Restless Multi-arm Bandits and Optimality of Whittle Index

February 25, 2017

Plan

- Background (Infinite horizon average-cost MDPs)
- RMABs and Whittle index
- Example problem

MDPs

- Framework to solve sequential decision making problems, e.g., uplink scheduling problem
- Described by a tuple: $\{\mathcal{S}, \mathcal{T}, \mathcal{A}\}$
- Example: uplink scheduling over N Gilbert-Elliot channels
- Infinite horizon average cost MDP objective:

$$R_{\pi}(i) = \lim_{N \to \infty} \frac{1}{N} \mathbb{E} \left\{ \sum_{k=0}^{N-1} c\left(x_k, \mu_k(x_k)\right) \Big| x_0 = i \right\}$$

Bellman Equation [Prop. 7.4.1, Bertsekas]

For average cost per stage problem:

- The optimal average cost R^* is the same for all initial states and together with some vector $f^* = \{f^*(1), \ldots, f^*(n)\}$ satisfies Bellman's equation

$$R^* + f^*(i) = \min_{u \in U(i)} \left[c(i,u) + \sum_{j=1}^{n} p_{ij}(u) f^*(j) \right]$$

  for all $i = 1, \ldots, n$, and f^* is unique such that $f^*(n) = 0$.
- If $\mu(i)$ attains the minimum in above for all i, the stationary policy is optimal
- If a scaler R and a vector f satisfy the Bellman's equation then R is the average optimal cost
- Policy iteration: **Curse of dimensionality**

Multi-armed Bandits

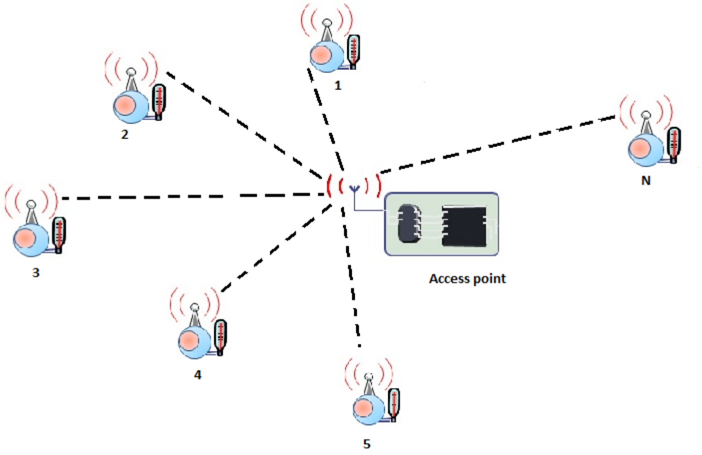
- 'L out of N' type sequential decision problems
- Simple Multi-armed bandits: only active projects/arms incur the cost and evolve.
- Restless multi-armed bandits: projects/arms which are not scheduled also evolve and incur the cost, e.g., N queues served by L servers
- Other variants: arm-acquiring bandits, hidden Markov bandits etc.
- In principle, can be solved using dynamic programming, but complexity increases exponentially in N
- There is an easier way (since arms are loosely coupled)

Whittle Index based policy for RMABs

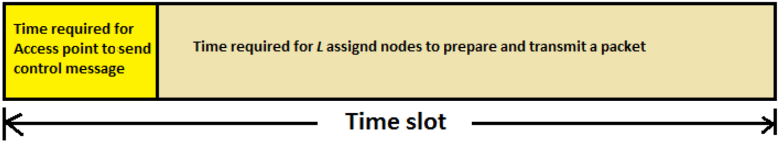
- Compute the Whittle index for each arm
- Choose arms with top L whittle index.
- Such policies are near-optimal, and can be shown to be asymptotically optimal as $N \to \infty$ with $\frac{L}{N}$ fixed.

Example

System Model



- **Assumption:** The Time is discrete.
- At most L sensors can simultaneously transmit in a time slot.



- **Channel :** unreliable

For client n :

- Packet success probability: $P_n \in (0, 1)$
- Each attempt consumes E_n units of energy

Problem statement

- **Ojectives:** regularity and energy-efficiency
- Designing a wireless scheduling policies that support the inter-delivery requirements of such wireless clients in an energy-efficient way.
- The QoS requirement of client n is specified through an integer , the *packet inter-delivery time threshold* τ_n.

Access point Goal: To select at most L clients to transmit in each time-slot from among the N clients, so as to minimize the cost function.

Cost function

The cost function incurred by the system during the time interval $\{0,1,2,...,T\}$ is given by,

$$E\Big[\sum_{n=1}^{N}\Big(\sum_{i=1}^{M_T^{(n)}}(D_i^{(n)}-\tau_n)^+ + \left(T - t_{D_{M_T^{(n)}}^{(n)}} - \tau_n\right)^+ + \eta \hat{M}_T^{(n)} E_n\Big)\Big] \qquad (1)$$

$D_i^{(n)}$: time between the deliveries of the i-th and (i+1)-th packets for client n.
$M_T^{(n)}$: The number of packets delivered for the n-th client by the time T.
$t_{D_i^{(n)}}$: Time slot in which the i-th packet for client n is delivered.
$\hat{M}_T^{(n)}$: Total number of slots in $\{0,1,...,T\text{-}1\}$ in which the n-th client is selected to transmit.
η : energy efficiency parameter.

Reduction to Finite state problem

- The system state at time-slot t is denoted by a vector $X(t) := (X_1(t), ..., X_N(t))$.
  where $X_n(t)$: Time elapsed since the latest delivery of client $n's$ packet.
- The Action at time t is $U(t) := (U_1(t), ..., U_N(t))$, with $\sum_{n=1}^{N} U_n(t) \leq L$

$$U_n(t) = \begin{cases} 1 & \text{if client n is selected to transmit in slot t,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The system state evolve as,

$$X_n(t+1) = \begin{cases} 0 & \text{if a packet of client n is delivered in t} \\ X_n(t)+1 & \text{otherwise.} \end{cases} \tag{3}$$

- The system forms a controlled Markov chain(MDP-1), with the transition probabilities given by,

$$\begin{aligned} P_{\mathbf{x},\mathbf{y}}^{MDP-1}(\mathbf{u}) &:= P[X(t+1) = \mathbf{y} | X(t) = \mathbf{x}, U(t) = \mathbf{u}] \\ &= \prod_{n=1}^{N} P[X_n(t+1) = y_n | X_n(t) = x_n, U_n(t) = u_n] \end{aligned} \quad (4)$$

$$P[X_n(t+1) = y_n | X_n(t) = x_n, U_n(t) = u_n] := \begin{cases} p_n & \text{if } y_n = 0 \text{ and } u_n = 1, \\ 1 - p_n & \text{if } y_n = x_n + 1 \text{ and } u_n = 1, \\ 1 & \text{if } y_n = x_n + 1 \text{ and } u_n = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

- The optimal cost-to-go function for MDP-2 is,

$$V_T(\mathrm{x}) := \min_{\pi:\Sigma_n U_n(t)\leq L} \mathrm{E}\{\sum_{t=0}^{T-1}\sum_{n=1}^{N}(\eta E_n U_n(t) + 1\{Y_n(t)=\tau_n\} | Y(0)=\mathrm{x}\}, \forall \mathbf{x} \in \mathbb{Y} \quad (5)$$

- **Theorem 4**: MDP-2 is equivalent to the MDP-1 in that:
  1. MDP-2 has the same transition probabilities as the accompanying process of MDP-1, i.e., the process $X(t) \wedge \tau$;
  2. Both MDPs satisfy the recursive relationship in (3); thus, their optimal cost-to-go functions are equal for each starting state x with $x_n \leq \tau_n$;
  3. Any optimal control for MDP-1 in state x is also optimal for MDP-2 in state $x \wedge \tau$

  The Dynamic Programming recursion for the optimal cost in MDP-2 is

$$V_T(\mathrm{x}) = \min_{\mathrm{u}:\Sigma_n u_n \leq L} \mathrm{E}\{\sum_n (\eta E_n u_n + 1\{x_n = \tau_n\}) + \sum_{\mathrm{y}} P_{\mathrm{x,y}}^{\mathrm{MDP-2}} V_{T-1}(\mathrm{y})\}. \quad (6)$$

Formulation of Restless Multi-armed bandit Problem

Notations:

- $\alpha = \frac{L}{N}$, Maximum fraction of clients that can simultaneously transmit.
- $Y_n(t)$ associated with client n is denoted as *project n*.
- $U_n(t) = 1$, if the project n is active in slot t.
- $U_n(t) = 0$, if the project n is passive in slot t.
  The infinite-horizon problem is to solve, with $Y(0) = \mathbf{x} \in \mathbb{Y}$,

$$\max_{\pi} \lim_{T \to +\infty} \inf \frac{1}{T} \mathrm{E}[\sum_{t=0}^{T-1} \sum_{n=1}^{N} -1\{Y_n(t) = \tau_n\} - \eta E_n U_n(t)] \tag{7}$$

$$s.t. \sum_{n=1}^{N} (1 - U_n(t)) \geq (1 - \alpha) N, \forall t. \tag{8}$$

Relaxations:

We consider an associated relaxation of the problem which puts a constraint only on the *time average* number of active projects allowed:

$$\max_{\pi} \lim_{T\to+\infty}\inf \frac{1}{T}\mathrm{E}[\sum_{t=0}^{T-1}\sum_{n=1}^{N} -1\{Y_n(t)=\tau_n\} - \eta E_n U_n(t)] \tag{9}$$

$$s.t. \lim_{T\to+\infty}\inf \frac{1}{T}\mathrm{E}[\sum_{t=0}^{T-1}\sum_{n=1}^{N}(1-U_n(t))] \geq (1-\alpha)N. \tag{10}$$

Let us consider the Lagrangian associated with the problem (9)-(10), with $Y(0)=\mathbf{x}\in\mathbb{Y}$,

$$\begin{aligned} l(\pi,\ \omega) := & \lim_{T\to+\infty}\inf \frac{1}{T}\mathrm{E}_{\pi}[\sum_{t=0}^{T-1}\sum_{n=1}^{N} -1\{Y_n(t)=\tau_n\} - \eta E_n U_n(t)] \\ & +\omega \lim_{T\to+\infty}\inf \frac{1}{T}\mathrm{E}_{\pi}[\sum_{t=0}^{T-1}\sum_{n=1}^{N}(1-U_n(t))] - \omega(1-\alpha)N, \end{aligned}$$

π: History dependent scheduling policy.

$\omega \geq 0$: Lagrangian multiplier

The Lagrangian dual function is $d(\omega) := max_{\pi} l(\pi, \omega)$:

$$
\begin{aligned}
d(\omega) \leq \max_{\pi} \liminf_{T \to +\infty} \frac{1}{T} \mathrm{E}[\sum_{t=0}^{T-1} \sum_{n=1}^{N} -1\{Y_n(t) = \tau_n\} \\
-\eta E_n U_n(t) + \omega(1 - U_n(t)) | Y(0) = \mathrm{x}] - \omega(1-\alpha)N \\
\leq \max_{\pi} \limsup_{T \to +\infty} \frac{1}{T} \mathrm{E}[\sum_{t=0}^{T-1} \sum_{n=1}^{N} -1\{Y_n(t) = \tau_n\} \\
-\eta E_n U_n(t) + \omega(1 - U_n(t)) | Y(0) = \mathrm{x}] - \omega(1-\alpha)N \\
\leq \max_{\pi} \sum_{n=1}^{N} \limsup_{T \to +\infty} \frac{1}{T} \mathrm{E}[\sum_{t=0}^{T-1} -1\{Y_n(t) = \tau_n\} \\
-\eta E_n U_n(t) + \omega(1 - U_n(t)) | Y(0) = \mathrm{x}] - \omega(1-\alpha)N, \qquad (11)
\end{aligned}
$$

equation (11) is the unconstrained problem.

It can be viewed as a composition of N independent ω-*subsidy* problems interpreted as follows: For each client n, besides the original reward -$\mathbf{1}\{Y_n(t) = \tau_n\} - \eta E_n U_n(t)$, when $U_n(t) = 0$, it receives a subsidy ω for being passive.
Thus, the ω-*subsidy problem* associated with client n is defined as,

$$R_n(\omega) = \max_{\pi_\mathrm{n}} \lim_{T\to+} \sup_{\infty} \frac{1}{T}\mathrm{E}[\sum_{t=0}^{T-1} -1\{Y_n(t) = \tau_n\}$$
$$-\eta E_n U_n!(t) + \omega(1 - U_n(t))|Y_n(0) = x_n], \qquad (12)$$

where π_n is a history dependent policy which decides the action $U_n(t)$ for client n in each time-slot.
We first solve this ω-subsidy problem, and then explore its properties to show that strong duality holds for the relaxed problem (9)-(10), and thereby determine the optimal relaxed policy.

- For $\theta \in \{0, 1, ..., \tau_n\}$ and $\rho \in [0, 1]$,
  we define $\sigma_n(\theta, \rho)$ to be a threshold policy for project n, as follows:The policy $\sigma_n(\theta, \rho)$
  at time t,
  $Y_n(t) < \theta$:Project is Passive i.e., $U_n(t) = 0$
  $Y_n(t) > \theta$:Project is Active i.e., $U_n(t) = 1$
  If $Y_n(t) = \theta$: then, Project stays Passive with Probability ρ, and is activated with probability $1 - \rho$.
- For each project n, associate a function defined as,

$$W_n(\theta) := p_n(\theta + 1)(1 - p_n)^{\tau_n - (\theta+1)} - \eta E_n, \tag{13}$$

- The Whittle Index $W_n(i)$of project n at state i is defined as the value of the subsidy that makes the passive and active actions equally attractive for the ω-subsidy problem associated with project n in state i.
  When $\omega = W_n(i)$ The following holds the optimality,

$$-\eta E_n + p_n f(0) + (1 - p_n) f((i + 1) \wedge \tau_n) = \omega + f((i + 1) \wedge \tau_n)$$

- The n-th project is said to be indexable if:
  - $B_n(\omega)$ be the set of states for which project n is passiveunder an optimal policy corresponding ω-subsidy problem.
  - Project n is indexable if, as ωincreases from $-\infty to + \infty$, the set $B_n(\omega)$ increases monotonically from ϕ to the whole space.
- **Lemma 5:** Consider the ω-subsidy problem(12), for project n. Then,
  - $\sigma_n(0,0)$ is optimal iff the subsidy $\omega \leq W_n(0)$.
  - For $\theta \in \{1, ..., \tau_n - 1\}$ is optimal iff the subsidy ω satisfies $W_n(\theta - 1) \leq \omega \leq W_n(\theta)$.
  - $\sigma_n(\tau_n, 0)$is optimal iff $\omega = W_n(\tau - 1)$.
  - $\sigma_n(\tau_n, 1)$is optimal iff $\omega \geq W_n(\tau - 1)$.
    In addition, for $\theta \in \{1, ..., \tau_n - 1\}$, the policies $\{\sigma_n(\theta, \rho) : \rho \in [0, 1]\}$ are optimal when,
    1. $0 \leq \theta \leq \tau - 1$ and $\omega = W_n(\theta)$,
    2. $\theta = \tau$ and $\omega = W_n(\tau - 1)$.

    Furthermore, for any $\theta \in \{1, ..., \tau\}$, under the $\sigma(\theta, 0)$ policy, the average reward earned is,

$$\frac{p_n\theta\omega - \eta E_n - (1 - p_n)^{\tau_n - \theta}}{1 + \theta p_n}. \quad (14)$$

- Consider the ω subsidy problem for project n,and denote by $a_n(\theta, \rho)$ the average proportion of time that the active action is taken under the policy $\sigma_n(\theta, \rho)$,i.e.,
  $a_n(\theta, \rho) := \lim_{T\to+\infty} \frac{1}{T} E_{\sigma_n(\theta,\rho)}[\sum_{t=0}^{T-1} U_n(t)]$.
  Let $a_{n,min}(\omega) := \min_{\theta,\rho}\{a_{n,min}(\theta, \rho)$:
  $\sigma_n(\theta, \rho)$is optimal when the subsidy is $\omega\}$.
- **Theorem 7:**For the relaxed problem (9)-(10) and its dual F$d(\omega)$, the following results hold:
  - The dual function $d(\omega)$ satisfies,

$$d(\omega) = \sum_{n=0}^{N-1} R_n(\omega) - \omega(1-\alpha)N. \tag{13}$$

  - Strong duality holds, i.e., the optimal average reward for the relaxed problem, denoted R_{rel}, satisfies,
    $R_{rel} = \min_{\omega\geq 0} d(\omega)$
  - Define policy $\sigma(\theta, \rho)$as the one that applies $\sigma_n(\theta_n, \rho_n)$ to each project n. Then, for each $\alpha \in [0, 1]$,there exist vectors $\theta^*$and ρ^* such that $\sigma(\theta^*, \rho^*)$
  - In addition, $d(\omega)$ is a convex and piecewise linear function of ω. Thus, the value of R_{rel} can be easily solved.

Properties of $d(\omega)$:

- Each $R_n(\omega)$ is a piecewise linear function.
- To prove convexity of $R_n(\omega)$, note that the reward earned by any policy is a linear function of ω, and the supremum of linear functions is convex. Thus, $d(\omega)$ is also convex and piecewise linear.
- The value of R_{rel}, which is the minimum value of this known,convex, and piecewise linear function $d(\omega)$,can be easily obtained.

Whittle index policy: At the beginning of each time slot t, client n is scheduled if its whittle index $W_n(Y_n(t))$ is positive, and moreover, is within the top αN index values of all clients in that slot. Now not more than αN clients are simultaneously scheduled.

Thank you