Achieving exponential asymptotic optimality in average-reward restless bandits without global attractor assumptions

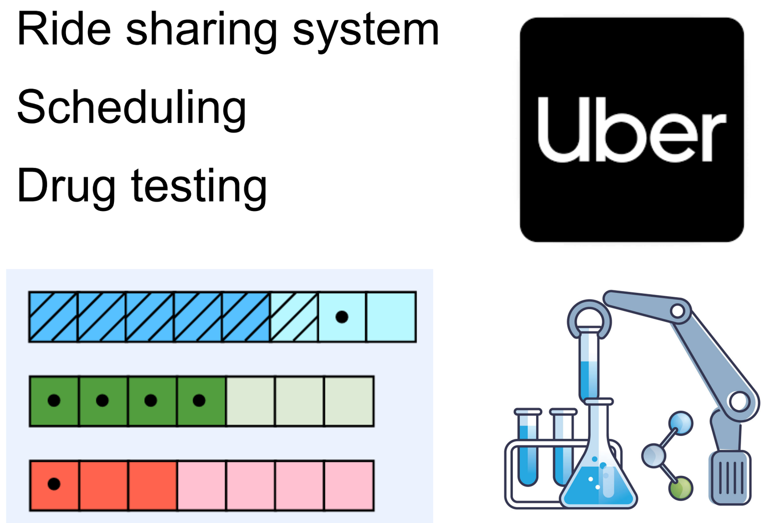
Yige Hong, Qiaomin Xie, Yudong Chen, Weina Wang

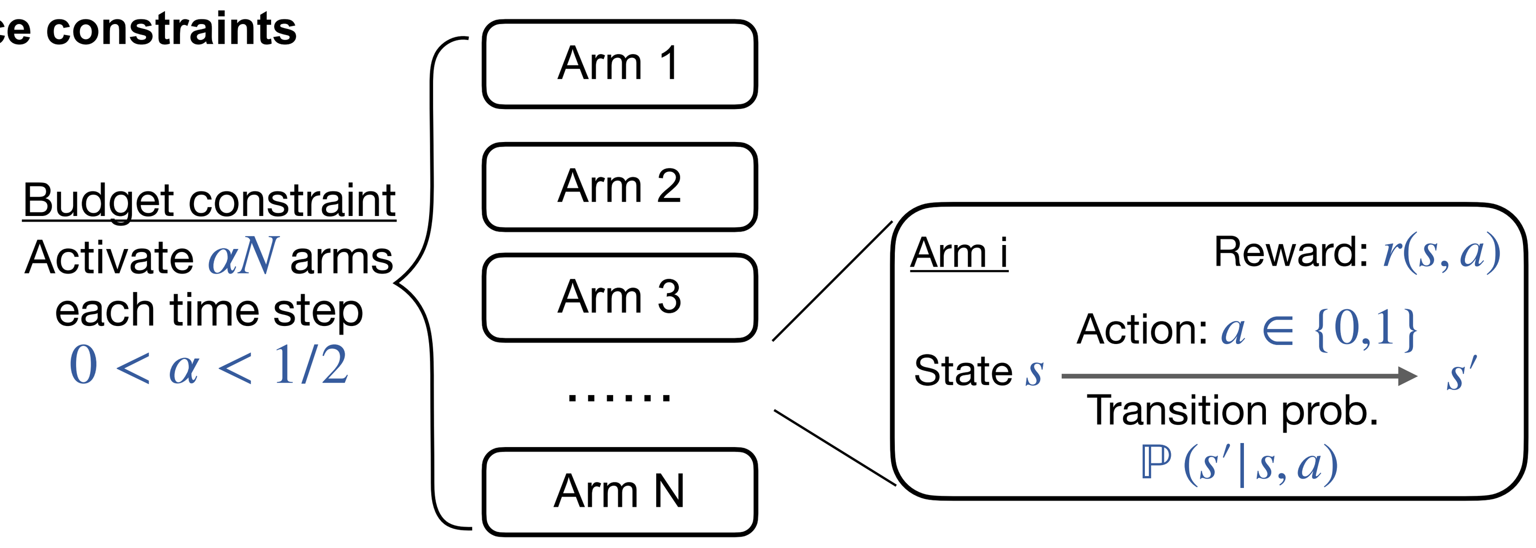
Carnegie Mellon University
WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

1 Restless bandits (RBs)

Multi-component control under resource constraints

- Ride sharing system
- Scheduling
- Drug testing





$\text{maximize}_\pi\ R_N^\pi \triangleq$ long-run avg reward per arm under policy π

- Focus on planning, i.e., model is known

Q: It's just a big MDP. Can we directly solve it?

A: N-dimensional state space; hard if N large

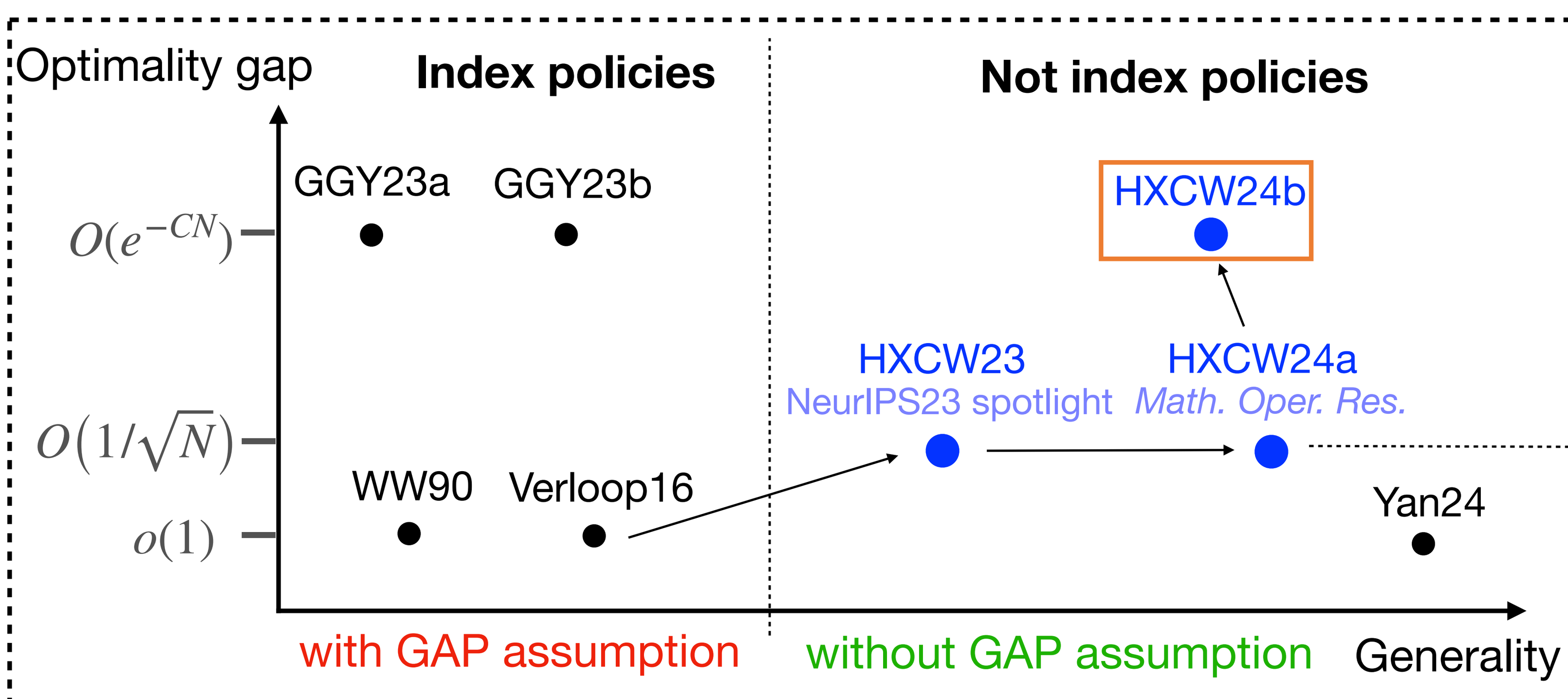
Q: Can we efficiently find a good policy?

Q: How to define a good policy?

A: Asymptotically optimal policy:

$\lim_{N\to\infty}\left(R_N^* - R_N^\pi\right) = 0$

2 Landscape of the RBs literature



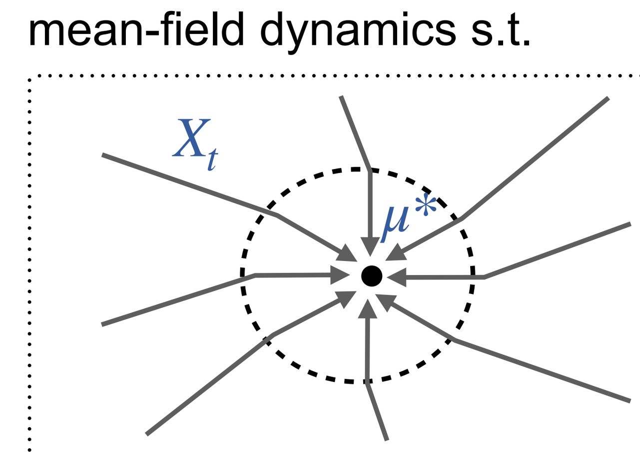
3 From global attractor to local stability

For each s,

$X_t(s)$ = fraction of arms in state s, $\mu^*(s)$ = ideal distr. (see below)

Global attractor assumption

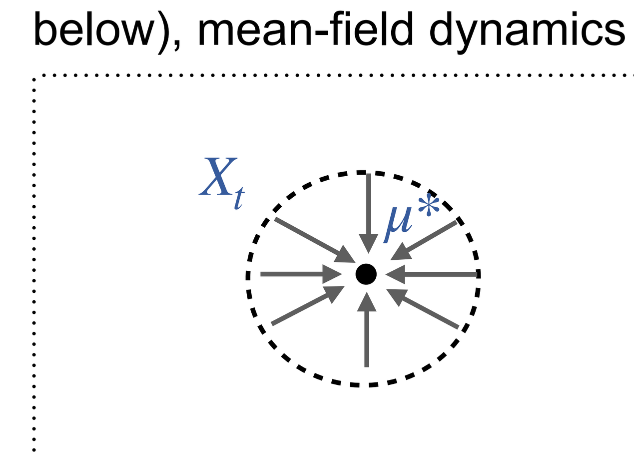
Under the given policy, mean-field dynamics s.t.



Non-linear, policy-dependent

Local stability assumption

Under LP priority policy (see below), mean-field dynamics s.t.



Linear, intrinsic to problem

4 Main theorems

Theorem 1 (achievability)

Assuming unichain, non-degeneracy, and local stability, we can efficiently obtain a policy π (Two-Set Policy below) such that

$R^{\text{rel}} - R_N^\pi = O\left(e^{-CN}\right)$.

Theorem 2 (converse)

Without any of unichain, non-degeneracy, or local stability, for any policy π,

$R^{\text{rel}} - R_N^\pi = \Omega\left(1/\sqrt{N}\right)$.

Comment:

- $R^{\text{rel}} \geq R_N^*$ is fluid upper bound (see below)
- Theorem 2: without a better upper bound, the three conditions are necessary for exp opt gap

5 Algorithm

LP relaxation

$y(s,a)$ = steady-state probability of (state, action) = (s,a)

$$\max_y \sum_{s,a} r(s,a)\, y(s,a)$$

$$\text{s.t.} \quad \sum_s y(s,1) = \alpha$$

Relaxed budget: satisfy in **expectation**

$$\sum_{s',a} \mathbb{P}(s \mid s', a)\, y(s',a) = \sum_a y(s,a) \quad \forall s \in \mathbb{S}$$

$$\sum_{s',a'} y(s',a') = 1; \quad y(s,a) \geq 0 \quad \forall s \in \mathbb{S}, a \in \mathbb{A}$$

$\Longrightarrow$ $y^*(s,a)$: ideal state-action frequency

R^{rel}: fluid upper bound

$\mu^*(s) \triangleq \sum_a y^*(s,a)$: ideal state distribution

$\bar{\pi}^*(a \mid s) \triangleq y^*(s,a) \big/ \mu^*(s)$: opt single-armed policy

Both optimal but infeasible

Subroutine 1: control distribution

Assume relaxed budget constraint,

Consistently following $\bar{\pi}^*$:

Each arm in state s, activate with prob. $\bar{\pi}^*(1 \mid s)$

$\Longrightarrow$ Drive X_t close to μ^* (By Markov chain mixing)

Subroutine 2: exploit reward

Assume X_t is sufficiently close to μ^*,

Consistently following "LP Priority policy":

i. Activate all arms in state s such that $\bar{\pi}^*(1 \mid s) = 1$

ii. Activate no arms in state s such that $\bar{\pi}^*(1 \mid s) = 0$

iii. Spend remaining budget on $0 < \bar{\pi}^*(1 \mid s) < 1$

$\Longrightarrow$ Expected reward = R^{rel}, use budget = αN

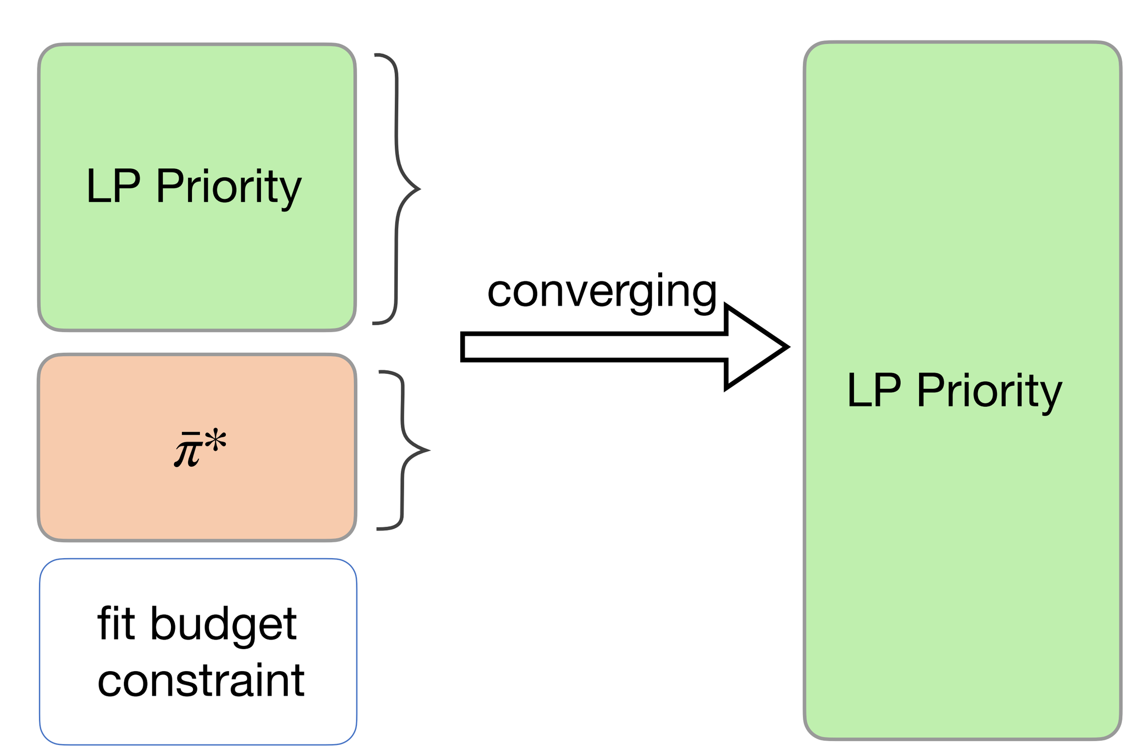
(Intuitively, activate all "valuable states"; "neutral states" are flexible)

Put together: Two-Set Policy

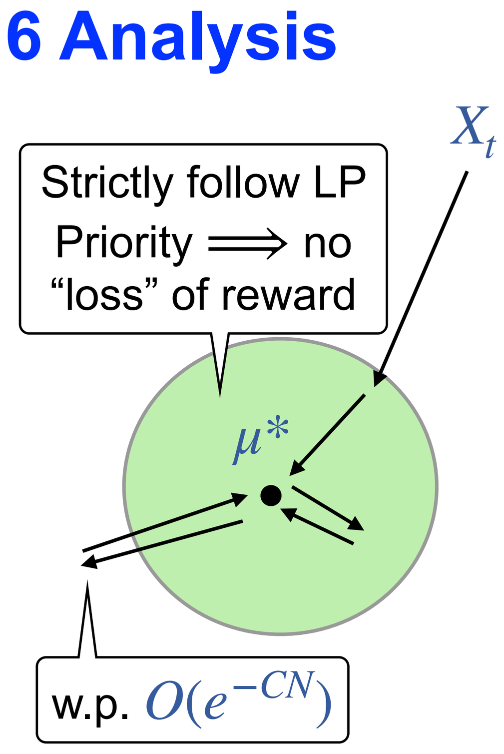
Partition arms into three sets:

- A maximal set of arms follows LP Priority, strictly expand or shrink in the next time step
- A subset follow $\bar{\pi}^*$, subject to budget
- The rest arms fit the constraint

Arms in [orange] merge into [green]



6 Analysis



$R^{\text{rel}} - R_N^\pi = R^{\text{rel}} - \mathbb{E}[r^\pi(X_t)]$, where $r^\pi(x)$ is instantaneous expected reward under π

There exists $V(x)$ that satisfies $R^{\text{rel}} - r^\pi(X_t) \leq \mathbb{E}[V(X_{t+1}) \mid X_t] - V(X_t) + O(e^{-CN})$

Specifically, $V(x) = \left(V_1(x) - 1/2\right)^+ + V_2(x)$

- $V_1(x)$: Lyapunov function for proving global attraction; it's "a multivariate Lyapunov function"
- $V_2(x)$: solution to the Poisson equation $V_2(x\Phi) - V_2(x) = -R^{\text{rel}} + r(x)$;
- Φ: mean-field transition map in the local region

We can show $\left(R^{\text{rel}} - r(X_t)\right)\mathbb{1}\{\text{inside}\} = \left(\mathbb{E}[V_2(X_{t+1}) \mid X_t] - V_2(X_t)\right)\mathbb{1}\{\text{inside}\}$

$$\left(R^{\text{rel}} - r(X_t)\right)\mathbb{1}\{\text{outside}\} \leq \mathbb{E}[(V_1(X_{t+1}) - 1/2)^+ \mid X_t] - (V_1(X_t) - 1/2)^+ + O(e^{-CN}) + \left(\mathbb{E}[V_2(X_{t+1}) \mid X_t] - V_2(X_t)\right)\mathbb{1}\{\text{outside}\}$$

$V_1(x) = \|X_t(D^{\text{LP}}) - m(D^{\text{LP}})\mu^*\|_U + \|X_t(D^{\bar{\pi}^*}) - m(D^{\bar{\pi}^*})\mu^*\|_W + L\left(1 - m(D^{\text{LP}}) - m(D^{\bar{\pi}^*})\right)$

where D^{LP} is the set of arms [green], $D^{\bar{\pi}^*}$ is the set of arms [orange], $m(D) = |D|/N$, U and W are weight matrices, L is a large const

7 Simulation