A Reinforcement Learning Algorithm for Restless Bandits[1]

Workshop on Learning Theory

Vivek Borkar IIT Bombay

January 3, 2019

[1] joint work with Karan Chadha

Model: Time-homogeneous finite-state Markov chain whose states are denoted by $1, \ldots, n$ with stationary distribution π.

X_t is the state at time t.

P is transition probability matrix.

$g(i, j)$ is cost of transitioning from state i to state j

g is the n length column vector with i^{th} entry $\sum_{j=1}^{n} P_{ij} g(i, j)$.

Average cost starting at state i:

$$\lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} E[g(X_k, X_{k+1})|X_0 = i],$$

is constant (η^*). $\eta^* = \pi g$.

Differential cost function:

$$h(i) = \lim_{t\to\infty} \sum_{k=0}^{t} E[g(X_k, X_{k+1}) - \eta^*|X_0 = i]$$

h satisfies the average cost Poisson equation:

$$h = g - \eta^* e + Ph$$

where e is the vector of all 1's.

Function Approximation:

$$h \approx \Phi r$$

Each column of Φ viewed as basis functions.

Each row describes attributes/features of corresponding state.

LSPE Algorithm:

H.Yu and D.P. Bertsekas, "Convergence Results for Some Temporal Difference Methods Based on Least Squares", IEEE Transactions on Automatic Control, 54 (7):1515–1531, 2009.

LSPE:

$(x_1, x_2, \ldots, x_t)$ is a infinitely long sample trajectory. Let

$$\eta_t = \frac{1}{t+1} \sum_{k=0}^{t} g(x_k, x_{k+1}).$$

Temporal differences:

$$d_t(m) = g(x_m, x_{m+1}) - \eta_m + \phi(x_{m+1})^T r_t - \phi(x_m)^T r_t.$$

Define $\tilde{r}_t$ as

$$\tilde{r}_t \quad = \quad \text{argmin}_{r \in \mathbb{R}^M} \sum_{k=0}^{t} (\Phi(x_k)^T r - \ \Phi(x_k)^T r_t - d_t(k))^2.$$

Solution can be found using:

$$r_{n+1} = r_n + c(n)(\tilde{r}_n - r_n),$$

where $c(n)$ satisfies Robbins-Monroe conditions.

Update:

$$\tilde{r}_n = r_n + c(n)\tilde{B}_n^{-1}(\tilde{A}_n r_n + \tilde{b}_n),$$

where

$$\tilde{B}_n = \frac{B_n}{n+1}, \ \tilde{A}_n = \frac{A_n}{n+1}, \ \tilde{b}_n = \frac{b_n}{n+1},$$

with

$$\begin{aligned} B_n &= \sum_{k=0}^{n} \Phi(x_k)\Phi(x_k)^T, \\ A_n &= \sum_{k=0}^{n} \Phi(x_k)(\Phi(x_{k+1})^T - \Phi(x_k)^T), \\ b_n &= \sum_{k=0}^{n} \Phi(x_k)(g(x_k, x_{k+1}) - \eta_k). \end{aligned}$$

Justification:

Introduce mapping T:

$$x \mapsto Tx := g - \eta^* e + Px$$

S: subspace spanned by basis vectors Φ.

LSPE computes r which is a fixed point of:

$$\Phi r = \Pi T \Phi r,$$

where Π is the projection matrix on S defined using the weighted Euclidean norm specified by the invariant distribution

Restless Bandits:

Multi-armed bandits:

N processes (Markov chains), out of which $M < N$ can be operated at a time ('active' arms), while the rest remain *frozen* ('passive' arms).

Problem:

Optimal scheduling

Typical solution: 'index rule' (Gittins): to each process is assigned a state-dependent index, use the bandit with the maximum index. (Optimal)

Restless bandits: Passive bandits drift according to a neutral dynamics.

Problem provably hard. Only a heuristic is available, even after relaxing the rigid condition of 'M out of N' to 'M out of N on average'.

Whittle index:

Consider the 'M out of N on average' version.

For each process i, introduce subsidy λ_i for remaining passive.

If the set of states at which it is optimal to remain passive monotonically increases from 'none' to 'all' as the subsidy increases from $-\infty$ to ∞ for each i, the problem is *Whittle indexable*.

Whittle index of i = the λ_i at which active and passive are equally desirable, as a function of state.

Index policy: Operate top M according to diminishing order of indices

This is suboptimal, but asymptotically optimal as $N \uparrow \infty$ (Weiss-Weber).

Known to perform well in practice.

Some applications:

Sensor scheduling (Nino-Mora and Villar)

Multi-UAV coordination (Ny, Dahleh and Feron)

Congestion control (Avrachenkov et al; Jacko and Sanso)

Cognitive radio (Liu and Zhao)

Real time wireless multicast (V. Raghunathan, VB, Cao, Kumar)

Other applications: (VB + friends)

Scheduling web crawlers

Cloud computing

File sharing networks

Opportunistic scheduling, etc.

Notation:

$\{X_n^i, n \geq 0\}$: $N > 1$ Markov chains

State spaces $S^i, 1 \leq i \leq N$, intervals in $\mathbb{R}$ or $\mathbb{Z}$

Two modes: active and passive.

	Active	Passive
Transition Kernels	$p_a^i(dx'\|x)$	$p_b^i(dx'\|x)$
Reward	$R_a^i(x)$	$R_a^i(x)$

Assume $R_a^i(x) > R_b^i(x), x \in S^i$.

Constraint: $M < N$ active at any time instant.

Problem is to choose which ones are active.

Maximize the time-averaged reward

$$\liminf_{n\uparrow\infty} \frac{1}{n} \sum_{m=0}^{n-1} \sum_{i=1}^{N} E\left[\nu^i(m) R_a^i(X_m^i) + (1-\nu^i(m)) R_b^i(X_m^i)\right]$$

subject to the constraint

$$\sum_{i=1}^{N} \nu^i(n) = M \ , \nu^i(n) \in \{0,1\} \quad \forall n \geq 0.$$

where $\nu^i(n) := I\{i\text{th bandit is active at time } n\}$

Whittle Relaxation:

Per stage constraint $\rightarrow$ averaged constraint.

$$\limsup_{n\uparrow\infty} \frac{1}{n} E\left[\sum_{m=0}^{n-1}\sum_{i=1}^{N} \nu^i(n)\right] \leq M, \ \forall n \geq 0.$$

This makes it a 'constrained Markov decision process'.

Important special feature: separable cost and separable constraint.

Using a Lagrange multiplier (λ), the problem decouples into individual agents.

$$\liminf_{n\uparrow\infty} \frac{1}{n} \sum_{m=0}^{n-1} E\left[\nu^i(m) R_a^i(X_m^i) + (1-\nu^i(m))(\lambda + R_b^i(X_m^i))\right].$$

λ : Subsidy for passivity

- Separability of cost and constraint $\Longrightarrow$ given the Lagrange multiplier λ, the problem decouples
- Each separate average cost control problem involves a binary decision variable:

  to be or not to be (active / passive)

  $\Longrightarrow$ decision boundary comes from an inequality of the type 'a function of state $\leq \lambda$' specifying the passive states. (Threshold policy)

- Whittle indexability $\Longrightarrow$ threshold increases from empty set to the whole space as $\lambda \uparrow \infty$.
- Set Whittle index := the value of λ for active/passive equally desirable.
- Choose the top M according to decreasing value of indices for the current profile of state variables.

SUMMARY:

Bandits are a pest

if they don't get much rest.

So do relax a little

your constraints *a la* Whittle.

Then all that it takes

is a simple index.

RL Scheme to learn Whittle Index:

A reinforcement learning algorithm for restless bandits (VB, Chadha, *Indian Control Conference* 2018)

DP equation is

$$V^i(x)+\beta = \max\Big[R_a^i(x)+\int_{S^i} p_a^i(dx'|x)V^i(x'), \lambda + R_b^i(x) + \int_{S^i} p_b^i(dx'|x)V^i(x')\Big].$$

Whittle index is λ when

$$R_a^i(x)+\int_{S^i} p_a^i(dx'|x)V^i(x') = \lambda^i(x) + R_b^i(x) + \int_{S^i} p_b^i(dx'|x)V^i(x').$$

Assuming fixed threshold policy, DP $\rightarrow$ Poisson Equation

$$\begin{aligned} V(x) &= R_a(x) - \beta + \int p_a(dx'|x)V(x'), \ x \geq \tilde{x}, \\ V(x) &= \lambda + R_b(x) - \beta + \int p_b(dx'|x)V(x'), \ x < \tilde{x}. \end{aligned}$$

Here, λ is the Whittle index for $\tilde{x}$, i.e.

$$\begin{aligned} \lambda(\tilde{x}) &= (R_a(\tilde{x}) + \int p_a(dx'|\tilde{x})V(x')) \\ &\quad - (R_b(\tilde{x}) + \int p_b(dx'|\tilde{x})V(x')). \end{aligned}$$

Whittle index for a fixed $\tilde{x}$ can be computed by minimizing

$$\Big(\lambda - (R_a(x) + \int p_a(dx'|x)\check{V}(\lambda, x')) + (R_b(x) + \int p_b(dx'|x)\check{V}(\lambda, x'))\Big)^2,$$

over λ, where $\check{V}$ is a solution of the Poisson equation above for given λ.

Need to consider for all λ. Write $V = \check{V}(x, \tilde{x})$.

Linear Function Approximation

$$\check{V}(x, \tilde{x}) \approx \Phi^T r = \sum_{i=1}^{m} r_i \phi^i(x, \tilde{x}),$$

where $\Phi^T(\cdot) = [\phi^1(\cdot) : \cdots : \phi^m(\cdot)]$, $r = [r_1, \cdots, r_m]^T$.

$\{\phi^k\}$: 'features' or 'basis functions'.

r_i: 'weights'.

Simulation Based Method

Simulate an i.i.d. process of candidate thresholds $\{\tilde{X}_n\}$ with law κ and a controlled Markov chain $\{X_n\}$ so that the joint transition kernel of the process $(\tilde{X}_n, X_n)$ is

$$I\{x \geq \tilde{x}\}p_a(dx'|x)\kappa(d\tilde{x}) + I\{x < \tilde{x}\}p_b(dx'|x)\kappa(d\tilde{x}).$$

Derive iterative updates using LSPE(0)(Yu & Bertsekas '09).

Define

$$g(x, \tilde{x}, \lambda) := R_a(x) I\{x \geq \tilde{x}\} + (\lambda + R_b(x)) I\{x < \tilde{x}\}.$$

and temporal differences as,

$$\begin{aligned} d_n(m) &= g(X_m) - \beta_m + \Phi(X_{m+1}, \tilde{X}_m)^T r_n - \\ & \quad \Phi(X_m, \tilde{X}_m)^T r_n, \ m \leq n. \end{aligned}$$

Define $\tilde{r}_n$ as

$$\tilde{r}_n = \text{argmin}_{r\in\mathbb{R}^M} \sum_{k=0}^{n} (\Phi(X_k, \tilde{X}_k)^T r - \Phi(X_k, \tilde{X}_k)^T r_n - d_n(k))^2.$$

Applying LSPE,

$$\tilde{r}_n = r_n + c(n)\tilde{B}_n^{-1}(\tilde{A}_n r_n + \tilde{b}_n),$$

where

$$\tilde{B}_n = \frac{B_n}{n+1}, \ \tilde{A}_n = \frac{A_n}{n+1}, \ \tilde{b}_n = \frac{b_n}{n+1}.$$

$$
\begin{aligned}
B_n &= \sum_{k=0}^{n} \Phi(X_k, \tilde{X}_k)\Phi(X_k, \tilde{X}_k)^T, \\
A_n &= \sum_{k=0}^{n} \Phi(X_k, \tilde{X}_k)(\Phi(X_{k+1}, \tilde{X}_k)^T - \Phi(X_k, \tilde{X}_k)^T), \\
b_n &= \sum_{k=0}^{n} \Phi(X_k, \tilde{X}_k)(g(X_k, \tilde{X}_k, \lambda_k) - \beta_n), \\
\beta_n &= \frac{1}{n+1} \sum_{k=0}^{n} g(X_k, \tilde{X}_k, \lambda_k).
\end{aligned}
$$

Function Approximation of the Whittle index:

$$\lambda(x) = \Psi^T y,$$

where $\Psi^T(\cdot) = [\psi_1(\cdot), ..., \psi_K(\cdot)]$, $\psi_i : S \mapsto \mathbb{R}$ represents K features.

Learning scheme to minimize the mean square error,

$$\begin{aligned} y_{n+1} &= y_n - a(n)\Psi(X_n)\Big(\Psi(X_n)^T y_n - \\ &\quad (R_a(X_n) - R_b(X_n)) - (\Phi^T(X'_{n+1}, X_n) r_n \\ &\quad - \Phi^T(X''_{n+1}, X_n) r_n)\Big), \end{aligned}$$

where $X'_{n+1} \approx p_a(\cdot|X_n)$, $X''_{n+1} \approx p_b(\cdot|X_n)$ and $X_{n+1} = X'_{n+1}$ if $X_n \geq \tilde{X}_n$ and $= X''_{n+1}$ otherwise.

Two-timescale update:

$\{a(n)\}, \{c(n)\}$ are positive stepsizes satisfying Robbins-Monroe and

$$\frac{a(n)}{c(n)} \stackrel{n\uparrow\infty}{\rightarrow} 0.$$

Set $\lambda_n := \Psi^T(X_n)y_n$ for use in $r-$update.

View $y_n \approx$ a constant(quasi static) for analysis of $r-$update.

This allows us to analyse r_n as converged in $y-$update.

Intuition:

Introduce $T := C_b(S) \mapsto C_b(S)$ given by

$$x \mapsto Tx := Px + \bar{g} - \beta e$$

The Poisson equation reads $V = TV$, i.e., V is a fixed point of T.

Letting $\pi = \pi(dx, d\tilde{x})$ denote the stationary distribution of the chain $\{(X_n, \tilde{X}_n)\}$ and $\Pi :=$ the projection operator w.r.t. the weighted norm $\|z\|_D := \left(\int \pi(dx, d\tilde{x}) |z_{(x,\tilde{x})}|^2 \right)^{\frac{1}{2}}$ given by

$$(\Pi f)(x, \tilde{x}) := \Phi^T(x, \tilde{x}) \times \left(\int \Phi(\breve{x}', \tilde{x}') \Phi^T(\breve{x}', \tilde{x}') \pi(d\breve{x}', d\tilde{x}') \right)^{-1} \times \int \Phi(\breve{x}'', \tilde{x}'') f(\breve{x}'', \tilde{x}'') \pi(d\breve{x}'', d\tilde{x}'').$$

r-update:

The $r - update$ converges to a solution of $\Phi^T r = \Pi T \Phi^T r$.

Good approximation of V when restricted to Range(Φ^T).

y-update:

Using the two-timescale ideology, y remains quasi-static in $r - update$.

Now, we can analyse $y - update$ using r_n as converged.

Let $\pi_1(dx) := \pi(dx, S)$ and define $F : \mathbb{R}^m \mapsto \mathbb{R}$ by:

$$F(z, r) := \int \pi_1(dx)\Big(\Psi^T(x)z - (R_a(x) - R_b(x) + \int p_a(dx'|x)\Phi^T(x', x)r - \int p_b(dx'|x)\Phi^T(x', x)r\Big)^2,$$

for $z \in \mathbb{R}^m$.

Theorem

If $\{y_n\}$ converge a.s., they converge to the unique minimizer of F.

Crawling for 'ephemeral' (trivia) content

(K. Avrachenkov and V. S. Borkar, "Whittle index policy for crawling ephemeral content," IEEE Transactions on Control of Network Systems, 2016)

Some news items:

- 'Just for animals' terminal at JFK Airport
- Half of Britain does not believe in God
- Messi motivates me to scale greater heights: Ronaldo
- Cyrus Broacha bonds with son over cricket

‘Ephemeral content’:

Web content of immediate interest, but interest rapidly decays with time.

Problem:

How to schedule web crawlers to capture ephemeral web content?

THE MODEL (Empirically validated)

1. $X_i(n) :=$ 'web content' at location i at time n, $1 \leq i \leq N$.
2. $\alpha_i \in (0, 1)$ decay rate of 'interest'
3. $u_i :=$ mean arrival rate of 'content' per epoch

Then the dynamics is:

$X_i(n+1) = \alpha_i X_i(n) + u_i$ if not crawled,

$X_i(n+1) = u_i$ if crawled.

Control variable: $v_i(t) = 1$ if crawled, 0 otherwise.

Objective: Maximize average reward

$$\limsup_{t\uparrow\infty} \sum_{i=1}^{N} \frac{1}{t} \sum_{m=0}^{t} X_i(m) v_i(m)$$

subject to

$$\sum_{i=1}^{N} v_i(m) = M \quad \forall\ m \geq 0.$$

Let

$$\zeta_i(x) := \left\lceil \log^+_{\alpha_i} \left(\frac{u_i - (1-\alpha_i)x}{\alpha_i u_i} \right) \right\rceil.$$

Then the Whittle index is (for non-boundary cases)

$$\gamma_i(x) := (1 + \zeta_i(x)((1-\alpha_i)x - u_i) + \left[\alpha_i^{\zeta_i(x)} + \left(\frac{1-\alpha_i^{\zeta_i(x)}}{1-\alpha_i} \right) \right] u_i.$$

'Boundary cases' (always/never crawl) can be treated separately.

Fully stochastic case

In the fully stochastic case, we replace u^i by i.i.d. random variables $\{U_n^i\}$ with law (say) φ.

In this case, there is no known closed form expression and we apply our algorithm to estimate the Whittle index.

Numerical Results

- Discrete time instants for crawling: $nT,\ n \geq 0$, for some $T > 0$.
- At the end of the interval $[nT, (n+1)T]$, the new utility seen is

$$U_{n+1} := \sum_{\{n:\tau_n \in [nT,(n+1)T]\}} \xi_n^i e^{-\mu_i((n+1)T-\tau_n)},$$

where content at source $i \in \{1, \ldots, N\}$ is published at times τ_n^i with an initial utility ($\approx$ interest level) ξ_n^i. $\{\xi_n^i\}$ are assumed to be i.i.d. exponential.

Function Approximations used

Value function:

$$\Phi = [1 \; x \; \tilde{x} \; x^2 \; \tilde{x}^2 \; x\tilde{x}]$$

where x represents the state and $\tilde{x}$ the threshold.

FA for the Whittle index is taken as perturbation around the deterministic Whittle index.

$$\gamma(x) = \gamma^*(x, y) + \Psi^T(x)y, \; arg = \frac{\pi(x_n - u_n)(1-\alpha)}{\alpha u_n}$$

$$\Psi_i = k\gamma^*(x)[cos(i * arg)].$$

Function Approximations used

- Apply a projection on the perturbation term $\Psi^T y$.
- The update equation is given by:

$$\begin{aligned}\tilde{y}_{n+1} &= y_n - a(n)(\gamma^*(x) + \Psi(X_n)'y - X_n - \\ &\quad V_n(U_{n+1}, X_n) + V_n(\alpha X_n + u_{n+1}, X_n))\Psi(X_n), \\ y_{n+1} &= \Gamma \tilde{y}_{n+1}.\end{aligned}$$

- Here Γ projects the updated y value so that $|\Psi^T y| \leq |p\gamma^*(x)|$.

Numerical Results

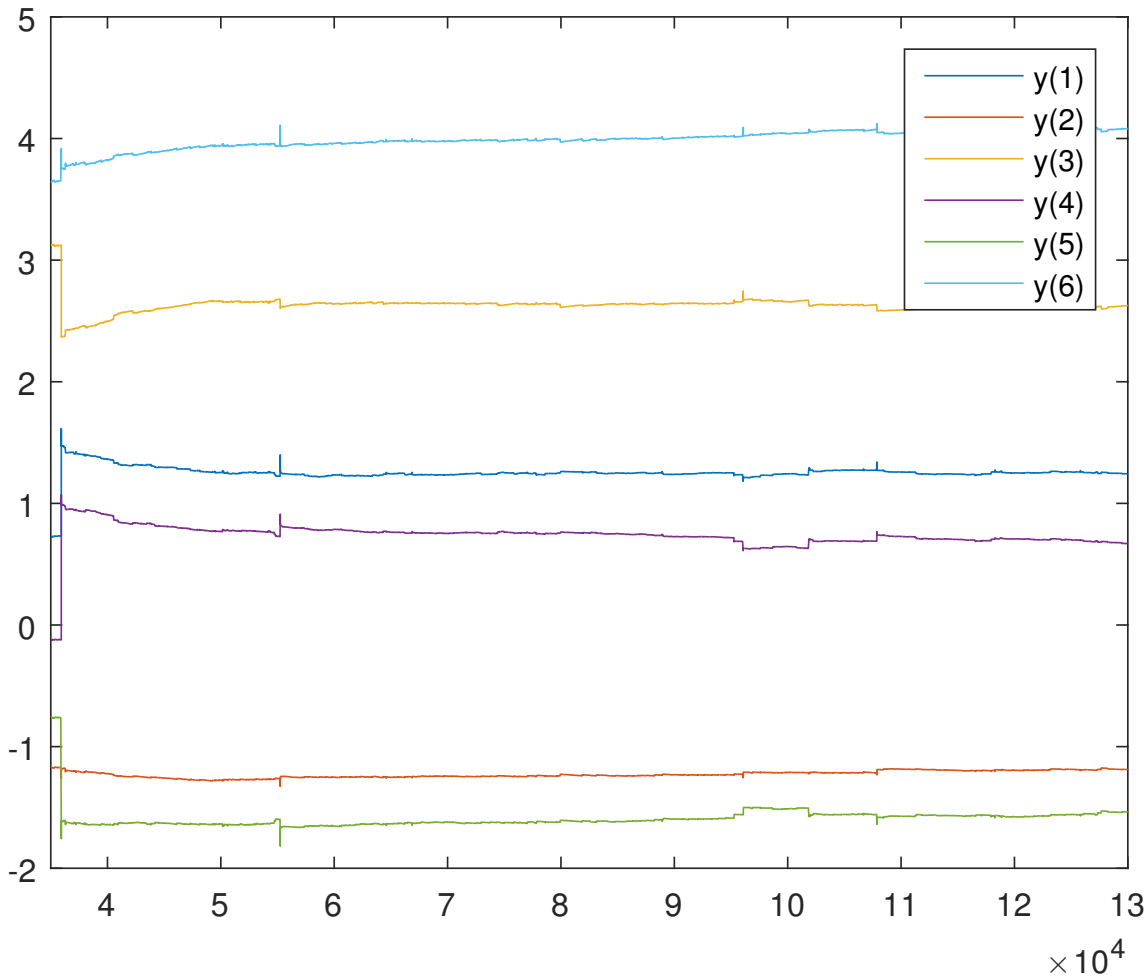

Figure: Convergence of weights of Function Approximation for the Policy with Projection

Numerical Results

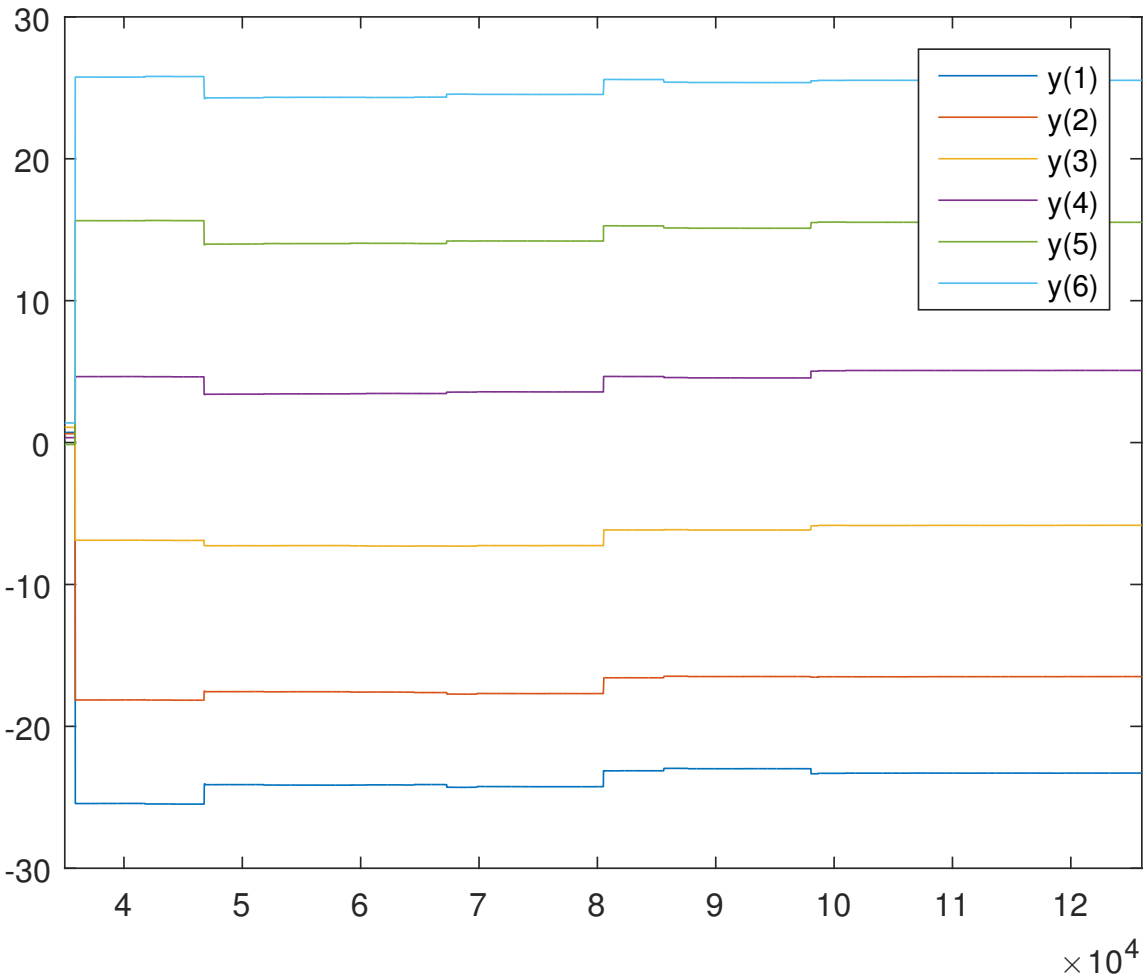

Figure: Convergence of weights of Function Approximation for the Policy without Projection

Numerical Results

Policy	β (M = 1)	β (M = 2)
Round Robin	208.13	281.53
Greedy	232.47	322.72
Deterministic	259.61	328.44
Stochastic without Projection	241.77	307.05
Stochastic with Projection	258.42	333.36

The greedy policy chooses the sources maximising

$$\frac{\Lambda_i \bar{\xi_i}}{\mu_i}(1 - exp(-\mu_i \times t_i))$$

where $\bar{\xi_i} = E\left[\xi_n^i\right]$ and $t_i =$ the time since last crawl for i.

TAKE HOME MESSAGE:

In reinforcement learning for approximate dynamic programming, use structural properties and established heuristics to beat down computation where possible.

For more in this vein, see:

A. Roy, V. S. Borkar, A. Karandikar, P. Chaporkar, "A structure-aware online learning algorithm for Markov decision processes", VALUETOOLS 2019.

With every mistake we must surely be learning,

still my guitar gently weeps.

- George Harrison