Two Arms, One Known

Let Arm 2 be the known arm. Then, if it is optimal to pull Arm 2 at any point, then it is optimal to keep pulling Arm 2 from then on (this assumes a regular discount sequence: $\gamma_n / \sum_n^\infty \gamma_j$ is non-decreasing. Two important regular discount sequences are the finite horizon uniform and geometric discount sequences).

Intuition: we don't get any new information once we start pulling the known arm

Therefore, our expected reward is always at least as great later on in the process as it is at the beginning of the process.

An observation: this isn't always true with all unknown arms (but the last reward in a finite horizon case is larger in expectation).

Regular discount sequences: let's think about geometric (exponential) discounting

What does the observation above about keeping on pulling Arm 2 tell us?

The form of the optimal strategy must be either that you always pull Arm 2, or you keep pulling Arm 1 until some time, then switch to Arm 2, and then keep pulling Arm 2 forever!

Important theorem: let's do it for Bernoulli arms, although it can be generalized to other distributions.

For any regular discount sequence, and each distribution F on the parameter of the unknown arm, there exists a unique $\Lambda(F) \in [0,1]$ such that Arm 1 is optimal initially iff $\lambda \leq \Lambda(F)$ and Arm 2 is optimal otherwise

$$\Lambda(F) = \max_{\tau:\tau(\Phi)=1} \frac{E_\tau \sum_{m=1}^{M} \alpha^{m-1} X_m | F}{E_\tau \sum_{m=1}^{M} \alpha^{m-1}}$$

where M is the stage at which Arm 1 is used for the last time (possibly $+\infty$) before switching to Arm 2 when following strategy τ.

Optimal Policies for Multi-Armed Bandits

The celebrated theorem of Gittins and Jones: for geometric discounting and n independent arms, we can solve the problem by treating it as n different 2-armed Bandits, and computing the dynamic allocation indices for each of the known arms in the 2-armed bandits. Then at any time pick the arm with highest index. The really cool thing: the allocation index for each arm only depends on that arm!

However, this only holds for the geometric discount sequence!

Exercise: consider a 2-period 2-armed Bandit with Bernoulli arms:

$F_1 : (1/2)\delta_0 + (1/2)\delta_1$

$F_2 : (5/7)\delta_{1/2} + (2/7)\delta_1$

1 is preferred to 2. But if you introduce a third, known arm with probability anywhere between 2/3 and 31/46, Arm 2 is suddenly optimal at Time 1! This violates the independence we were talking about (and the two period discount sequence is $(1, 1, 0, 0, \ldots)$, which is regular

Style of the optimal strategy: keep playing an arm with highest index until it becomes lower than the second highest. Then switch to the second highest, and so on...