

Sequential Resource Allocation Under Uncertainty: An Index Policy Approach

Weici Hu
Adviser: Peter Frazier

Cornell ORIE

July 7, 2017

Problem Setup

We consider an MDP $(\mathbb{S}^K, \mathbb{A}^K, \mathbb{P}, R)$ that consists of K identical sub-processes $(\mathbb{S}, \mathbb{A}, P, r)$, specifically,

- Time horizon $T < \infty$.
- State space \mathbb{S}^K is the cross-product of K \mathbb{S} . \mathbb{S} is assumed finite.
- Action space \mathbb{A}^K is the cross-product of K \mathbb{A} . $\mathbb{A} = \{0, 1\}$.
- Reward $R_t(\mathbf{s}, \mathbf{a}) = \sum_{x=1}^K r_t(s_x, a_x)$, $1 \leq t \leq T$, is additive of the reward of individual sub-processes.
- Transition probability $\mathbb{P}^{\mathbf{a}}(\mathbf{s}', \mathbf{s}) = \prod_{x=1}^K P^{a_x}(s'_x, s_x)$.
- Starts at an initial state \mathbf{s}_1 .

Problem Setup Con't

- A Markov policy $\pi : \mathbb{S}^K \times \mathbb{A}^K \times \{1, \dots, T\} \rightarrow [0, 1]$, with $\pi(\mathbf{s}, \mathbf{a}, t) = P(\mathbf{a} | \mathbf{S}_t = \mathbf{s})$ (Our decision).
We require $\sum_{\mathbf{a} \in \mathbb{A}^K} \pi(\mathbf{s}, \mathbf{a}, t) = 1, \forall \mathbf{s} \in \mathbb{S}^K, \forall 1 \leq t \leq T$.
- Objective

$$\begin{aligned} & \underset{\pi \in \Pi}{\text{maximize}} && \mathbb{E}^{\pi} \left[\sum_{t=1}^T R_t(\mathbf{S}_t, \mathbf{A}_t) \right] \\ & \text{subject to} && P^{\pi}(|\mathbf{A}_t| = m_t) = 1, \quad \forall 1 \leq t \leq T. \end{aligned} \tag{1}$$

Difficulty: Optimal solutions are computationally infeasible

Optimal solutions of (1) can be obtained with Bellman optimality equations.

But it requires $O(|\mathbb{S}|^K |\mathbb{A}|^K T)$ time complexity and $O(|\mathbb{S}|^K |\mathbb{A}|^K T)$ storage complexity.

The complexity grow exponentially with the number of sub-processes K , and becomes computationally infeasible for large K .

Past attempts

Pre-computations: 1. Optimal Lagrange Multiplier of (2)

Relax the original problem (1) to

$$\begin{aligned} & \underset{\pi \in \Pi}{\text{maximize}} && \mathbb{E}^{\pi} \left[\sum_{t=1}^T R_t(\mathbf{S}_t, \mathbf{A}_t) \right] \\ & \text{subject to} && \mathbb{E}^{\pi}(|\mathbf{A}_t|) = m_t, \quad \forall 1 \leq t \leq T. \end{aligned} \quad (2)$$

The Lagrangian relaxation of (2)

$$P(\lambda) = \max_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{t=1}^T R_t(\mathbf{S}_t, \mathbf{A}_t) \right] - \sum_{t=1}^T \lambda_t (\mathbb{E}^{\pi}[|\mathbf{A}_t|] - m_t). \quad (3)$$

Pre-computations: 1. Optimal Lagrange Multiplier of (2)

Decomposition of the Lagrangian relaxation

$$P(\lambda) = KQ(\lambda) + \sum_t \lambda_t m_t, \quad (4)$$

where

$$Q(\lambda) = \max_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{t=1}^T r_t(S_t, A_t) - \lambda_t A_t \right], \quad (5)$$

is the objective function for sub-process $(\mathbb{S}, \mathbb{A}, P, r)$. Definition of policy π is similar to π , with $\pi(s, a, t) = P(a|S_t = s)$.

Pre-computations: 1. Optimal Lagrange Multiplier of (2)

Optimal Lagrange Multiplier λ^* is a solution to the Lagrange dual

$$\min_{\lambda} P(\lambda) = K \left(\min_{\lambda} Q(\lambda) + \sum_t \lambda_t \frac{m_t}{K} \right), \quad (6)$$

which can be solved by the following linear program (LP):

$$\begin{aligned} & \min_{\{V(s,t), \lambda_t: s \in \mathbb{S}, t \in \{1, \dots, T\}\}} V(s_1, 1) + \frac{1}{K} \sum_t \lambda_t m_t \\ & \text{subject to} \quad V(s, t) - \sum_{s' \in \mathbb{S}} P^a(s, s') V(s', t+1) \geq r_t(s, a) - \lambda_t a, \\ & \quad \forall s \in \mathbb{S}, a \in \mathbb{A}, 1 \leq t \leq T-1 \\ & \quad V(s, T) \geq r_T(s, a) - \lambda_1 a \quad \forall s \in \mathbb{S}, a \in \mathbb{A} \end{aligned} \quad (7)$$

where V^* corresponds to the value function of problem (5).

Pre-computations: 1. Optimal Lagrange Multiplier of (2)

Remarks:

- Problem (7) has $O(|\mathcal{S}|T)$ variables and $O(|\mathcal{S}||\mathcal{A}|)T$ constraints, which is manageable.
- $P(\lambda)$ is a point-wise maximum of a group of affine functions of λ , hence convex in λ . (6) can also be solved by sub-gradient descent.

Pre-computations: 2. Occupation Measure ρ^* of an optimal policy of $Q(\lambda^*)$

Occupation measure of a policy π : the fraction of the time a process spent in each state-action pair at a time step under π .

To compute ρ^* , we solve the following linear program (LP):

$$\begin{aligned} \max_{\rho} \quad & \sum_{t=1}^T \sum_{a \in \mathbb{A}} \sum_{s \in \mathbb{S}} \rho(s, a, t) r_t(s, a) \\ \text{subject to} \quad & \sum_{s \in \mathbb{S}} \rho(s, 1, t) = \frac{m_t}{K}, \forall t = 1, \dots, T \\ & \sum_{a \in \mathbb{A}} \rho(s, a, t) - \sum_{a \in \mathbb{A}} \sum_{s' \in \mathbb{S}} \rho(s', a, t-1) P^a(s', s) = 0, \forall s \in \mathbb{S}, 2 \leq t \leq T \\ & \sum_{a \in \mathbb{A}} \rho(s, a, 1) = \mathbb{1}(s = s_1) \quad \forall s \in \mathbb{S} \\ & \rho(s, a, t) \geq 0, \quad \forall s \in \mathbb{S}, a \in \mathbb{A}, t = 1, \dots, T. \end{aligned}$$

(8)

Pre-computations: 2. Occupation Measure ρ^* of an optimal policy of $Q(\lambda^*)$

Lemma

Let ρ^ be an optimal solution to (8), then ρ^* is the occupation measure of an optimal policy to $Q(\lambda^*)$.*

A quick justification of Lemma 1:

(8) is the dual of (7) attains $Q(\lambda^*)$. By dynamic program theory ??????(Don't know what theory), solutions to (8) form the occupation measure of optimal solutions of $Q(\lambda^*)$.

Pre-computations: 2. Occupation Measure ρ^* of an optimal policy of $Q(\lambda^*)$

Remark:

- Any policy π^* constructed using ρ^* is an optimal solution to $Q(\lambda^*)$ and satisfies

$$\mathbb{E}^{\pi}[|A_t|] = \frac{m_t}{K}. \quad (9)$$

- Solving for ρ^* requires solving an LP with $|\mathcal{S}||\mathcal{A}|T$ variables and at most $T|\mathcal{S}|$ constraints.

Pre-computations: 3. Indices of states

We first describe a specific way of computing an optimal policy π^λ of $Q(\lambda)$:

Define value functions $V^\lambda : \mathbb{S} \times \{1, \dots, T\} \mapsto \mathbb{R}$ of $Q(\lambda)$ recursively by

$$V^\lambda(s, t) = \begin{cases} \max_{a \in \mathbb{A}} \{r_T(s, a) - a\lambda_T\} & \text{if } t = T, \\ \max_{a \in \mathbb{A}} \{r_t(s, a) - a\lambda_t + \sum_{s' \in \mathbb{S}} P^a(s, s') V^\lambda(s', t+1)\} & \text{o.w.} \end{cases} \quad (10)$$

Construct π^λ by

$$\pi^\lambda(s, 1, t) = \begin{cases} 1 & \text{if one-step lookahead value for } a = 1 \text{ is} \\ & \text{greater than or equal to } a = 0 \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Pre-computations: 3. Indices of states

Use $\mathbf{v}[c, t]$ to denote $\mathbf{v} + (c - v_t) * \mathbf{e}_t$.

The index of a state $s \in \mathbb{S}$ at time t is defined by

$$\beta_t(s) = \sup\{\beta : \pi^{\lambda^*[\beta, t]}(s, 1, t) = 1\} \quad (12)$$

We compute $\beta = \{\beta_t(s) : s \in \mathbb{S}, 1 \leq t \leq T\}$

Remark: Computing π^{λ} takes $O(|\mathbb{S}||\mathbb{A}|T)$ time and space.
Computational complexity for β is $O(|\mathbb{S}|^2|\mathbb{A}|T^2)$

Algorithm of Index policy $\hat{\pi}$

Pre-compute: λ^* ; β ; ρ . (Refer to earlier discussions for computation details)

for $t = 1, \dots, T$ **do**

Let $\beta_{t,[i]}$ be the i^{th} largest element in the list $\beta_t(\mathbf{S}_{t,1}), \dots, \beta_t(\mathbf{S}_{t,K})$, so $\beta_{t,[1]} \geq \dots \geq \beta_{t,[K]}$.

Let $\bar{\beta}_t = \beta_{t,[m_t]}$

Let $I_t = \{s : \beta_t(s) = \bar{\beta}_t \text{ and } s = \mathbf{S}_{t,x} \text{ for some } x\}$

Let $N_t(s) = |\{x : \mathbf{S}_{t,x} = s\}|$, for all s .

For $s \in I_t$, let

$$q(s) = \begin{cases} \frac{\rho(s,1,t)}{\sum_{s' \in I_t} \rho(s',1,t)}, & \text{if } \sum_{s' \in I_t} \rho(s',1,t) > 0 \\ \frac{N_t(s)}{\sum_{s' \in I_t} N_t(s')}, & \text{otherwise} \end{cases}$$

Let $b = \text{Rounding}(m_t - \sum_{s': \beta_t(s') > \bar{\beta}_t} N_t(s'), (q(s) : s \in I_t), (N_t(s) : s \in I_t))$

for all s **do**

If $\beta_t(s) > \bar{\beta}_t$, set all $N_t(s)$ sub-processes in s active.

If $\beta_t(s) = \bar{\beta}_t$, set $b(s)$ sub-processes in s active.

If $\beta_t(s) < \bar{\beta}_t$, set 0 sub-processes in s active.

end for

end for

Asymptotic optimality of the index policy $\hat{\pi}$

Notations:

- For $\alpha \in \mathbb{R}^T$, and $c \in \mathbb{R}$, define $\lfloor \alpha K \rfloor = (\lfloor \alpha_1 K \rfloor, \dots, \lfloor \alpha_T K \rfloor)$.
- Let $Z(\pi, \mathbf{m}, K)$ denote the expected reward of the original MDP (1), with constraint $\mathbf{m} = (m_1, \dots, m_T)$ and K sub-processes.
- Let $\Pi_{\mathbf{m}, K}$ denote the set of all feasible Markov policies of the original MDP (1), with constraint $\mathbf{m} = (m_1, \dots, m_T)$ and K sub-processes.

Theorem (1)

For any $\alpha \in [0, 1]^T$,

$$\lim_{K \rightarrow \infty} \frac{1}{K} \left(\max_{\pi \in \Pi_{\lfloor \alpha K \rfloor, K}} Z(\pi, \lfloor \alpha K \rfloor, K) - Z(\hat{\pi}, \lfloor \alpha K \rfloor, K) \right) = 0. \quad (13)$$

Asymptotic optimality of the index policy $\hat{\pi}$

Notations:

- Define $N_t(s)$ as the number of sub-processes in state s at time t , and $M_t(s)$ the number of sub-processes taking active actions in state s at time t , both under $\hat{\pi}$.
- Let π^* be a policy constructed using ρ^* .
- Let $P_t(s)$ denote the probability of being in state s at time t under π^* .

Theorem (2)

For every $s \in \mathbb{S}$ and $1 \leq t \leq T$,

$$\lim_{K \rightarrow \infty} \frac{N_t(s)}{K} = P_t(s), \quad P^{\hat{\pi}} - a.s., \quad (14)$$

and

$$\lim_{K \rightarrow \infty} \frac{M_t(s)}{K} = P_t(s) * \pi^*(s, 1, t), \quad P^{\hat{\pi}} - a.s., \quad (15)$$

Proof of Theorem 1

That $\hat{\pi}$ being feasible implies

$$Z(\hat{\pi}, \lfloor \alpha K \rfloor, K) \leq \max_{\pi \in \Pi_{\lfloor \alpha K \rfloor, K}} Z(\pi, \lfloor \alpha K \rfloor, K)$$

. Thus,

$$\lim_{K \rightarrow \infty} \frac{1}{K} Z(\hat{\pi}, \lfloor \alpha K \rfloor, K) \leq \lim_{K \rightarrow \infty} \frac{1}{K} \sup_{\pi \in \Pi_{\lfloor \alpha K \rfloor, K}} Z(\pi, \lfloor \alpha K \rfloor, K).$$

Proof of Theorem 1 Cont'

On the other hand,

$$\begin{aligned}\lim_{K \rightarrow \infty} \frac{1}{K} Z(\hat{\pi}, \lfloor \alpha K \rfloor, K) &= \lim_{K \rightarrow \infty} \frac{1}{K} \mathbb{E}^{\hat{\pi}} \left[\sum_{t=1}^T \sum_{s \in \mathbb{S}} r_t(s, 1) M_t(s) + r_t(s, 0) (N_t(s) - M_t(s)) \right] \\&= \sum_{t=1}^T \sum_{s \in \mathbb{S}} [r_t(s, 1) \rho(s, 1, t) + r_t(s, 0) \rho(s, 0, t)] \\&\quad - \mathbb{E}^{\pi^*} \left[\sum_t \lambda_t^* (A_t - \alpha_t) \right] \\&= Q(\lambda^*) + \alpha \sum \lambda_t^* \\&= \lim_{K \rightarrow \infty} \frac{1}{K} (K Q(\lambda^*) + \lfloor \alpha K \rfloor \sum \lambda_t^*) \\&= \lim_{K \rightarrow \infty} \frac{1}{K} P(\lambda^*, \lfloor \alpha K \rfloor, K) \\&\geq \lim_{K \rightarrow \infty} \frac{1}{K} \sup_{\pi \in \Pi_{\lfloor \alpha K \rfloor, K}} Z(\pi, \lfloor \alpha K \rfloor, K).\end{aligned}$$