Maximizing Expected Reward with State-Specific Controls: Closed-Form Results

Pavel Izhutov[1][0000−0002−1028−2297] and Haim Mendelson[2][0000−0003−0431−6072]

[1] Altos Protocol, Axio Inc and Stanford University,
[2] Stanford University

Abstract. Consider a decision problem with a finite set of states. The decision-maker takes a state-specific action that affects both the state probability and the reward of that state. We find a closed-form solution to the problem of maximizing the expected reward under the condition that an action that changes the probability of a state i does not affect the conditional probability of any other state given that state i does not occur.

1 Introduction

In a variety of applications, we seek to maximize the expected reward when the actions are state-specific and each action influences both the state probability and the reward of that state. In general, these problems are intractable. In this paper we provide a closed-form solution to this problem under the condition that when action a_i affects the probability of state i, all other state probabilities change proportionately so as to accommodate the change in the probability of state i. In other words, for any state $j \neq i$, the probability of j conditional on i not occurring remains the same.

2 Setting

There are N mutually exclusive states of the world $i = 1, 2, ..., N$ that occur with probabilities $\phi_i \geq 0 : \sum_{i=1}^{N} \phi_i = 1$. The system reward (value) in state i is v_i. Then, the expected system reward is given by

$$\sum_{i=1}^{N} v_i \phi_i.$$

Both the probabilities and the rewards depend on the actions taken by a decision-maker in each state i, which we denote by f_i. We assume that for all $i = 1, 2, \ldots, N$

$$v_i = v_i(f_i)$$

and

$$\phi_i(f_1, \ldots, f_N) = \frac{h_i(f_i)}{\sum_{j=1}^{N} h_j(f_j)},$$

where $h_i(\cdot), i = 1, \ldots, N$ are arbitrary non-negative functions. Under this structure,

$$\frac{\phi_j}{\phi_k} = \frac{h_j(f_j)}{h_k(f_k)} \text{ for all } j, k = 1, 2, \ldots, N,$$

in other words, the relative likelihood of two states depends only on the actions taken in those two states and are independent of the actions taken in other states. Thus, we assume that action a_i affects only what happens in state i, except that a change in the probability of state i has to adjust the probabilities of all other states proportionately.

Consider now the problem of maximizing the expected system reward:

$$\max_{f_1, \cdots, f_N} \sum_{i=1}^{N} v_i(f_i)\phi_i(f_1, \ldots, f_N) \tag{1}$$

$$s.t.(f_1, \ldots, f_N) \in \mathcal{A},$$

where $\mathcal{A}$ is the set of feasible actions. The dimensionality of the problem is $N \times D$, where N is the number of states and D is the dimensionality of the vector f. In general, this type of the problem cannot be decomposed due to the interdependencies among the state probabilities and actions. However, our problem structure allows us to decompose it into N D-dimensional optimization problems. And, if the D-dimensional problems can be solved as functions of a single parameter, the global solution is then reduced to solving one scalar equation. For example, when $D = 1$, the solution involves inverting separately N scalar functions and solving a scalar equation.

3 Key result

Theorem 1. *Consider the optimization problem (1), where* $\phi_i(f_1, \ldots, f_N) = \frac{h_i(f_i)}{\sum_{j=1}^{N} h_j(f_j)}$ *and the functions* $v(\cdot)$, $h_i(\cdot)$ *are twice continuously-differentiable for* $i = 1, 2, \ldots, N$. *Also let* $T_i(\cdot) = \frac{(v_i(\cdot)h_i(\cdot))'}{h_i'(\cdot)}$. *Then, if problem 1 has an optimal solution, its value is given by the largest root of the scalar equation*

$$V = \frac{\sum_{i=1}^{N} v_i(f_i(V))h_i(f_i(V))}{\sum_{i=1}^{N} h_i(f_i(V))}$$

among all potential roots V *obtained by plugging in all possible combinations* $(f_1(V), \ldots, f_N(V))$, *where* $f_i(V)$ *is on the boundary of* $\mathcal{A}$ *or it solves* $T_i(f_i) = V$. *The vector* $(f_1(V^*), \ldots, f_N(V^*))$ *that yields the highest* V^*, *is the optimal action vector.*

Remark 1. When actions are vector-valued, the equation $T_i(f_i) = V$ is understood as $\nabla f_i(v_i h_i) = V \nabla_{f_i} h_i$.

We can simplify the interpretation of the Theorem and derive additional practical insights using the following proposition.

Proposition 1. *Assume that the state space $\mathcal{A}$ is decomposable, i.e., $\mathcal{A} = \Pi_{i=1}^{N} \mathcal{A}_i$ and that the problem (1) has an optimal solution with value V^*. Then for $i = 1, 2, \ldots, N$, each component f_i^* of the solution $(f_1^*, \ldots, f_N^*)$ to problem (1) can be represented as a solution to the i-th subproblem*

$$\begin{aligned} max_{f_i} h_i\left(f_i\right)\left(v_i\left(f_i\right) - V^*\right) \\ s.t.\ f_i \in \mathcal{A}_i. \end{aligned} \tag{2}$$

That is, the optimal action in each state maximizes the weighted deviation from the global optimal value V^*.

More generally, we can decompose the space of system controls into state-specific and state-agnostic controls. If the optimization problem permits the decomposition into state-specific and state-agnostic optimization problems, we can use the state-specific optimization method provided here as the first step in the global optimization.

Proof of Theorem 1:

Proof. The first-order condition w.r.t. f_i is given by

$$\frac{\nabla\left(v_i h_i\right)\sum_{i=1}^{N} h_i - \nabla\left(h_i\right)\sum_{i=1}^{N}\left(h_i v_i\right)}{\left(\sum_{i=1}^{N} h_i\right)^2} = 0,$$

which can be rewritten as

$$\frac{1}{\sum_{i=1}^{N} h_i}\left[\nabla\left(v_i h_i\right) - \nabla\left(h_i\right)\frac{\sum_{i=1}^{N} h_i v_i}{\sum_{i=1}^{N} h_i}\right] = 0,$$

which implies

$$\nabla\left(v_i h_i\right)\Bigg|_{f_i} = \frac{\sum_{i=1}^{N} h_i v_i}{\sum_{i=1}^{N} h_i}\Bigg|_{f_i} \nabla h_i\Bigg|_{f_i} = V^* \nabla h_i\Bigg|_{f_i}, \tag{3}$$

where $V^* = V\left(f_1, \ldots, f_N\right)$ is the optimal value of the objective function. Let $f^* = \left(f_1^*, \ldots, f_N^*\right)$ denote the optimal action vector, then $f_i^*\left(V^*\right)$ either solves (3) or is on the boundary of $\mathcal{A}$. As a result, there is a set of candidates $\left\{f^j\left(V\right)\right\}$ for an optimal action vector f^*, where $f^j\left(V\right) = \left(f_1^j\left(V\right), \ldots, f_N^j\left(V\right)\right)$, and $f_i^j \in \mathcal{A}_i$ or f_i^j solves (3) for $i = 1, \ldots, N$. For each j we plug $f^j\left(V\right)$ in the expression for the objective $V = \frac{\sum_{i=1}^{N} v_i\left(f_i^j(V)\right)h_i\left(f_i^j(V)\right)}{\sum_{i=1}^{N} h_i\left(f_i^j(V)\right)}$ and solve for V. Then we arrive at the set of values $\left\{V^j\right\}$ corresponding to optimal action candidates $\left\{f^j\left(V\right)\right\}$. Obviously, the largest value $V^* \equiv V^{j^*} = \max_j V^j$ is the optimal value of the objective function, and, hence, the corresponding vector $f^{j^*}\left(V^{j^*}\right)$ is the optimal action vector.

Proof of Proposition 1:

Proof. If f_i^* is the solution to (2), then for all f_i

$$\left(v_i^* - V^*\right) h_i^* \geq \left(v_i - V^*\right) h_i, \tag{4}$$

where $v_i^* = v_i\left(f_i^*\right),\ h_i^* = h_i\left(f_i^*\right),\ v_i = v_i\left(f_i\right),\ h_i = h_i\left(f_i\right)$.
Plugging in

$$V^* = \frac{v_i^* h_i^* + \left(VH\right)_{-i}^*}{h_i^* + H_{-i}^*},$$

where $\left(VH\right)_{-i} = \sum_{-i} v_j h_j,\ H_{-i} = \sum_{-i} h_j$, into (4) we get

$$\left(v_i^* - \frac{v_i^* h_i^* + \left(VH\right)_{-i}^*}{h_i^* + H_{-i}^*}\right) h_i^* \geq \left(v_i - \frac{v_i^* h_i^* + \left(VH\right)_{-i}^*}{h_i^* + H_{-i}^*}\right) h_i$$

Rearranging this expression yields

$$\left(\frac{v_i^* h_i^* + \left(VH\right)_{-i}^* h_i^*}{h_i^* + H_{-i}^*}\right) \geq \frac{v_i h_i^* h_i + v_i H_{-i}^* h_i - v_i^* h_i^* h_i + \left(VH\right)_{-i}^* h_i}{h_i^* + H_{-i}^*}$$

$$\left(v_i^* - v_i\right) h_i^* h_i + \left(v_i^* H_{-i}^* - \left(VH\right)_{-i}^*\right) h_i^* - \left(v_i H_{-i}^* - \left(VH\right)_{-i}^*\right) h_i \geq 0$$

$$\left(v_i^* h_i^* + \left(VH\right)_{-i}^*\right)\left(h_i + H_{-i}^*\right) \geq \left(v_i h_i + \left(VH\right)_{-i}^*\right)\left(h_i^* + H_{-i}^*\right)$$

$$\frac{\left(v_i^* h_i^* + \left(VH\right)_{-i}^*\right)}{h_i^* + H_{-i}^*} \geq \frac{\left(v_i h_i + \left(VH\right)_{-i}^*\right)}{h_i + H_{-i}^*}.$$

Thus, f_i^* is the optimal solution for (1) as well.