalities

is full-dimensional, and invertible. As of course $\overline{x}^{\pi,N,M_0}$ is also non-negative, it follows that incredibly, when $\pi$ is a $\tau$-priority policy, $\overline{x}^{\pi,N,M_0}$ is the unique solution to the set of equations

$$\sum_{i \in \mathcal{S} \setminus \bigcup_{k=1}^{j} \tau(k)} A_i^{\mathcal{S} \setminus \bigcup_{k=1}^{j} \tau(k)} \mathbf{x}_i = b_{M_0}(\mathcal{S} \setminus \bigcup_{k=1}^{j} \tau(k)) \quad , j = 1, \ldots, |\mathcal{S}| - 1;$$

$$\sum_{i \in \mathcal{S}} A_i^{\mathcal{S}} \mathbf{x}_i = b_{M_0}(\mathcal{S}).$$

One may have thought that analyzing the performance of such a priority-policy was very challenging, but amazingly we may compute its performance by solving this simple $|\mathcal{S}|$-dimensional system of equations. Furthermore, it follows that for every $\tau$-priority policy $\pi$, $\overline{x}^{\pi,N,M_0}$ is a basic feasible solution of our LP relaxation. Of course, the concern is that since we merely wrote down a relaxation, our relaxation may have other vertices that do not even correspond to any policy, not-to-mention a priority policy.

We now prove that remarkably, the aforementioned basic feasible solutions are the ONLY vertices of our LP relaxation. Let us reflect on what this means. It would mean that we had written down an LP relaxation, then proven that certain vertices correspond to vectors acheived by simple policies, then proven that THERE ARE NO OTHER VERTICES. Namely, although the vector for every feasible policy appears in the polytope, it would mean that none of those vectors are vertices, i.e. they can be acheived as convex combinations of what can be acheived by priority policies. Furthermore, it implies that for any given reward vector, there is an optimal priority policy, i.e. Gittin's index theorem.

We proceed as follows. First, we write down the dual. Second, we construct a particular feasible solution to the dual. Third, we construct a feasible primal solution corresponding to a particular priority policy (with the priority not depending on $N$ or $M_0$, only on the underlying Markov chain of each arm, the discount factor, and the reward vector) which, along with our particular dual solution, satisfies complementary slackness. It will then follow from the general complementary slackness theorem that our primal solution is optimal. As our procedure will work for any vector of rewards, it will follow that for any vector of rewards there exists an optimal priority policy. As for any given vertex there must exist some cost vector for which that vertex is uniquely optmal, it will follow that the basic feasible solutions corresponding to the priority policies are in fact the only vertices of our LP relaxation, i.e. our relaxation was actually not a relaxation at all and Gittin's index theorem holds.

Thus returning to our final LP formulation, let us take the dual. We find that the dual of our LP relaxation is:

$$\min_{\mathbf{y} \in \mathcal{R}^{2^{\mathcal{S}}}} \sum_{S \in 2^{\mathcal{S}}} b_{M_0}(S) \mathbf{y}_S,$$

s.t.

$$\sum_{S \in 2^{\mathcal{S}} : i \in S} A_i^S \mathbf{y}_S \geq R(i) \quad \text{for all} \quad i \in \mathcal{S};$$

$$\mathbf{y}_S \leq 0 \quad \text{for all} \quad S \in 2^{\mathcal{S}}, S \neq \mathcal{S}, \emptyset.$$

Note that $Y_{\mathcal{S}}$ is not sign-restricted. As discussed, we begin by identifying a particular dual-feasible solution $\mathbf{y}$ to use in a primal-dual complimentary slackness pair. Our dual solution will have exactly $|\mathcal{S}|$ non-zero values, one for each state $i$. Those $y_S$ with non-zero value will have a very special structure: namely, the sets will be nested. In particular, the non-zero elements of our dual solution will be $y_{\mathcal{S}}, y_{\mathcal{S} \setminus i_1}, y_{\mathcal{S} \setminus i_1, i_2}, \ldots, y_{\mathcal{S} \setminus i_1, \ldots, i_{|\mathcal{S}|-1}}$ for some ordering of the states $i_1, \ldots, i_n$. As it will be convenient, let us denote this "ordering" using the permutation $\tau$. Thus the identity of state $i_1$ equals $\tau(1)$, etc. In that case, the non-zero elements of our dual solution will be $y_{\mathcal{S}}, y_{\mathcal{S} \setminus \tau(1)}, y_{\mathcal{S} \setminus \tau(1), \tau(2)}, \ldots$. For such a vector $\mathbf{y}$, dual feasibility becomes:

$$A_{\tau(1)}^{\mathcal{S}} \mathbf{y}_{\mathcal{S}} \geq R\big(\tau(1)\big);$$

$$A_{\tau(2)}^{\mathcal{S}} \mathbf{y}_{\mathcal{S}} + A_{\tau(2)}^{\mathcal{S} \setminus \tau(1)} \mathbf{y}_{\mathcal{S} \setminus \tau(1)} \geq R\big(\tau(2)\big);$$

$$A^{\mathcal{S}}_{\tau(3)}\mathbf{y}_{\mathcal{S}} + A^{\mathcal{S}\setminus\tau(1)}_{\tau(3)}\mathbf{y}_{\mathcal{S}\setminus\tau(1)} + A^{\mathcal{S}\setminus\tau(1),\tau(2)}_{\tau(3)}\mathbf{y}_{\mathcal{S}\setminus\tau(1),\tau(2)} \geq R\big(\tau(3)\big);$$

$$\dots$$

$$A^{\mathcal{S}}_{\tau(k)}\mathbf{y}_{\mathcal{S}} + \sum_{i=1}^{k-1} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{\tau(k)}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)} \geq R\big(\tau(k)\big) \quad , \quad k = 4, \dots, |\mathcal{S}| - 1;$$

$$\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)} \leq 0, i = 1, \dots, |\mathcal{S}| - 1.$$

We will go one step further, and actually find a dual solution satisfying the above with equality (except for the non-positivity constraints, which may be satisfied with inequality), i.e. satisfying

$$A^{\mathcal{S}}_{\tau(k)}\mathbf{y}_{\mathcal{S}} + \sum_{i=1}^{k-1} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{\tau(k)}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)} = R\big(\tau(k)\big) \quad , \quad k = 1, \dots, |\mathcal{S}| - 1;$$

$$\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)} \leq 0, i = 1, \dots, |\mathcal{S}| - 1.$$

As before, for any given $\tau$ this gives a full-dimensional set of equalities. We note that apriori, it is not clear that for the given reward vector, there is any way to select $\tau$ (possibly depending on the reward vector) such that when one solves this system, one gets a feasible (i.e. satisfying the required non-positivity) solution. We now prove that for any given reward vector, there is a systematic way to identify such a $\tau$, and this shall yield the dual solution in our prima-dual pair.

We proceed as follows. The first equality tells us that

$$\mathbf{y}_{\mathcal{S}} = \frac{R\big(\tau(1)\big)}{A^{\mathcal{S}}_{\tau(1)}};$$

while the second equality tells us that

$$A^{\mathcal{S}}_{\tau(2)}\mathbf{y}_{\mathcal{S}} + A^{\mathcal{S}\setminus\tau(1)}_{\tau(2)}\mathbf{y}_{\mathcal{S}\setminus\tau(1)} = R\big(\tau(2)\big);$$

equivalently

$$\mathbf{y}_{\mathcal{S}\setminus\tau(1)} = \frac{R\big(\tau(2)\big) - A^{\mathcal{S}}_{\tau(2)}\mathbf{y}_{\mathcal{S}}}{A^{\mathcal{S}\setminus\tau(1)}_{\tau(2)}}.$$

Similarly, the third equality tells us that

$$A^{\mathcal{S}}_{\tau(3)}\mathbf{y}_{\mathcal{S}} + A^{\mathcal{S}\setminus\tau(1)}_{\tau(3)}\mathbf{y}_{\mathcal{S}\setminus\tau(1)} + A^{\mathcal{S}\setminus\tau(1),\tau(2)}_{\tau(3)}\mathbf{y}_{\mathcal{S}\setminus\tau(1),\tau(2)} = R\big(\tau(3)\big);$$

equivalently

$$\mathbf{y}_{\mathcal{S}\setminus\tau(1),\tau(2)} = \frac{R\big(\tau(3)\big) - \big(A^{\mathcal{S}}_{\tau(3)}\mathbf{y}_{\mathcal{S}} + A^{\mathcal{S}\setminus\tau(1)}_{\tau(3)}\mathbf{y}_{\mathcal{S}\setminus\tau(1)}\big)}{A^{\mathcal{S}\setminus\tau(1),\tau(2)}_{\tau(3)}}.$$

Note that, more generally, by a simple induction, the value of $\mathbf{y}_{\mathcal{S}\setminus\tau(1),\tau(2),\dots,\tau(k)}$ is determined by our choices of $\tau(1), \dots, \tau(k+1)$, with $y_{\mathcal{S}}$ determined by our choice of $\tau(1)$. Ideally, we would like a mechanism which ensures that, supposing we have picked $\tau(1), \dots, \tau(k)$ in such a manner to ensure that $\mathbf{y}_{\mathcal{S}\setminus\tau(1),\tau(2),\dots,\tau(i)} \leq 0$ for $i = 1, \dots, k-1$, it is somehow "easy" to select $\tau(k+1)$ from $\mathcal{S} \setminus \tau(1), \dots, \tau(k)$ in a manner to ensure that $\mathbf{y}_{\mathcal{S}\setminus\tau(1),\tau(2),\dots,\tau(k)} \leq 0$. If we had such a procedure, we could simply apply it inductively to derive a dual-feasible solution. We now derive an "extreme" version of such a procedure. In particular, we derive a procedure for iteratively selecting the elements of $\tau$, one at a time, such that once we have selected $\tau(1), \dots, \tau(k)$, we may select ANY element for $\tau(k+1)$ from $\mathcal{S} \setminus \tau(1), \dots, \tau(k)$ and still be certain that $\mathbf{y}_{\mathcal{S}\setminus\tau(1),\tau(2),\dots,\tau(k)} \leq 0$. Of course, instead of selecting "any element", we will want to select the element that will allow us "continue" the induction to the remaining equalities (not just the immediately next equality).

Let us begin by thinking on just $\tau(1)$. We would like to select $\tau(1)$ in such a way that any choice for

$\tau(2)$, from $\mathcal{S} \setminus \tau(1)$, will ensure that $\mathbf{y}_{\mathcal{S}\setminus\tau(1)} \leq 0$; equivalently that for all choices for $\tau(2) \in \mathcal{S} \setminus \tau(1)$, it holds that

$$\frac{R\big(\tau(2)\big) - A^{\mathcal{S}}_{\tau(2)}\mathbf{y}_{\mathcal{S}}}{A^{\mathcal{S}\setminus\tau(1)}_{\tau(2)}} \leq 0,$$

equivalently

$$R\big(\tau(2)\big) - A^{\mathcal{S}}_{\tau(2)}\mathbf{y}_{\mathcal{S}} \leq 0;$$

equivalently

$$y_{\mathcal{S}} \geq \frac{R\big(\tau(2)\big)}{A^{\mathcal{S}}_{\tau(2)}};$$

equivalently

$$\frac{R\big(\tau(1)\big)}{A^{\mathcal{S}}_{\tau(1)}} \geq \frac{R\big(\tau(2)\big)}{A^{\mathcal{S}}_{\tau(2)}}.$$

Of course, this is accomplished by setting $\tau(1)$ equal to the state MAXIMIZING (over ALL states $i$) $\frac{R\big(i\big)}{A^{\mathcal{S}}_{i}}$. Suppose we have done this, and select $\tau(1)$ as the corresponding argmax.

Next, let us consider $\tau(2)$. We would like to select $\tau(2)$ in such a way that for all $\tau(3) \in \mathcal{S} \setminus \tau(1), \tau(2)$,

$$\frac{R\big(\tau(3)\big) - \big(A^{\mathcal{S}}_{\tau(3)}\mathbf{y}_{\mathcal{S}} + A^{\mathcal{S}\setminus\tau(1)}_{\tau(3)}\mathbf{y}_{\mathcal{S}\setminus\tau(1)}\big)}{A^{\mathcal{S}\setminus\tau(1),\tau(2)}_{\tau(3)}} \leq 0,$$

equivalently

$$R\big(\tau(3)\big) - \big(A^{\mathcal{S}}_{\tau(3)}\mathbf{y}_{\mathcal{S}} + A^{\mathcal{S}\setminus\tau(1)}_{\tau(3)}\mathbf{y}_{\mathcal{S}\setminus\tau(1)}\big) \leq 0;$$

equivalently

$$\mathbf{y}_{\mathcal{S}\setminus\tau(1)} \geq \frac{R\big(\tau(3)\big) - A^{\mathcal{S}}_{\tau(3)}\mathbf{y}_{\mathcal{S}}}{A^{\mathcal{S}\setminus\tau(1)}_{\tau(3)}};$$

equivalently

$$\frac{R\big(\tau(2)\big) - A^{\mathcal{S}}_{\tau(2)}\mathbf{y}_{\mathcal{S}}}{A^{\mathcal{S}\setminus\tau(1)}_{\tau(2)}} \geq \frac{R\big(\tau(3)\big) - A^{\mathcal{S}}_{\tau(3)}\mathbf{y}_{\mathcal{S}}}{A^{\mathcal{S}\setminus\tau(1)}_{\tau(3)}}.$$

Of course, we can ensure this by selecting $\tau(2)$ to be the index which maximizes

$$\frac{R\big(i\big) - A^{\mathcal{S}}_{i}\mathbf{y}_{\mathcal{S}}}{A^{\mathcal{S}\setminus\tau(1)}_{i}}$$

over all states $i$ in $S \setminus \tau(1)$. More generally, note that

$$\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{k}\tau(j)} = \frac{R\big(\tau(k+1)\big) - A^{\mathcal{S}}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}} - \sum_{i=1}^{k-1} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}}{A^{\mathcal{S}\setminus\bigcup_{j=1}^{k}\tau(j)}_{\tau(k+1)}}.$$

Thus

$$\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{k}\tau(j)} \leq 0$$

is equivalent to

$$R\big(\tau(k+1)\big) - A^{\mathcal{S}}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}} - \sum_{i=1}^{k-1} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)} \leq 0,$$

itself equivalent to

$$\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{k-1}\tau(j)} \geq \frac{R\big(\tau(k+1)\big) - A^{\mathcal{S}}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}} - \sum_{i=1}^{k-2} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}}{A^{\mathcal{S}\setminus\bigcup_{j=1}^{k-1}\tau(j)}_{\tau(k+1)}},$$

itself equivalent to

$$\frac{R\big(\tau(k)\big) - A^{\mathcal{S}}_{\tau(k)}\mathbf{y}_{\mathcal{S}} - \sum_{i=1}^{k-2} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{\tau(k)}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}}{A^{\mathcal{S}\setminus\bigcup_{j=1}^{k-1}\tau(j)}_{\tau(k)}} \geq \frac{R\big(\tau(k+1)\big) - A^{\mathcal{S}}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}} - \sum_{i=1}^{k-2} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{\tau(k+1)}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}}{A^{\mathcal{S}\setminus\bigcup_{j=1}^{k-1}\tau(j)}_{\tau(k+1)}}.$$

Of course, this can be ensured by selecting $\tau(k)$ to be the state $l \in \mathcal{S}\setminus\tau(1),\ldots,\tau(k-1)$ which maximizes

$$\frac{R\big(l\big) - A^{\mathcal{S}}_{l}\mathbf{y}_{\mathcal{S}} - \sum_{i=1}^{k-2} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{l}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}}{A^{\mathcal{S}\setminus\bigcup_{j=1}^{k-1}\tau(j)}_{l}}.$$

The above may be formalized through a straightforward induction, the details of which we omit. Namely, we arrive at a feasible dual solution satisfying our desired requirements if we select $\tau$ as follows.

$\tau(1)$ is selected to the state maximizing $\frac{R\big(i\big)}{A^{\mathcal{S}}_{i}}$ over all states $i \in \mathcal{S}$, where we note that this is equivalent to the state of highest reward. As we wish to allow for general reward vectors, we note that this may result in $y_{\mathcal{S}}$ being positive or negative, but that this is fine as it is unrestricted in sign. Next, supposing we have selected $\tau(1),\ldots,\tau(k-1)$, we select $\tau(k)$ to be the state $l \in \mathcal{S}\setminus\tau(1),\ldots,\tau(k-1)$ maximizing

$$\frac{R\big(l\big) - A^{\mathcal{S}}_{l}\mathbf{y}_{\mathcal{S}} - \sum_{i=1}^{k-2} A^{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}_{l}\mathbf{y}_{\mathcal{S}\setminus\bigcup_{j=1}^{i}\tau(j)}}{A^{\mathcal{S}\setminus\bigcup_{j=1}^{k-1}\tau(j)}_{l}}.$$

We thus have our dual-feasible solution $\mathbf{y}^{\tau}$, where we note that (as per our above algorithm for constructing the dual solution) $\mathbf{y}^{\tau}$, and more generally the particular permutation $\tau$, depends only on the the specific Markov chain (for each state), the discount factor, and the specific vector of rewards, but NOT $N$ or $M_0$. We note that $\mathbf{y}^{\tau}_{S}$ is possibly non-zero only for $S = \mathcal{S}, \mathcal{S}\setminus\tau(1), \mathcal{S}\setminus\tau(1),\tau(2)$, etc. However, if we consider the primal feasible solution corresponding to the $\tau$-priority policy, this primal solution is binding for exactly the corresponding inequalities in the primal. As the dual solution is binding for every (non-sign-related) inequality, the dual solution is trivially binding for the corresponding non-zero components in the $\tau$-priority policy primal solution (which may be any number of the primal components). Thus it follows from complementary slackness that this primal-dual pair satisfies complementary slackness, and hence respective optimality (for the dual and primal). Thus the $\tau$-priority policy is optimal.

We further note that the above procedure not only proves existence of an optimal index policy for any given reward vector, but actually yields an EFFICIENT algorithm for computing the correspoding priority. Note that since the original LP had an exponential number of inequalities, efficient optimization was not a given (as even efficient separation was not obvious). Although we will not formally analyze the runtime, such an analysis has been done and it can be shown to yield a formally polynomial-time algorithm, whose runtime scales completely independent of N (and is polynomial in the input size of the Markov chain, rewards, etc.). As already stated, this then yields a very efficient algorithm for actually solving the Bayesian multi-arm bandit problem, which scales almost linearly in the number of arms (we will not conduct a precise analysis here). We also note that the above proof can be extended to the setting of an infinite (even uncountably infinite) number of states, although we do not pursue that here (we will use the fact that Gittin's theorem holds in essentially total generality).

Let us do part of the calculation for the example we considered previously. Namely, suppose there are

five states, 0,1,2,3,4. $R(0) = 2.5, P_{0,i} = \frac{1}{4}$ for $i = 1, 2, 3, 4$; $R(i) = i$ for $i = 1, 2, 3, 4$; $P_{i,i} = 1$ for $i = 1, 2, 3, 4$. Let us compute $\tau(1)$ and $\tau(2)$. We proceed as follows. We set $\tau(1) = 4, y_{\mathcal{S}} = 4$. We then solve

$$\max_{l=0,1,2,3} \frac{R(l) - y_{\mathcal{S}}}{A_l^{\{0,1,2,3\}}} = \max_{l=0,1,2,3} \frac{R(l) - 4}{E[\sum_{k=0}^{T_l^{\{0,1,2,3\}}-1} \beta^k]}.$$

To do this, we must compute $E[\sum_{k=0}^{T_l^{\{0,1,2,3\}}-1} \beta^k]$ for $l = 0, 1, 2, 3$. Note that $T_l^{\{0,1,2,3\}} = 1$ (w.p.1) for $l = 1, 2, 3$. $T_0^{\{0,1,2,3\}} = 1$ w.p. $\frac{3}{4}$, and equals $\infty$ w.p. $\frac{1}{4}$. Thus $E[\sum_{k=0}^{T_l^{\{0,1,2,3\}}-1} \beta^k] = 1$ for $l = 1, 2, 3$, and equals $\frac{3}{4} + \frac{1}{4} \times \frac{1}{1-\beta}$ for $l = 0$. Thus $\tau(2) = 0$ if $\frac{2.5-4}{\frac{3}{4}+\frac{1}{4}\times\frac{1}{1-\beta}}$ is larger than $\max_{l=1,2,3}(l-4) = -1$. Equivalently,

$$-\frac{3}{2} \geq -(\frac{3}{4} + \frac{1}{4} \times \frac{1}{1-\beta}),$$

equivalently

$$-\frac{3}{4} + \frac{1}{4} \times \frac{1}{1-\beta} \geq 0,$$

equivalently $3(1-\beta) \leq 1$, equivalently $\beta \geq \frac{2}{3}$, in agreement with our previous answer.

Although we could proceed mechanically, the calculations become somewhat cumbersome for a human (although can be easily coded up efficiently on a computer).

### 1.8.5 Review of Gittin's theorem

To review, we now formally state the Gittin's index theorem. We only state for the case in which the rewards are bounded and the underlying Markov chain is countable, but note that these results typically hold for much more general systems. We state only for the case in which all arms see the same chain, as we have already argued that this is w.l.o.g. We note as we want a statement that covers countably infinite Markov chains, and also clearly states how a policy should handle ties, we will map states to a table instead of a list (with ties broken by the second dimension). We note that when the Markov chain is finite, no such additional complexity is needed.

Gittin's index theorem (for the MCSP): For any given discount factor and any given underlying countable-state Markov chain (and associated set of state-dependent rewards) for which the rewards (associated to the different states) are uniformly bounded (over all states), the following is true. There exists a deterministic mapping $\pi$, from the states of the Markov chain to $[0, 1]^2$ (i.e. the unit box), such that an optimal policy is (in each period t) to examine the states of the chains in all arms, and play the arm whose state has the highest first component under this mapping. In the case of ties, of all arms whose state has the highest first component, play the arm whose state has the highest second component. In case two arms have the same exact state, play the arm with lowest index. Importantly, $\pi$ can be taken independent of the number of arms, the initial conditions, the random realizations of the process, and time itself, depending only on the discount factor, the underlying Markov chain of any one arm, and the reward vector associated to the states of that chain.

We now restate the theorem in the context of the original Bayesian MAB problem. We again w.l.o.g. restrict to the setting in which all arms have the same prior on reward distributions.

Gittin's index theorem (for the Bayesian MAB): For any given discount factor and any given prior over reward distributions s.t. there exists a finite or countably infinite and bounded set $\mathcal{Q}$ of real numbers s.t. the unconditional probability that a reward belongs to $\mathcal{Q}$ equals 1, the following is true. There exists a deterministic mapping $\pi$, from the collection of finite unordered subsets of $\mathcal{Q}$ to $[0, 1]^2$, such that an optimal policy is (in each period t) to examine the sequence of realized observations on each arm (as a finite unordered subsets of $\mathcal{Q}$), and play the arm whose set has the highest first component under this mapping. In the case of ties, of all arms whose set has the highest first component, play the arm whose set has the highest second component. In case two arms have the same exact history, play the arm with the lowest index. Importantly, $\pi$ can be taken independent of the number of arms, the initial conditions, the random realizations of the process, and time itself, depending only on the discount factor, and the underlying prior of any one arm.

## 1.9 Understanding the index as an optimization over stopping times

OK, so we have an elegant polyhedral way to derive the optimal permutation for any given problem. However, this sheds little inuitive light on e.g. how to think about, for the Beta-Bernoulli problem, how "good" any given state $(n_0, n_1)$ is. Once we know that an index policy is optimal, we can easily derive another "more convenient" way to think about how to rank the states as follows. Let us begin by thinking on the so-called $1\frac{1}{2}$-arm problem, in which one has two arms, one of which is associated with some non-trivial Markov chain (i.e. regular arm), the other of which is associated with a state with reward $x$ that loops to itself (i.e. known arm), and is effectively disconnected from the Markov chain of the other arm. In this case, when is it optimal to pull the non-trivial arm first? Note that as we restrict to stationary optimal policies, if we ever play the known arm, we will continue to play that known arm forever. Thus the overall MCSP optimization problem is equivalent to determining a. Do you pull the regular arm first, and b. If you pull the regular arm first, when do you stop pulling the regular arm and switch to the known arm (at which time you play the known arm forever). Note that any such decision must be based solely on which states the regular arm has visited so far, i.e. it must be adapted to the corresponding filtration generated by the states visited by the regular arm's Markov chain. In particular, w.r.t. this filtration, it must be a stopping time $\tau$. Namely, in the $1\frac{1}{2}$-arm bandit problem, the associated optimization reduces to deciding to either play the known arm the entire time, or play "the best" stopping time on the regular arm, and then switching to the known arm. What is the performance of a given stopping time $\tau$, letting $s_t$ denote the state of the regular arm's Markov chain after it has advanced $t$ time periods (with the initial state $s_0$)? One gets $\sum_{t=1}^{\tau} R(s_{t-1})\beta^t + \beta^\tau x\frac{\beta}{1-\beta}$, with expectation $E[\sum_{t=1}^{\tau} R(s_{t-1})\beta^t + \beta^\tau x\frac{\beta}{1-\beta}]$. Now, what does it mean for it to be optimal to play the regular arm first? It should be that there exists some stopping time $\tau$, which is at least 1 with probability 1, such that this expectation is at least $x\frac{\beta}{1-\beta}$, or at least that the supremum over all such stopping times is at least $x\frac{\beta}{1-\beta}$. For a given stopping time $\tau$,

$$E[\sum_{t=1}^{\tau} R(s_{t-1})\beta^t + \beta^\tau x\frac{\beta}{1-\beta}] \geq x\frac{\beta}{1-\beta}$$

iff

$$E[\sum_{t=1}^{\tau} R(s_{t-1})\beta^t + \beta^\tau x\frac{\beta}{1-\beta}] - x\frac{\beta}{1-\beta} \geq 0$$

iff

$$E[\sum_{t=1}^{\tau} R(s_{t-1})\beta^t + x\sum_{t=\tau+1}^{\infty}\beta^t - x\sum_{t=1}^{\infty}\beta^t] \geq 0$$

iff

$$E[\sum_{t=1}^{\tau} R(s_{t-1})\beta^t - x\sum_{t=1}^{\tau}\beta^t] \geq 0$$

iff

$$\frac{E[\sum_{t=1}^{\tau} R(s_{t-1})\beta^t]}{E[\sum_{t=1}^{\tau}\beta^t]} \geq x$$

iff

$$\frac{E[\sum_{t=0}^{\tau-1} R(s_t)\beta^{t+1}]}{E[\sum_{t=0}^{\tau-1}\beta^{t+1}]} \geq x$$

Thus let $T$ denote the set of all stopping times which are at least 1 w.p.1 for the initial condition $s_0$ (note that a stopping time may equal $\infty$). Note that by general results from the theory of MDP we may take $\tau$ to be a function only of the current state, not the entire history, and we further restrict T to such stopping times. In that case, it is strictly optimal to play the regular arm, starting in state $s_0$, first if

$$\sup_{\tau\in T}\frac{E[\sum_{t=0}^{\tau-1} R(s_t)\beta^{t+1}]}{E[\sum_{t=0}^{\tau-1}\beta^{t+1}]} > x.$$

To make the dependence on $s_0$ explicit, we further denote this as

$$\sup_{\tau \in T} \frac{E[\sum_{t=0}^{\tau-1} R(s_t)\beta^{t+1}|s_0]}{E[\sum_{t=0}^{\tau-1} \beta^{t+1}|s_0]} > x.$$

Similarly, it must be strictly optimal to play the known arm first if

$$\sup_{\tau \in T} \frac{E[\sum_{t=0}^{\tau-1} R(s_t)\beta^{t+1}|s_0]}{E[\sum_{t=0}^{\tau-1} \beta^{t+1}|s_0]} < x.$$

Now, a simple argument shows that we can in fact take the order on $\mathcal{S}$ induced by $\sup_{\tau \in T} \frac{E[\sum_{t=0}^{\tau-1} R(s_t)\beta^{t+1}|s_0]}{E[\sum_{t=0}^{\tau-1} \beta^{t+1}|s_0]}$ to BE the Gittin's order. Indeed, Note that for each $x \in \mathcal{R}$, we can augment the states of the Markov chain by a state which loops to itself and has reward x. As Gittin's theorem applies in total generality, it will also apply to this extended Markov chain. Note that since if none of the initial states are special in this way, the relevant optimization reduces to that with none of the x added (as the original Markov chain does not communicate with these states and visa-versa), the relative order of the original states in a priority policy which is optimal for this augmented chain must also be optimal for the original problem. As Gittin's theorem says that our preference list does not depend on the number of arms, the following must be true: If faced with just two arms, one of which is state $s_1$ and one of which is state $x$, it is strictly optimal to play state $s_1$ first; AND if faced with just two arms, one of which is state $s_2$ and one of which is state $x$, it is strictly optimal to play state $x$ first; THEN it must be that state $s_1$ appears before state $s_2$ in some optimal solution to the original problem. Now, suppose that (simultaneously) for every pair of states $(s_i, s_j)$, there existed some $x_{i,j}$ which determined the relative order of $s_i$ and $s_j$ in the same way. Then the same logic would dictate the entire ordering on the original states. Combining the above, and assuming no ties, we find that (as we know some index policy is optimal) we can thus take our optimal ordering to be dictated by the aforementioned index. Namely, to each state $s$, we assign the index

$$\sup_{\tau \in T} \frac{E[\sum_{t=0}^{\tau-1} R(s_t)\beta^{t+1}|s_0 = s]}{E[\sum_{t=0}^{\tau-1} \beta^{t+1}|s_0 = s]}$$

We then take the ordering of the indices as the optimal permutation. Indeed, this precise index is what is formally known as the Gittin's index. We note that in the case of ties, it can be shown that one gets an optimal policy as long as one imposes any fixed order on those states at time 0 and consistently plays the resulting order.