Sequential Resource Allocation Under Uncertainty: An Index Policy Approach

Weici Hu
Adviser: Peter Frazier

Cornell ORIE

July 11, 2017

Problem Setup

We consider an MDP $(\mathbb{S}^K, \mathbb{A}^K, \mathbb{P}^\cdot, R)$ that consists of K identical sub-processes $(\mathbb{S}, \mathbb{A}, P^\cdot, r)$, 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 $\boldsymbol{\pi} : \mathbb{S}^K \times \mathbb{A}^K \times \{1, ..., T\} \to [0, 1]$, with $\boldsymbol{\pi}(\mathbf{s}, \mathbf{a}, t) = P(\mathbf{a}|\mathbf{S}_t = \mathbf{s})$ (Our decision).
  We require $\sum_{\mathbf{a}\in\mathbb{A}^K} \boldsymbol{\pi}(\mathbf{s}, \mathbf{a}, t) = 1.\ \forall \mathbf{s} \in \mathbb{S}^K, \forall 1 \le t \le T$.
- Objective

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

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

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

The Lagrangian relaxation of (2)

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

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

Decomposition of the Lagrangian relaxation

$$P(\boldsymbol{\lambda}) = KQ(\boldsymbol{\lambda}) + \sum_t \lambda_t m_t, \tag{4}$$

where

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

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

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

Optimal Lagrange Multiplier $\boldsymbol{\lambda}^*$ is a solution to the Lagrange dual

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

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

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

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

Remarks:

- Problem (7) has $O(|\mathbb{S}|T)$ variables and $O(|\mathbb{S}||\mathbb{A}|)T$ constraints, which is manageable.
- $P(\lambda)$ is an point-wise maxdimum of a group of affine functions of $\boldsymbol{\lambda}$, hence convex in $\boldsymbol{\lambda}$. (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):

$$\begin{aligned}
\max_{\rho} \quad & \sum_{t=1}^{T}\sum_{a\in\mathbb{A}}\sum_{s\in\mathbb{S}}\rho(s,a,t)r_t(s,a) \\
\text{subject to} \quad & \sum_{s\in\mathbb{S}}\rho(s,1,t) = \frac{m_t}{K}, \forall t = 1\ldots, T \\
& \sum_{a\in\mathbb{A}}\rho(s,a,t) - \sum_{a\in\mathbb{A}}\sum_{s'\in\mathbb{S}}\rho(s',a,t-1)P^a(s',s) = 0, \forall s\in\mathbb{S}, 2\le t\le T \\
& \sum_{a\in\mathbb{A}}\rho(s,a,1) = \mathbb{1}(s=s_1) \ \ \forall s\in\mathbb{S} \\
& \rho(s,a,t)\ge 0, \ \forall s\in\mathbb{S}, a\in\mathbb{A}, t=1,\ldots,T.
\end{aligned} \tag{8}$$

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

Lemma

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

A quick justification of Lemma 1:
(8) is the dual of (7) attains $Q(\boldsymbol{\lambda}^*)$. By dynamic program theory ??????(Don't know what theory), solutions to (8) form the occupation measure of optimal solutions of $Q(\boldsymbol{\lambda}^*)$.

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

Remark:

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

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

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

We first describe a specific way of computing an optimal policy $\pi^{\boldsymbol{\lambda}}$ of $Q(\boldsymbol{\lambda})$:

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

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

Construct $\pi^{\boldsymbol{\lambda}}$ by

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

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

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

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

Remark: Computing $\pi^{\boldsymbol{\lambda}}$ takes $O(|\mathbb{S}||\mathbb{A}|T)$ time and space.
Computational complexity for $\boldsymbol{\beta}$ is $O(|\mathbb{S}|^2|\mathbb{A}|T^2)$

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

Pre-compute: $\boldsymbol{\lambda}^*$; $\boldsymbol{\beta}$; ρ. (Refer to earlier discussions for computation details)

for $t = 1, ..., T$ **do**

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

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

Let $I_t = \{s : \beta_t(s) = \bar{\beta}_t$ and $s = \mathbf{S}_{t,x}$ 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 $b = \mathrm{Rounding}(m_t - \sum_{s' : \beta_t(s') > \bar{\beta}_t} N_t(s'), (q(s) : s \in I_t), (N_t(s) : s \in I_t))$

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

Notations:

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

Theorem (1)

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

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

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 $\hat{\boldsymbol{\pi}}$.
- 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 $1 \le t \le T$,

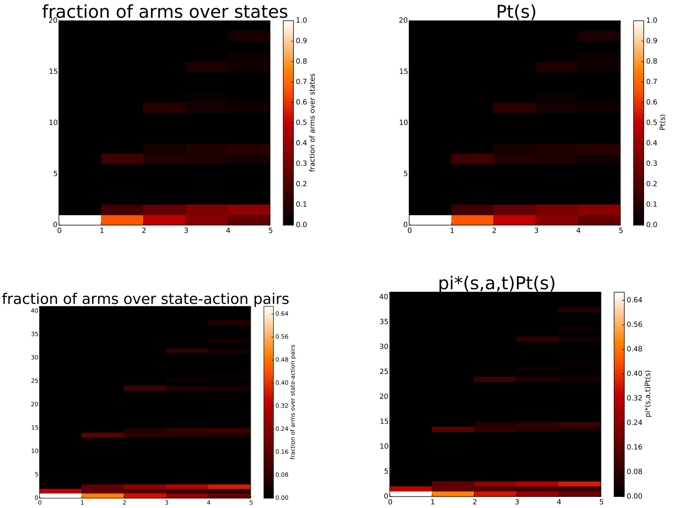
$$\lim_{K\to\infty} \frac{N_t(s)}{K} = P_t(s), \quad P^{\hat{\boldsymbol{\pi}}} - a.s., \tag{14}$$

and

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

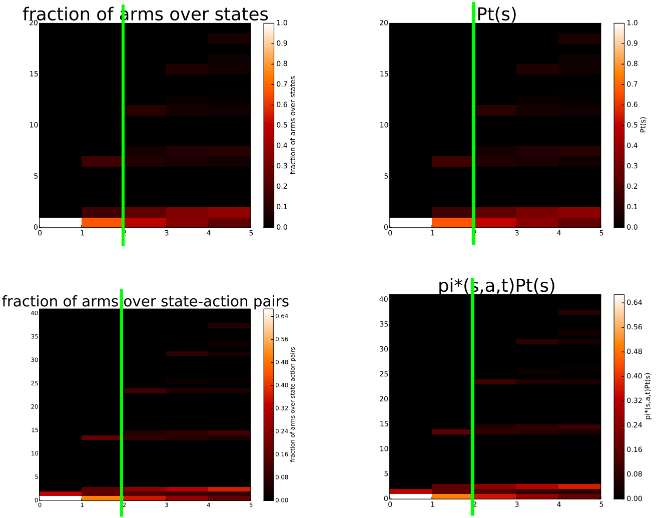
Theorem 2 is proven by induction over time

An numerical illustration of Theorem 2 using an instance of MAB



Theorem 2 is proven by induction over time

An numerical illustration of Theorem 2 using an instance of MAB



Proof of Theorem 1

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

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

. Thus,

$$\lim_{K \to \infty} \frac{1}{K} Z(\hat{\boldsymbol{\pi}}, \lfloor \alpha K \rfloor, K) \leq \lim_{K \to \infty} \frac{1}{K} \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\lfloor \alpha K \rfloor, K}} Z(\boldsymbol{\pi}, \lfloor \alpha K \rfloor, K).$$

On the other hand,

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