Asymptotic Optimality in Restless Bandit

Nicolas Gast

joint work with Bruno Gaujal, Dheeraj Narasimha and Chen Yan

Inria

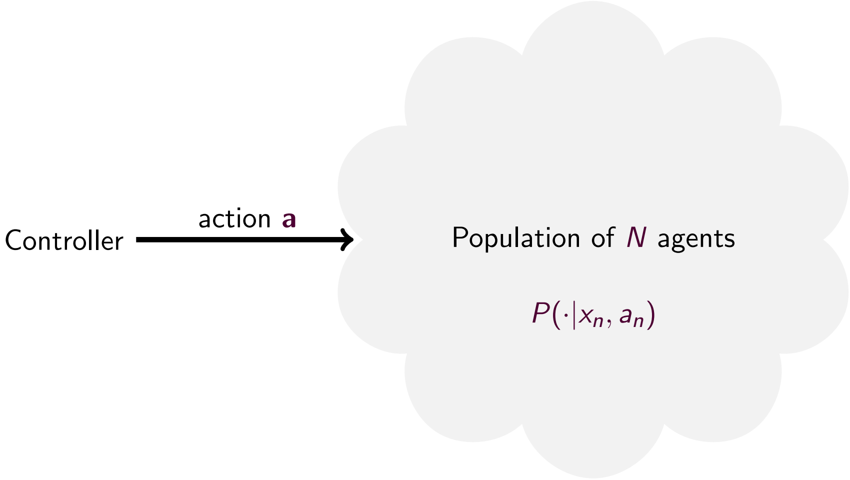
AEP 13, Toulouse, December 2024

Mean field control

Controller $\xrightarrow{\text{action } \mathbf{a}}$ Population of N agents

$P(\cdot|x_n, a_n)$

Mean field control



The computational difficulty increases with N but "$N = \infty$" is easy.

- How to use the $N = +\infty$ solution for finite N?
- How efficient is this? (i.e., how fast does it become optimal?)

This talk will focus on *Markovian bandits*

N statistically identical arms (=agents)

- Discrete time, finite state space.
- $P(\cdot|s_n, a_n)$ and $r(s_n, a_n)$.

Maximize expected reward

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} r(s_n(t), a_n(t)).$$

This talk will focus on *Markovian bandits*

N statistically identical arms (=agents)

- Discrete time, finite state space.
- $P(\cdot|s_n, a_n)$ and $r(s_n, a_n)$.

Maximize expected reward

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} r(s_n(t), a_n(t)).$$

Hard constraint: $\forall t : \sum_{n=1}^{N} a_n(t) \leq \alpha N.$

- If $a_n(t) \in \{0, 1\}$: Markovian bandit (this talk)
- If $a_n(t) \in \{0, 1\}^d$: Weakly coupled MDP.

Example: Resource allocation

Arm/agent can be:

- Tasks (e.g., scheduling)
- Workers (e.g., maintenance problems)
- Electric vehicles (e.g., charging)

Outline

1. The (relaxed) mean-field control problem
2. Three types of policies
   - Index policies
   - FTVA
   - Model predictive control
3. Performance guarantee
4. Conclusion

The mean-field control problem (Whittle's relaxation)

Replace "For all t, $\sum_{n=1}^{N} a_n(t) \leq \alpha N$" by in steady-state: $\sum_{n=1}^{N} \mathbb{E}[a_n] \leq \alpha N$"

$\Rightarrow$ This is a constrained MDP and can be solved by an LP (Altman 99).

The mean-field control problem (Whittle's relaxation)

Replace "For all t, $\sum_{n=1}^{N} a_n(t) \leq \alpha N$" by in steady-state: $\sum_{n=1}^{N} \mathbb{E}[a_n] \leq \alpha N$"
$\Rightarrow$ This is a constrained MDP and can be solved by an LP (Altman 99).

$$
\begin{aligned}
V_{rel} := \max_{x \in \Delta, y \geq 0} & \sum_{s,a} r_{s,a} y_{s,a} & \\
\text{s.t.} \quad & x_{s'} = \sum_{s} y_{s,a} P(s'|s,a) & \text{Markov transitions} \\
& x_s = \sum_{a} y_{s,a} & \text{action taken} \\
& \sum_{s} y_{s,1} = \alpha & \text{relaxed budget contraint}
\end{aligned}
$$

where $x_s = \mathbf{P}[s_n = s]$ and $y_{s,a} = \mathbf{P}[s_n = s, a_n = a]$.

How does a solution look like?

`bandit_lp.BanditRandom(4, seed=1).relaxed_lp_average_reward(alpha=0.4)`

$$y^* = \begin{array}{cc} \text{Action } 0 & \text{Action } 1 \end{array}$$
$$y^* = \begin{bmatrix} & 0.232 \\ 0.028 & 0.168 \\ 0.210 & \\ 0.171 & \\ 0.191 & \end{bmatrix}$$

Note: $0.232 + 0.168 = \alpha = 0.4$.

How does a solution look like?

```
bandit_lp.BanditRandom(4, seed=1).relaxed_lp_average_reward(alpha=0.4)
```

$$
\begin{array}{cc}
 & \begin{array}{cc} \text{Action } 0 & \text{Action } 1 \end{array} \\
y^* = & \begin{bmatrix} & 0.232 \\ 0.028 & 0.168 \\ 0.210 & \\ 0.171 & \\ 0.191 & \end{bmatrix}
\end{array}
\quad \Rightarrow \quad
\pi^* = \begin{bmatrix} 1 \\ 0.857 \\ 0 \\ 0 \\ 0 \end{bmatrix}
$$

Note: $0.232 + 0.168 = \alpha = 0.4$.

Can I apply this to the original (non-relaxed) problem?

π^* is optimal for the constrained MDP $\sum_n \mathbb{E}[A_n] = \alpha N$.

On an example:

$$\text{If } S(t) = [0,0,0,0,0,0,1,1,1,2,2,2,3,3,3,4]$$

$$\Downarrow \text{ Sample } A_n(t) \sim \pi^*(S_n(t)) \text{ (indep.)}$$

$$\tilde{A}_{\pi^*}(t) = [1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0]$$

Problem: here $8 = \sum_{n=1}^{N} \tilde{A}_n(t) \neq \alpha N = 6$.

Historical perspective

and possible solutions

1. Whittle index (88) (Nino-Mora, 90s-2000s) / LP-index (Verloop 15)
   - Works extremely well in practice
   - Often asymptotically optimal (UGAP, Weber and Weiss 91).
   - When they are: exponentially fast. (G, Gaujal, Yan 2023).
2. FTVA – Follow the virtual advice (Hong et al, 2023, 2024)
   - Whittle index can fail (when UGAP fails)
   - Asymptotically optimal in theory, not in practice.
3. Model predictive control (G., Narasimha 2024, G, Gaujal, Yan 2023)
   - Best of both worlds

Outline

1. Index policy: LP-index (and Whittle index)

$$\begin{array}{cc} & \begin{array}{cc} \text{Action } 0 & \text{Action } 1 \end{array} \\ y^* = & \begin{bmatrix} & 0.232 \\ 0.028 & 0.168 \\ 0.210 & \\ 0.171 & \\ 0.191 & \end{bmatrix} \end{array} \xrightarrow{LPindex} I = \begin{bmatrix} 1.216 \\ 0 \\ -0.418 \\ -0.878 \\ -0.237 \end{bmatrix}$$

Index policy: priority to largest index: $0 > 1 > 4 > 2 > 3$.

1. Index policy: LP-index (and Whittle index)

$$
\begin{array}{cc}
 & \begin{array}{cc}\text{Action } 0 & \text{Action } 1\end{array}
\end{array}
$$

$$
y^* = \begin{bmatrix} & 0.232 \\ 0.028 & 0.168 \\ 0.210 & \\ 0.171 & \\ 0.191 & \end{bmatrix} \xrightarrow{LPindex} I = \begin{bmatrix} 1.216 \\ 0 \\ -0.418 \\ -0.878 \\ -0.237 \end{bmatrix}
$$

Index policy: priority to largest index: $0 > 1 > 4 > 2 > 3$.

$$
S(t) = [0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4]
$$

$$
A_{Idx}(t) = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
$$

References: Whittle 88, Verloop 16, Yan et al. 22.

Where does the LP-index comes from?

The $N = \infty$ is a constraint MDP:

- $P(\cdot|s_n, a_n)$ and $r(s_n, a_n)$ s.t. in steady-state, $\mathbf{P}[a_n] = \alpha$.

Where does the LP-index comes from?

The $N = \infty$ is a constraint MDP:

- $P(\cdot|s_n, a_n)$ and $r(s_n, a_n)$ s.t. in steady-state, $\mathbf{P}[a_n] = \alpha$.

Idea: use a Lagrangian relaxation:

- $P(\cdot|s_n, a_n)$ and $r(s_n, a_n) - \underbrace{\lambda}_{\text{Penalty for activation}} a_n$.

Index of state s: $I_s = Q_\lambda(s, 1) - Q_\lambda(s, 0)$.

2. FTVA (Follow the virtual advice, Hong et al. 2023)

	Real	Virtual (uses π^*)
S	0 0 0 0 0 1 1 1 2 2 2 3 3 3 4	[0 0 0 0 0 1 1 1 2 2 2 3 3 3 4]
A		[1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
S		[2 3 4 4 4 3 2 0 0 3 1 2 2 2 0]
A		[0 0 0 0 0 0 0 1 1 0 1 0 0 0 1]
S		[3 2 1 0 1 2 1 1 2 4 0 3 3 3 3]
A		[0 0 1 1 1 0 1 1 0 0 1 0 0 0 0]
S		[3 4 2 4 3 3 0 1 1 0 4 3 4 4 2]
A		[0 0 0 0 0 0 1 1 1 1 0 0 0 0 0]
S		[2 0 3 3 2 1 1 0 4 3 3 2 0 1 4]
A		[0 1 0 0 0 1 1 1 0 0 0 0 1 1 0]
S		[4 4 2 2 1 2 4 4 0 2 4 4 1 0 3]
A		[0 0 0 0 1 0 0 0 1 0 0 0 1 1 0]

FTVA:

Sample $\tilde{A}_n \sim \pi(\tilde{S}_n)$ and $A \leftarrow \mathrm{cap}(\tilde{A})$

If $S_n = \tilde{S}_n$ and $A_n = \tilde{A}$: **then**: couple $S_n(t+1)$ and $\tilde{S}_n(t+1)$

else: wait for them to synchronize.

2. FTVA (Follow the virtual advice, Hong et al. 2023)

	Real	Virtual (uses π^*)
S	0 0 0 0 0 1 1 1 2 2 2 3 3 3 4	[0 0 0 0 0 1 1 1 2 2 2 3 3 3 4]
A	0 0 1 1 1 1 1 1 0 0 0 0 0 0 0	[1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
S		[2 3 4 4 4 3 2 0 0 3 1 2 2 2 0]
A		[0 0 0 0 0 0 0 1 1 0 1 0 0 0 1]
S		[3 2 1 0 1 2 1 1 2 4 0 3 3 3 3]
A		[0 0 1 1 1 0 1 1 0 0 1 0 0 0 0]
S		[3 4 2 4 3 3 0 1 1 0 4 3 4 4 2]
A		[0 0 0 0 0 0 1 1 1 1 0 0 0 0 0]
S		[2 0 3 3 2 1 1 0 4 3 3 2 0 1 4]
A		[0 1 0 0 0 1 1 1 0 0 0 0 1 1 0]
S		[4 4 2 2 1 2 4 4 0 2 4 4 1 0 3]
A		[0 0 0 0 1 0 0 0 1 0 0 0 1 1 0]

FTVA:

Sample $\tilde{A}_n \sim \pi(\tilde{S}_n)$ and $A \leftarrow \mathrm{cap}(\tilde{A})$

If $S_n = \tilde{S}_n$ and $A_n = \tilde{A}$: **then**: couple $S_n(t+1)$ and $\tilde{S}_n(t+1)$

else: wait for them to synchronize.

2. FTVA (Follow the virtual advice, Hong et al. 2023)

	Real	Virtual (uses π^*)
S	0 0 0 0 0 1 1 1 2 2 2 3 3 3 4	[0 0 0 0 0 1 1 1 2 2 2 3 3 3 4]
A	0 0 1 1 1 1 1 1 0 0 0 0 0 0 0	[1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
S	3 0 4 4 4 3 2 0 0 3 1 2 2 2 0	[2 3 4 4 4 3 2 0 0 3 1 2 2 2 0]
A		[0 0 0 0 0 0 0 1 1 0 1 0 0 0 1]
S		[3 2 1 0 1 2 1 1 2 4 0 3 3 3 3]
A		[0 0 1 1 1 0 1 1 0 0 1 0 0 0 0]
S		[3 4 2 4 3 3 0 1 1 0 4 3 4 4 2]
A		[0 0 0 0 0 0 1 1 1 1 0 0 0 0 0]
S		[2 0 3 3 2 1 1 0 4 3 3 2 0 1 4]
A		[0 1 0 0 0 1 1 1 0 0 0 0 1 1 0]
S		[4 4 2 2 1 2 4 4 0 2 4 4 1 0 3]
A		[0 0 0 0 1 0 0 0 1 0 0 0 1 1 0]

FTVA:

Sample $\tilde{A}_n \sim \pi(\tilde{S}_n)$ and $A \leftarrow \mathrm{cap}(\tilde{A})$

If $S_n = \tilde{S}_n$ and $A_n = \tilde{A}$: **then**: couple $S_n(t+1)$ and $\tilde{S}_n(t+1)$

else: wait for them to synchronize.

2. FTVA (Follow the virtual advice, Hong et al. 2023)

	Real	Virtual (uses π^*)
S	0 0 0 0 0 1 1 1 2 2 2 3 3 3 4	[0 0 0 0 0 1 1 1 2 2 2 3 3 3 4]
A	0 0 1 1 1 1 1 1 0 0 0 0 0 0 0	[1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
S	3 0 4 4 4 3 2 0 0 3 1 2 2 2 0	[2 3 4 4 4 3 2 0 0 3 1 2 2 2 0]
A	1 1 0 0 0 0 0 1 1 0 1 0 0 0 1	[0 0 0 0 0 0 0 1 1 0 1 0 0 0 1]
S	4 4 1 0 1 2 1 1 2 4 0 3 3 3 3	[3 2 1 0 1 2 1 1 2 4 0 3 3 3 3]
A	0 0 1 1 1 0 1 1 0 0 1 0 0 0 0	[0 0 1 1 1 0 1 1 0 0 1 0 0 0 0]
S	3 0 2 4 3 3 0 1 1 0 4 3 4 4 2	[3 4 2 4 3 3 0 1 1 0 4 3 4 4 2]
A	1 1 0 0 0 0 1 1 1 1 0 0 0 0 0	[0 0 0 0 0 0 1 1 1 1 0 0 0 0 0]
S	4 3 3 3 2 1 1 0 4 3 3 2 0 1 4	[2 0 3 3 2 1 1 0 4 3 3 2 0 1 4]
A	0 1 0 0 0 1 1 1 0 0 0 0 1 1 0	[0 1 0 0 0 1 1 1 0 0 0 0 1 1 0]
S	0 4 2 2 1 2 4 4 0 2 4 4 1 0 3	[4 4 2 2 1 2 4 4 0 2 4 4 1 0 3]
A	1 1 0 0 1 0 0 0 1 0 0 0 1 1 0	[0 0 0 0 1 0 0 0 1 0 0 0 1 1 0]

FTVA:

Sample $\tilde{A}_n \sim \pi(\tilde{S}_n)$ and $A \leftarrow \mathrm{cap}(\tilde{A})$

If $S_n = \tilde{S}_n$ and $A_n = \tilde{A}$: **then**: couple $S_n(t+1)$ and $\tilde{S}_n(t+1)$

else: wait for them to synchronize.

3. Model predictive control (aka "LP-update")

We define a finite-horizon deterministic problem:

$$V_\tau(\mathbf{S}) := \max_{y \geq 0} \sum_{t=0}^{\tau} \sum_{s,a} r_{s,a} y_{s,a}(t)$$

$$\text{s.t.} \quad \sum_a y_{s,a}(t+1) = \sum_s y_{s,a}(t) P(s'|s,a) \qquad \text{Markov transitions}$$

$$\sum_s y_{s,1}(t) = \alpha \qquad \text{relaxed budget contraint}$$

$$\sum_a y_{s,a}(0) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{\{S_n(t)=s\}} \qquad \text{initial state}$$

We then apply $y_{s,a}(0)$ to all states.

- Finite-horizon or rolling horizon.

Outline

Assumptions

We consider the following deterministic dynamical system:

$$\phi(\mathbf{x}) = \mathbb{E}\left[\mathbf{X}(t+1) \mid \mathbf{X}(t) = \mathbf{x} \wedge \ A \ \sim \text{index}\right],$$

and we call y^* the solution of V_{rel}, with $x^*_s = \sum_a y^*_{sd,a}$.

We define the following conditions:

- **UGAP** $\lim_{t\to\infty} x_{t+1} = \phi(x_t)$ converges to x^* uniformly for all x.
- **Local stability** ϕ is locally stable around x^*.
- **Degenerate** $y_{s,1} = 0$ or $y_{s,0} = 0$ for all s.

Theoretical guarantees

Theorem (Weber-Weiss, G,G,Y23)

Under UGAP and non-degenerate: $V_{index} \geq V_{rel} - e^{-\Omega(N)}$.

Theorem (Hong et al. 23)

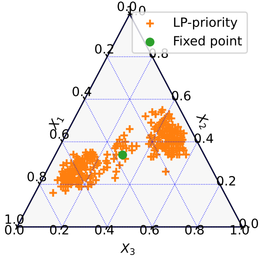
If P *is ergodic, then:* $V_{FTVA} \geq V_{rel} - O(1/\sqrt{N})$.

Theorem (G,N 24)

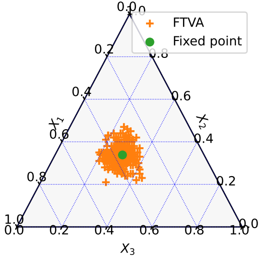
1. *If* P *is ergodic:* $V_{MPC} \geq V_{rel} - O(1/\sqrt{N})$.
2. *Under non-degenerate and local stability:* $V_{MPC} \geq V_{rel} - e^{-\Omega(N)}$.

UGAP is not alwqys satisfied

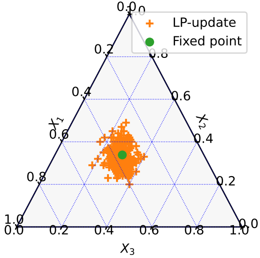
Example from Yan 2023 (3D example)



(a) Index



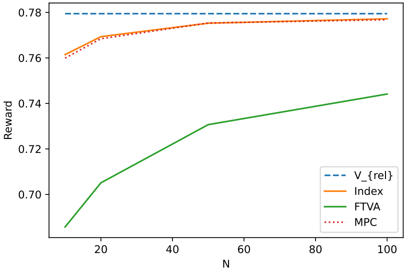
(b) FTVA



(c) MPC

Illustration

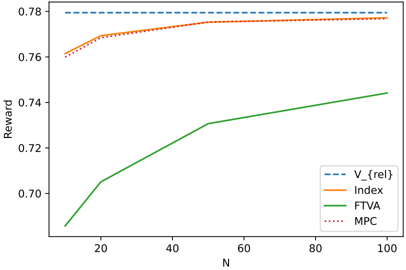
The random example.



UGAP + non-degenerate.

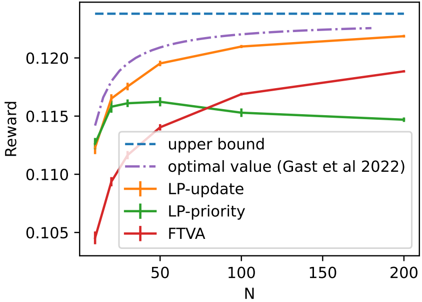
Illustration

The random example.



UGAP + non-degenerate.

Example from Yan 2023.



No UGAP nor local stability.

Outline

Conclusion

For Markovian bandits, mean-field control can be solved by an LP.

- Can be generalized to weakly coupled MDPs.

Simple policies (priority rule) are not always optimal.

- When they are, they become optimal exponentially fast.

- This talk: comparison of various approaches.

Conclusion

For Markovian bandits, mean-field control can be solved by an LP.

- Can be generalized to weakly coupled MDPs.

Simple policies (priority rule) are not always optimal.

- When they are, they become optimal exponentially fast.

- This talk: comparison of various approaches.

- Open questions: learning, continuous state-spaces.

http://polaris.imag.fr/nicolas.gast/

- *LP-based policies for restless bandits: necessary and sufficient conditions for (exponentially fast) asymptotic optimality*. G. Gaujal Yan. MMOR 2023. https://arxiv.org/abs/2106.10067
- *Restless Bandits with Average Reward: Breaking the Uniform Global Attractor Assumption*. Hong, Xie, Chen, and Wang. NeurIPS 2023.
- *Model Predictive Control is Almost Optimal for Restless Bandit*. G, Narasimha. 2024. Under review.