Whittle index based Q-Learning for restless bandits with average reward by Avrachenkov and Borkar

Navdeep Kumar

Indian Institute of Science

September 21, 2020

Multi Arm Bandit (rested and restless)

- N Controlled Markov chains $\{X_n^i, n \geq 0\}$ on state space $\mathrm{S} = \{1,2,..,d\}$, where $1 \leq i \leq N, \quad X_n^i \in S$
- Binary Control: $U = \{0, 1\}$
- Reward dependant on control, arm, and state: $\mathrm{R}_u^i(X)$
- Transition probabilities:
  $p_u^i(j,k) = p(X_{n+1}^i = k | X_n^i = j, U_n^i = u)$
  $i, j \in S, u \in U, F_n = \sigma(X_m^a, U_m^a, 1 \leq a \leq N, m \leq n), p(X_{n+1}^i = k | F_n) = p_u^i(X_n^i, k)$
- Objective to maximize: $liminf_{n \to \infty} \frac{1}{n} \mathbf{E}[\sum_{i=1}^N \sum_{m=0}^{n-1} R_{U_m^i}^i(X_m^i)]$
- Constraint: On number of active arm. To be discussed later.

Rested MAB (Special Case)

- Reward is zero for control u=0: $R_0^i(X) = 0$
- States freezes for control u=0: $p_0^i(j,k) = \delta(j,k)$
- Constraint: m active arm each time: $\sum_{i=1}^{N-1} U_n^i = m \quad \forall n$
- Optimal solution: Gittins Index Policy
- Gittins value at state s, for arm i:
  $g^i(s) = sup_{T>0} \quad \frac{1}{T}\mathbf{E}[\sum_{t=0}^{T-1} R_{U_t^i}^i(X_t^i) \quad |X_0^i = s]$
- Gittins Index Policy : Arrange arms in decreasing order according to Gittins value in their current states, pick top m arms for u=1, rest u=0
- Proof : Intuitive greedy approach.

reference: R. Weber, On the Gittins Index for Multiarmed Bandits, 1992.

Restless MAB (General Case)

- Reward may not be zero for control u=0: $R_0^i(X) \neq 0$
- States evolves for both control u=0,1: $p_0^i(j,k) \neq \delta(j,k)$
- Constraint: m active arm each time: $\sum_{i=1}^{N-1} U_n^i = m \quad \forall n$
- Optimal solution: Provably hard : PSPACE Complex class

  The Complexity of Optimal queuing network control, Christos H. Papadimitriou and John N. Tsitsiklis
- Heuristic for relaxed version : Whittle Index Policy

Restless MAB (Relaxed Case)

Objective: $lim_{n\to\infty}\frac{1}{n}\mathbf{E}[\sum_{i=1}^{N}\sum_{m=0}^{n-1}R^i_{U^i_m}(X^i_m)]$

Original Problem

- Constraints: $\sum_{i=0}^{N-1}U^i_m = M \quad \forall n$
- Lagrange: $L(U,\lambda) =$ $lim_{n\to\infty}\mathbf{E}[\sum_{i=1}^{N}\sum_{m=0}^{n-1}(\frac{1}{n}R^i_{U^i_m}(X^i_m) - \lambda_n(U^i_m - M/N))]$
- Remark: Intercoupled constraints and lagrange.

Relaxed Version

- Constraints: $lim_{n\to\infty}\frac{1}{n}\mathbf{E}\sum_{i=0}^{N-1}U^i_m = M \quad \forall i$
- Lagrange: $L(U,\lambda) =$ $lim_{n\to\infty}\frac{1}{n}\mathbf{E}[\sum_{i=1}^{N}\sum_{m=0}^{n-1}(R^i_{U^i_m}(X^i_m) - \lambda_i(U^i_m - M/N))]$
- Lagrange splits for each arm i: $L(U,\lambda) =$ $\sum_{i=1}^{N}lim_{n\to\infty}\frac{1}{n}\mathbf{E}[\sum_{m=0}^{n-1}(R^i_{U^i_m}(X^i_m) - \lambda_i(U^i_m - M/N))]$

Dynamic Programming Equation for Relaxed Restless MAB's Lagrange

- Objective: maximize $L(U, \lambda) = \sum_{i=1}^{N} sup_{U^i} \quad lim_{n\to\infty} \frac{1}{n}\mathbf{E}[\sum_{m=0}^{n-1}(R^i_{U^i_m}(X^i_m) - \lambda_i(U^i_m - M/N))]$
- To match whittle setup, ignore $\lambda * constant$ term: maximize $L(U, \lambda) = \sum_{i=1}^{N} sup_{U^i} \quad lim_{n\to\infty} \frac{1}{n}\mathbf{E}[\sum_{m=0}^{n-1}(R^i_{U^i_m}(X^i_m) + \lambda_i(1 - U^i_m))]$
- For a each arm i, λ_i is fixed, and its in standard RL objective form (Q Learning, reward $R'(s, u) = R_u(s) + \lambda(1 - u)$
- Potential equation (Bellman Equation): $V^i(k) = max_{u\in\{0,1\}}(R^i_u(k) + \sum_j p^i(j|k, u)V^i(j) - \beta^i + (1 - u)\lambda_i)$
- Q value equation : $Q^i(k, u) = R^i_u(k) + \sum_j p^i(j|k, u) max_{v\in\{0,1\}} Q^i(k, v) - \beta^i + (1 - u)\lambda_i)$

Whittle Index

Let R, u, Q, p, β^i have usual meaning. And λ be some constant.
$Q_\lambda(k,u) = R_u(k) + \sum_j p(j|k,u) max_{v \in \{0,1\}} Q_\lambda(k,v) - \beta^i + (1-u)\lambda$

Let $\lambda^*(k)$ denote whittle index for state k for equation Q_0.
$\lambda^*(k) := min\{\lambda \quad | \quad Q_\lambda(k,0) = Q_\lambda(k,1)\}$

Whittle Index base Q Learning for restless bandits with average reward by Avrachenkov and Borkar

- Q update: $Q^i_{n+1}(k,u) = Q^i_n(k,u) + \alpha(i,k,n)[R^i_u(k) + (1-u)\lambda^i_n - f(Q^i) - Q^i_n(k,u) + max_{v\in\{0,1\}} Q^i_n(X_{n+1},v)]$

  $f(Q) = np.mean(Q)$
- Whittle update: $\lambda^i_{n+1} = \lambda^i_n + \gamma(i,n)[Q^i(k,1) - Q^i(k,0)]$
- Control (Policy) : Arrange arms in decreasing order according to λ^i (whittle value) in their current states, pick top m arms for u=1, rest u=0

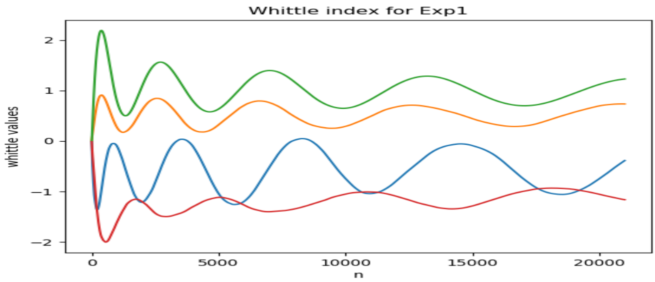
Few Remarks about above algorithm by Avrachenkov and Borkar

- It solves original Restless MAB by getting motivation/intuition from Lagrange of relaxed Restless MAB
- It requires $d = |S|$ set of Q parameters, one for each states. When state space is large, things can get difficult.
- Convergence to Whittle index: $\lambda_n^i \to \lambda^*(i)$

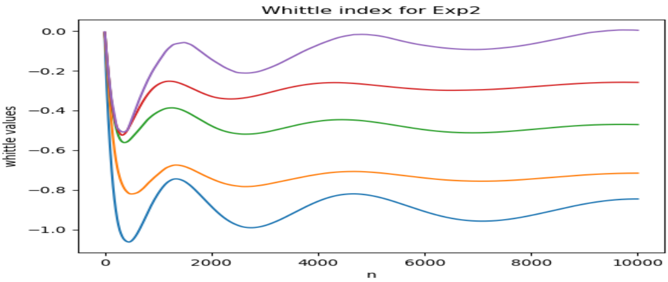
  Proof: Whittle Index base Q Learning for restless bandits with average reward by Avrachenkov and Borkar

Experiment: Whittle Index

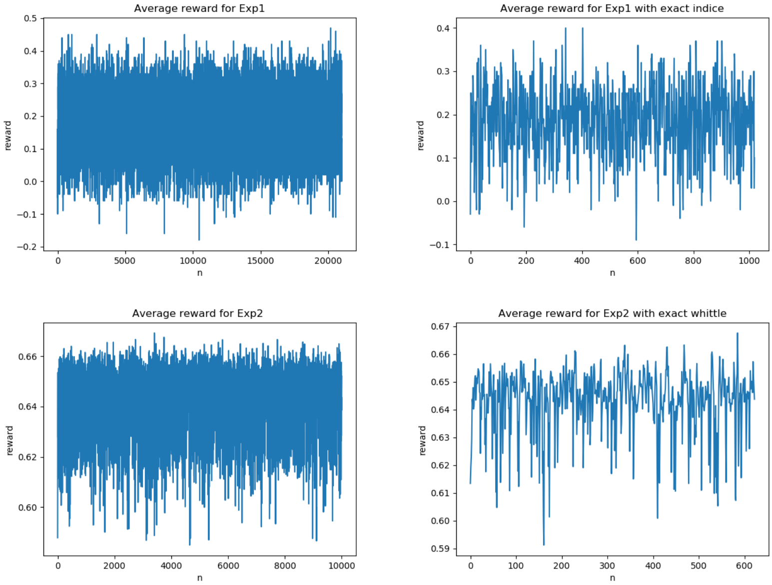
Exact Whittle index = [-0.5, 0.5, 1, -1]



Exact Whittle index = [-0.9, -0.73, -0.5, -0.26, -0.01]



Experiment: Reward



Remark About Above Experiment

- Both examples are taken from the paper by Avrachenkov and Borkar
- Code is available at https://github.com/navdeepkumar12/SA
- Paper's result and my result are in agreement (not in best).
- Learning rate parameters differs. I have put $\gamma = 0.01, \alpha(i, u, n) = 1/v(i, u, n) = 1/\text{no.of times state,}$ control i,u encountered before.
- Damped Oscilatory behaviour of Whittle indices.

Possible Improvements/Generalizations on existing work

- If its a resource allocation problem, why waste resources to get inferior reward. $\sum_i^{N-1} U_m^i \leq M$ instead of $\sum_i^{N-1} U_m^i = M$
- Generalization while keeping analysis similar.
  - Instead of only two controls $U = \{0, 1\}$, have two class of control $U = \{A, B\}, A = \{a_1, a_2, .., a_r\}, B = \{b_1, b_2, ..., b_s\}$.
  - Same objective to maximize.
  - Constraints changes: $\sum_i^{N-1} U_m^i = M$ to $\sum_i^{N-1} \mathbf{1}(U_m^i \in A) = M$
  - Transition probabilities: No freezing, each control $u \in U = \{A, B\}$ may have different transitions probabilities.
- More generalization:
  - Control $U = \{u_1, u_2, .., u_r\}$, weight function: $f : U \to \mathbf{R}$
  - Constraints: $\sum_i^{N-1} f(U_m^i) = M$

Continued ...

- Simplification of algorithm: Present algorithm requires $d = |S|$ sets of Q parameters. May be we can do away with only one Q, derive control using $Q_n(k,0) - Q_n(k,1)$.

Things next in order to do

- I have not yet fully understood the proof of convergence of Q-learning algorithm presented by Avrachenkov and Borkar in their paper.
- Do more experiments, observe carefully to get more intuitions.
- Understand complexity of Restless MAB.

  The Complexity of optimal queuing network control by Papaditmitriou and Tsitsiklis.
- Work on feasibility of ideas on the page above.

Thank You