a

Show how 2.7.3 and 2.7.4 follow from 2.7.2.

$$V_\alpha(i) = \max\{\lambda + \alpha V_\alpha(i), r_i(c) + \alpha \sum_j p_{ij}(c) V_\alpha(j)\} \tag{2.7.2}$$

$$V_\alpha(i) = \max\{\frac{\lambda}{1-\alpha}, r_i(c) + \alpha \sum_j p_{ij}(c) V_\alpha(j)\} \tag{2.7.3}$$

$$V_\alpha(i) = \max\{\frac{\lambda}{1-\alpha}, \sup_{\tau>0} \mathbf{E}_i(\sum_{n=0}^{\tau-1} \alpha^n r X_n(c) + \alpha^\tau \frac{\lambda}{1-\alpha})\} \tag{2.7.4}$$

2.7.3 follows from 2.7.2

The state i denotes whether we are currently operating bandit 1 $i = 0$ or we are operating bandit 2 $i \in \{1, ..., S\}$. From the moment we start operating bandit one, we have that the maximum reward is $\lambda + \alpha V_\alpha(i)$. In the subsequent epochs, this will not change (because the state doesn't change) and it will remain optimal to keep operating bandit 1. To see why this is the case, notice that the second arm freezes, hence the right hand side in the maximization function does not change anymore. So once the it is optimal to pull arm 1, this will remain the case. Hence, we can rewrite the reward we will receive at the moment we switch from bandit 1 to bandit 2 as an infinite amount of discounted λ's:

$$V_\alpha(i) = \lambda + \alpha V_\alpha(i) = \lambda + \alpha(\lambda + \alpha V_\alpha(i)) = \lambda \sum_{s=0}^{\infty} \alpha^s = \frac{\lambda}{1-\alpha} \tag{2.7.2}$$

Hence 2.7.3 follows from 2.7.2.

2.7.4 follows from 2.7.2

Following the same logic, we can prove that 2.7.4 follows from 2.7.3 we can reformulate our bandit problem to a stopping problem where we receive a reward $rXn(c)$ to continue and reward $\frac{\lambda}{1-\alpha}$ to stop. Then it follows that there is a certain epoch, say τ in which we expect it would be optimal to switch from bandit 2 to bandit 1, hence when it is optimal to pull arm 2, we have the following expected reward:

$$V_\alpha(i) = \sup_{\tau>0} \mathbf{E}_i(\sum_{n=0}^{\tau-1} \alpha^n r X_n(c) + \alpha^\tau \frac{\lambda}{1-\alpha}) \tag{1}$$

The first term in the expectation, $\sum_{n=0}^{\tau-1} \alpha^n r X_n(c)$, represents the discounted rewards up until period $\tau - 1$ from pulling arm 1. The second term, $\alpha^\tau \frac{\lambda}{1-\alpha}$, represents the stopping reward discounted by α^τ because we switch to arm 1 in period τ.

b

We can view this problem as a 2 armed bandit problem where we keep administering medicine A (bandit arm 2), and receive expected reward $\frac{x+1}{x+y+2}$ until we stop administering medicine 2 and start using medicine B with reward p. Similarly to the two armed bandit problem, once we stop using medicine A, the state freezes and we receive no update in x and y. Therefore, if it is optimal to start using the known medicine B, it will remain optimal to do so. The optimality equation is therefore:

$$V_\alpha(x, y) = \max\{p + \alpha V_\alpha(x, y), \frac{x+1}{x+y+2} + \alpha \sum_{(x',y') \in \Xi} P_{(x,y)(x',y')}(c) V_\alpha(x', y')\} \tag{2}$$

Where Ξ represents the possible next states when administering medicine A, $\Xi = \{(x+1, y), (x, y+1)\}$. Similar to the two armed bandit, if at a certain point we start using medicine B, it will remain optimal to do so. Hence we can rewrite the left part of the maximization function as an infinite stream of p rewards discounted by α, which simplifies to $\frac{p}{1-\alpha}$ because it is a geometric series. So we can rewrite:

$$V_\alpha(x,y) = \max\{\frac{p}{1-\alpha}, \frac{x}{x+y+1} + \alpha \sum_{(x',y')\in\Xi} P_{(x,y)(x',y')}(c)V_\alpha(x',y')\} \tag{3}$$

Since there are only two possible states we after (x, y), we can write the right term with ease:

$$V(x,y)_\alpha = \max\left\{\frac{p}{1-\alpha}, \frac{x}{x+y+1} + \alpha\left[\frac{x+1}{x+y+2}V_\alpha(x+1,y) + \frac{y+1}{x+y+2}V_\alpha(x,y+1)\right]\right\} \tag{4}$$

We can also express the optimality equations using the Gittins index (with discount factor α) $G(x, y)$ of state (x, y):

$$V(x,y)_\alpha = \max\left\{\frac{p}{1-\alpha}, \frac{G(x,y)}{1-\alpha}\right\} \tag{5}$$

Which shows that the optimal policy is to continue with medicine A as long as the Gittins index of the state (x, y) is greater than p.

c

When the estimated cure probability $\frac{(x+1)}{(x+y+2)}$ of medicine A is much higher than the cure probability $p = 0.4$ of medicine B, then the optimal policy will be to continue with medicine A and the Gittins index of the corresponding state (x, y) will be higher than p. Reversely, when the estimated cure probability $\frac{(x+1)}{(x+y+2)}$ of medicine A is much lower then the cure probability $p = 0.4$ of medicine B then the optimal policy will be to stop with medicine A and the Gittins index of the corresponding state (x, y) will be lower than $p = 0.4$.
Therefore we can expect that for a given number of experiments $n = x_n + y_n$ with x_n cures and y_n non-cures there will be a critical state (x_n^*, y_n^*) with $\frac{(x_n^*+1)}{(x_n^*+y_n^*+2)} \approx 0.4$ such that $G(x_n, y_n) > 0.4$ for $x_n > x_n^*$ and $G(x_n, y_n) \leq 0.4$ for $x_n \leq x_n^*$.
Therefore only the Gittins indexes for states close to (x_n^*, y_n^*) are relevant for deciding to continue with medicine A or to stop.
Because we don't need the Gittins index for all states but for a (small) subset of states only, for the calculation of the Gittins indexes we prefer to use an online method instead of an offline method.

From the paper "2013-bandit-computations-annotated" we have selected the online-method Restart Formulation (Katehakis and Veinott).

According to the Restart Formulation we can calculate the Gittins index for a specific state $(\tilde{x}, \tilde{y})$ by formulating the optimality equation for a restart problem:

$$v = \max\{r^{(0)} + \alpha Q^{(0)} v, r^{(1)} + \alpha Q^{(1)} v\} \tag{6}$$

Here we have:

- v is the optimal value function (on the original state space (x, y)) for the restart problem.
- $r^{(0)}$ is the reward when restarting from $(\tilde{x}, \tilde{y})$.
- $Q^{(0)}$ is the transition matrix when restarting from $(\tilde{x}, \tilde{y})$.
- $r^{(1)}$ is the reward when continuing from (x, y).
- $Q^{(1)}$ is the transition matrix for continuing from (x, y).

We can then use Value Iteration (but also Policy Iteration or LP) to compute the value function v of the restart problem.
The Gittins index of $(\tilde{x}, \tilde{y})$ then follows from the computed value function as:

$$G(\tilde{x}, \tilde{y}) = (1-\alpha)v(\tilde{x}, \tilde{y}) \tag{7}$$

For our specific problem we have:

$$r^{(0)}(x,y) = \frac{\tilde{x}+1}{\tilde{x}+\tilde{y}+2}$$

$$Q^{(0)}_{(x,y),(x',y')} = \begin{cases} \frac{\tilde{x}+1}{\tilde{x}+\tilde{y}+2} & x' = \tilde{x}+1, y' = \tilde{y} \\ \frac{\tilde{y}+1}{\tilde{x}+\tilde{y}+2} & x' = \tilde{x}, y' = \tilde{y}+1 \\ 0 & \text{otherwise} \end{cases}$$

$$r^{(1)}(x,y) = \frac{x+1}{x+y+2}$$

$$Q^{(1)}_{(x,y),(x',y')} = \begin{cases} \frac{x+1}{x+y+2} & x' = x+1, y' = y \\ \frac{y+1}{x+y+2} & x' = x, y' = y+1 \\ 0 & \text{otherwise} \end{cases}$$

We have used Value Iteration to compute the value functions of the restart formulations.

The table below shows the computed Gittins index for selected states. The states are selected around critical points where the optimal action changes from "stop" to "continue" (with medicine A). For each state the table shows the cure probability $\frac{x_n+1}{x_n+y_n+2}$ as well.

n	(x_n, y_n)	$G(x_n, y_n)$	$\frac{x_n+1}{x_n+y_n+2}$	optimal action
1	$(0,1)$	0.500	0.333	continue
	$(1,0)$	0.800	0.667	continue
10	$(3,7)$	0.390	0.333	stop
	$(4,6)$	0.475	0.417	continue
20	$(7,13)$	0.397	0.364	stop
	$(8,12)$	0.443	0.409	continue
30	$(11,19)$	0.399	0.375	stop
	$(12,18)$	0.430	0.406	continue
40	$(15,25)$	0.399	0.381	stop
	$(16,24)$	0.423	0.405	continue
50	$(18,32)$	0.380	0.365	stop
	$(19,31)$	0.400	0.385	stop
	$(20,30)$	0.419	0.404	continue

When $G(x_n, y_n) > 0.4$ it's optimal to continue with medicine A and when $G(x_n, y_n) \leq 0.4$ it's optimal to stop.
The Gittins index is always higher or equal than the immediate reward from a state. Therefore for each state the Gittins index should be higher or equal than the cure probability (immediate reward). We see this confirmed in the table values.
We also see that for small n the difference between the Gittins index and the cure probability is higher than for larger values of n.
This makes sense because for small n the chance to get into a state with higher cure probability than 0.4 is higher than for larger values of n.
This means for lower n when there is more uncertainty about the cure probability there is a higher willingness to accept lower cure numbers (more exploration), which makes sense intuitively.

Code solutions

All code is available on our Github repo