Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

INVERSE OPTIMIZATION IN FINITE-STATE SEMI-MARKOV DECISION PROCESSES

Nhi Nguyen[1], and Archis Ghate[1]

[1]Dept. of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN, USA

ABSTRACT

Inverse optimization involves finding values of problem parameters that would render given values of decision variables optimal. For finite-state, finite-action Markov decision processes (MDPs), inverse optimization literature has focused on imputing two types of inputs: reward parameters and transition probabilities. All published work focuses on discrete-time MDPs. We study inverse optimization in semi-Markov decision processes (SMDPs) with continuous-time MDPs (CTMDP) as a special case. Specifically, we wish to find reward parameters that are as close as possible to given estimates and that render a given policy optimal. We utilize Bellman's equations for the forward SMDP to present a formulation of this inverse problem. This formulation is often convex. Simulation results on a batch manufacturing problem are included.

EXTENDED ABSTRACT

We study inverse optimization (Ahuja and Orlin 2001) in semi-Markov decision processes (SMDPs). SMDPs are an extension of continuous-time MDPs (CTMDPs), where times between state-transitions can have general distributions (Bertsekas 2005). We consider a discounted finite-state finite-action infinite-horizon SMDP. Assuming that the inter-transition distributions are well-behaved, an optimal stationary deterministic policy exists. Such an optimal policy is formed by actions that attain the maxima in Bellman's equations.

Our inverse problem is an extension of the CTMDP work in Nguyen and Ghate (2025). A target policy is given and we seek reward parameters that make it optimal. Specifically, rewards upon choosing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ are characterized by parameters $\theta \in \mathbb{R}^k$. Here, $\mathcal{S}$ is the state-space with $|\mathcal{S}| = n$ and A is a finite action-space. We denote rewards by $r(s, a; \theta)$. Assume the set $\Theta \subset \mathbb{R}^k$ of possible θ values is compact, and $r(s, a; \theta)$ is continuous in θ. We are given an estimate $\hat{\theta} \in \Theta$, a vector $\beta \in \mathbb{R}^k$ of positive weights, and a stationary deterministic policy $\hat{\pi}$. We wish to find $\theta \in \Theta$ such that the given policy $\hat{\pi}$ is optimal to the SMDP with rewards $r(s, a; \theta)$, and the β-weighted sum of absolute componentwise differences between θ and $\hat{\theta}$ is the smallest possible. All other problem-attributes are known. We formulate this as

$$\min_{\theta \in \Theta \subset \mathbb{R}^k, V \in \mathbb{R}^n} \sum_{j=1}^{k} \beta_j |\theta_j - \hat{\theta}_j|$$

$$V(s) = r(s, \hat{\pi}(s); \theta) + \sum_{j=1}^{n} m_{sj}(\hat{\pi}(s))\, V(j), \ \forall \mathcal{S}$$

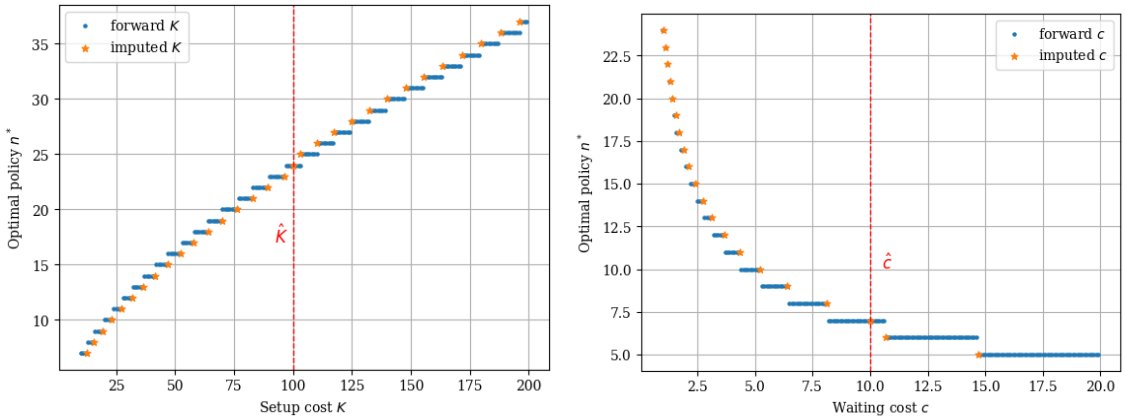
$$V(s) \geq r(s, a; \theta) + \sum_{j=1}^{n} m_{sj}(a)\, V(j), \ \forall (s, a) \in \mathcal{S} \times \mathcal{A} \text{ where } a \neq \hat{\pi}(s).$$

The first two constraints ensure that V satisfies Bellman's equations and that actions $\hat{\pi}(s)$, for all $s \in \mathcal{S}$, attain the maxima therein. This guarantees that V is the optimal value function and policy $\hat{\pi}$ is optimal. Quantities of the form $m_{sj}(a)$ in these constraints are given by a standard formula that depends on the continuous-time discount factor γ and parameters of the inter-transition distribution. It plays a role similar to discounted transition probabilities in discrete-time MDPs.

We illustrate with an example adapted from Bertsekas (2005). A manufacturer receives orders with inter-arrival times i.i.d. uniformly distributed on $[0, \tau_{\max}]$. While unfilled, each order incurs a waiting cost of $c > 0$ per unit time. A fixed setup cost of $K > 0$ is incurred immediately upon filling any orders, no matter the batch size. The state s equals the number of waiting orders. The manufacturer cannot accumulate more than $n-1$ orders. We are given cost estimates $\hat{c}, \hat{K}$ that belong to known intervals $[c_L, c_H]$ and $[K_L, K_H]$. A particular stationary deterministic policy $\hat{\pi}$ is also given. This $\hat{\pi}$ has an intuitive threshold structure: process all accumulated orders if $s \geq \hat{n}$ and wait if $s < \hat{n}$. For brevity, let $\alpha = \int_{\tau=0}^{\tau=\tau_{\max}} \frac{1-e^{-\gamma\tau}}{\gamma\tau_{\max}} d\tau$ and $\mu = \frac{1-e^{-\gamma\tau_{\max}}}{\gamma\tau_{\max}}$. Then the above generic inverse formulation reduces to the convex problem

$$
\begin{aligned}
&\min_{c\in[c_L,c_H],K\in[K_L,K_H],V\in\mathbb{R}^n} && \beta_1|c-\hat{c}| + \beta_2|K-\hat{K}| \\
&V(s) = -\alpha cs + \mu V(s+1), && s = 0,\ldots,\hat{n}-1 \\
&V(s) = -K + \mu V(1), && s = \hat{n},\hat{n}+1,\ldots,n-1 \\
&V(s) \geq -K + \mu V(1), && s = 1,\ldots,\hat{n}-1 \\
&V(s) \geq -\alpha cs + \mu V(s+1), && s = \hat{n},\hat{n}+1,\ldots,n-2.
\end{aligned}
$$

We report simulation results here. First, we solved multiple instances of the forward problem with $c = 1$ fixed and different $10 \leq K \leq 200$ to obtain a set of optimal thresholds. For each unique optimal threshold n^*, we then solved the inverse problem with $\hat{n} = n^*$, $\hat{c} = c = 1$, and $\hat{K} = 100$. Figure 1a shows that, for each n^*, the imputed K is the closest value to $\hat{K}$ within the set of K for which n^* was optimal. This is consistent with our objective of minimizing distance to the estimate. We repeated these simulations for variable $1 \leq c \leq 20$, $\hat{c} = 10$, and fixed $K = 100$, with similar results (Figure 1b).



(a) Fixed $c = 1$ and $10 \leq K \leq 200$.

(b) Fixed $K = 100$ and $1 \leq c \leq 20$.

Figure 1: Simulation results for imputing one cost while holding the other fixed.

This research was funded in part by the United States National Science Foundation grant #CMMI-2449675.

REFERENCES

Ahuja, R. K., and J. B. Orlin. 2001. "Inverse optimization." *Operations Research* 49(5):771–783.

Bertsekas, D. P. 2005. *Dynamic Programming and Optimal Control Vol. 1 and 2.*. 3rd ed., Belmont: Athena Scientific.

Nguyen, N., and A. Ghate. Forthcoming. "Inverse optimization in finite-state continuous-time Markov decision processes." In *Proceedings of the 2025 INFORMS Conference on Service Science*, July 1st–3rd, Oxford, United Kingdom.