For simplicity, in the below we’ll assume the utilities $U_i$ are bounded, without loss of generality in $[0, 1]$.

**Example 1** (Bernoulli sampling). In the Bernoulli metalevel probability model, each arm will either succeed or not $U_i \in \{0, 1\}$, with an unknown latent frequency of success $\Theta_i$, and a set of stochastic simulations of possible consequences $\mathcal{E} = \{ E_{ij} \mid 1 \leq i \leq k, j \in \mathbb{N} \}$ that can be performed:

- $\Theta_i \overset{iid}{\sim} \text{Uniform}[0, 1]$ for $i \in \{1, \ldots, k\}$
- $U_i \mid \Theta_i \sim \text{Bernoulli}(\Theta_i)$ for $i \in \{1, \ldots, k\}$
- $E_{ij} \mid \Theta_i \overset{iid}{\sim} \text{Bernoulli}(\Theta_i)$ for $i \in \{1, \ldots, k\}, j \in \mathbb{N}$

The one-armed Bernoulli metalevel probability model has $k = 2$, $\Theta_1 = \lambda \in [0, 1]$ a constant, and $\Theta_2 \sim \text{Uniform}[0, 1]$.

A metalevel probability model, when combined with a cost of computation $c > 0$, defines a metalevel decision problem: what is the optimal strategy with which to choose a sequence of computations $E \in \mathcal{E}$ in order to maximize the agent’s net utility? Intuitively, this strategy should choose the computations that give the most evidence relevant to deciding which arm to use, stopping when the cost of computation outweighs the benefit gained. We formalize the selection problem as a Markov Decision Process (see, e.g., Puterman (1994)):

**Definition 2.** A (countable state, undiscounted) Markov Decision Process (MDP) is a tuple $M = (S, s_0, A_s, T, R)$ where: $S$ is a countable set of states, $s_0 \in S$ is the fixed initial state, $A_s$ is a countable set of actions available in state $s \in S$, $T(s, a, s')$ is the transition probability from $s \in S$ to $s' \in S$ after performing action $a \in A_s$, and $R(s, a, s')$ is the expected reward received on such a transition.

To formulate the metalevel decision problem as an MDP, we define the states as sequences of computation outcomes and allow for a terminal state when the agent chooses to stop computing and act:

**Definition 3.** Given a metalevel probability model $\langle U_1, \ldots, U_k, \mathcal{E} \rangle$ and a cost of computation $c > 0$, a corresponding metalevel decision problem is any MDP $M = (S, s_0, A_s, T, R)$ such that

$$S = \{ \bot \} \cup \{ e_1, \ldots, e_n \} : e_i \in E_i \text{ for all } i, \text{ for finite } n \geq 0 \text{ and distinct } E_i \in \mathcal{E}$$

$$s_0 = \{}$$

$$A_s = \{ \bot \} \cup \mathcal{E}$$

where $\bot \in S$ is the unique terminal state, where $\mathcal{E} \subseteq \mathcal{E}$ is a state-dependent subset of allowed computations, and when given any $s = (e_1, \ldots, e_n) \in S$, computational action $E \in \mathcal{E}$, and $s' = (e_1, \ldots, e_n, e) \in S$ where $e \in E$, we have:

$$T(s, E, s') = P(E = e \mid E_1 = e_1, \ldots, E_n = e_n)$$

$$R(s, E, s') = -c$$

$$R(s, \bot, \bot) = \max_i \mu_i(s)$$

where $\mu_i(s) = \mathbb{E}[U_i \mid E_1 = e_1, \ldots, E_n = e_n]$.

Note that when stopping in state $s$, the expected utility of action $i$ is by definition $\mu_i(s)$, so the optimal action to take is $i^* = \arg\max_i \mu_i(s)$ which has expected utility $\mu_{i^*}(s) = \max_i \mu_i(s)$.

One can optionally add an external constraint on the number of computational actions, or their total cost, in the form of a deadline or budget. This bridges with the related area of budgeted learning (Madani et al., 2004). Although this feature is not formalized in the MDP, it can be added by including either time or past total cost as part of the state.

**Example 2** (Bernoulli sampling). In the Bernoulli metalevel probability model (Example 1), note that:

$$\Theta_i \mid E_{i1}, \ldots, E_{i\text{in}} \sim \text{Beta}(s_i + 1, f_i + 1)$$

$$E_{i(n_i+1)} \mid E_{i1}, \ldots, E_{i\text{in}} \sim \text{Bernoulli} \left( \frac{s_i + 1}{n_i + 2} \right)$$

$$\mathbb{E}[U_i \mid E_{i1}, \ldots, E_{i\text{in}}] = (s_i + 1)/(n_i + 2)$$

by standard properties of these distributions, where $s_i = \sum_{j=1}^{n_i} E_{ij}$ is the number of simulated successes of arm $i$, and $f_i = n_i - s_i$ the failures. By Equation (1), the state space is the set of all $k$ pairs $(s_i, f_i)$; Equations (2) and (3) suffice to give the transition probabilities and terminal rewards, respectively. The one-armed Bernoulli case is similar, requiring as state just $(s, f)$ defining the posterior over $\Theta_2$.

Given a metalevel decision problem $M = (S, s_0, A_s, T, R)$ one defines policies and value functions as in any MDP. A (deterministic, stationary) metalevel policy $\pi$ is a function mapping states $s \in S$ to actions to take in that state $\pi(s) \in A_s$. 

---