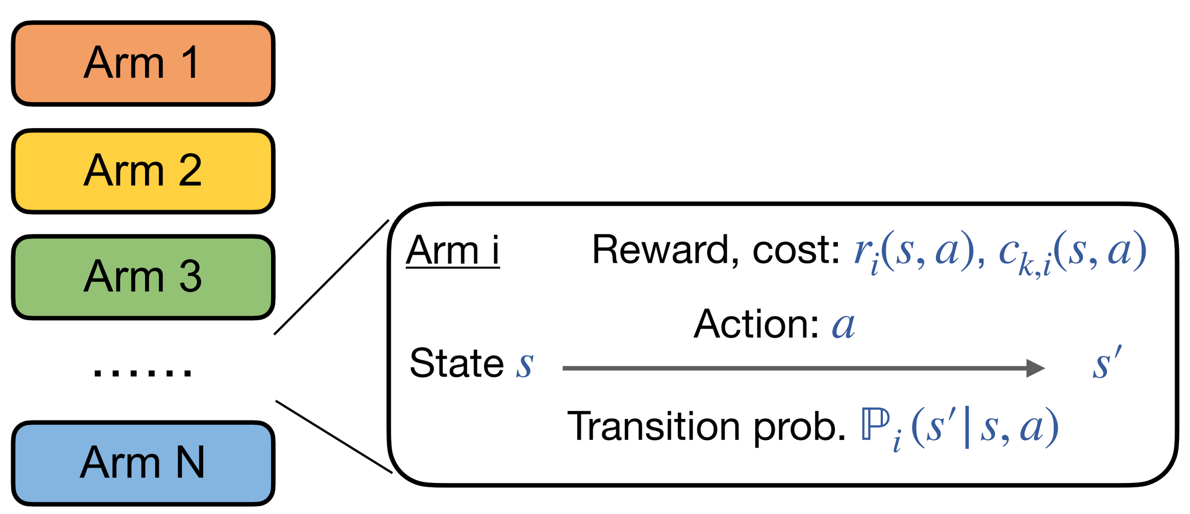
Projection-based Lyapunov method for fully heterogeneous weakly-coupled MDPs

Xiangcheng Zhang*, **Yige Hong***, Weina Wang

1 Weakly-Coupled Markov Decision Processes (WCMDPs)



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

s.t. total type-k cost $\le \alpha_k N$, **each time step**, for $k \in [K]$

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

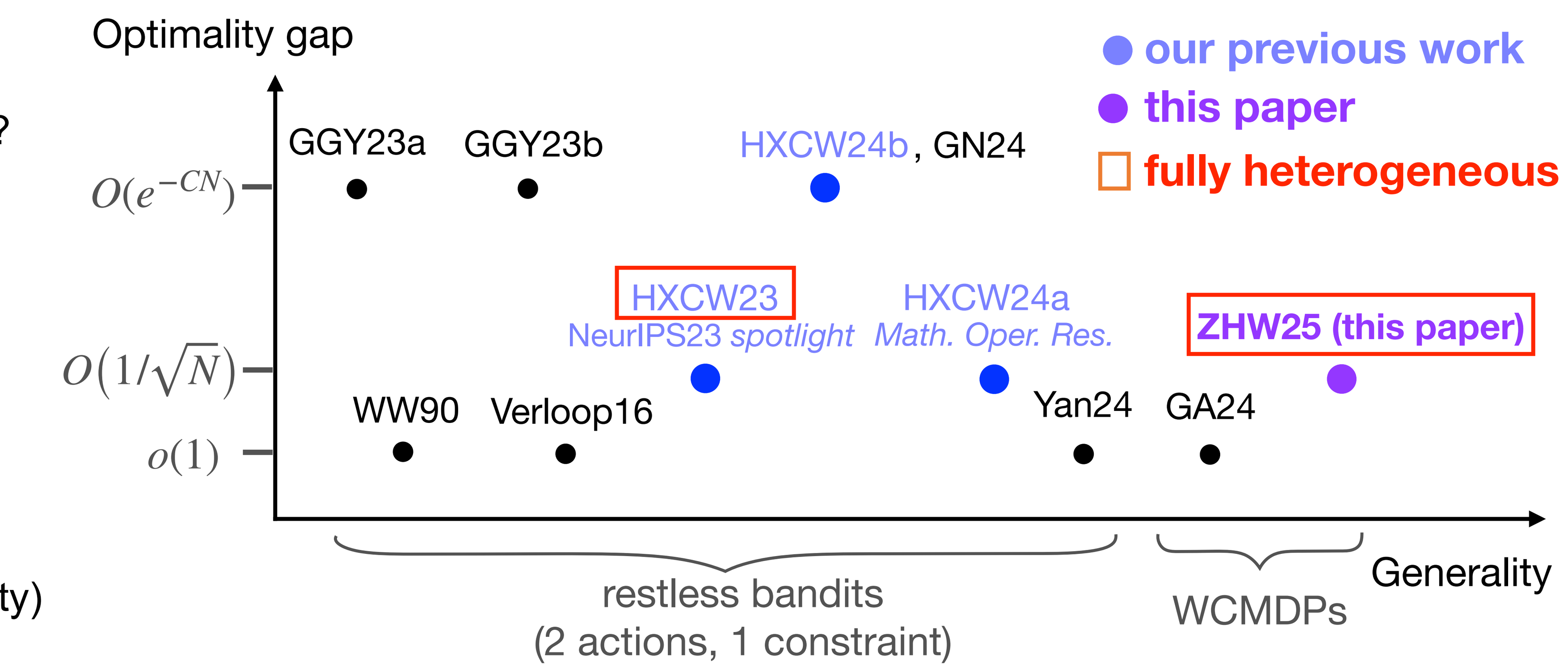
Q: Can we efficiently find a good policy?

Q: How to define a good policy?

A: We want to efficiently find a policy s.t.

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

2 Prior work and this paper



3 Challenge: heterogeneity

Different system state representations:

Homogeneous: for each s

$X_t(s)$ = fraction of arms in state s

Heterogeneous: for each s and $i \in [N]$

$X_{i,t}(s) = 1\{$state of arm $i = s\}$

State aggregation fails

4 Algorithm

4.1 Linear programming relaxation

$y_i(s,a)$ = steady-state probability of arm i 's (state, action) = (s,a)

$$\max_y \sum_{i,s,a} r_i(s,a)\, y_i(s,a)$$

$$\text{s.t.} \sum_{i,s,a} c_{k,i}(s,a)\, y_i(s,a) \le \alpha_k N \quad \forall k \in [K]$$

$$\sum_{s',a} \mathbb{P}_i(s \mid s',a)\, y_i(s',a) = \sum_a y_i(s,a) \quad \forall s \in \mathbb{S}, i \in [N]$$

$$\sum_{s',a'} y_i(s',a') = 1; \quad y_i(s,a) \ge 0 \quad \forall s \in \mathbb{S}, a \in \mathbb{A}, i \in [N]$$

Relaxed constraints on **expected** cost

$\Longrightarrow$ $y_i^*(s,a)$ = ideal state-action frequency

4.2 What LP relaxation gives us…

Each arm:

i. Ideal state distribution: $\mu_i^*(s) = \sum_a y_i^*(s,a)$

ii. Ideal policy:: $\bar\pi_i^*(a \mid s) = y_i^*(s,a) / \mu_i^*(s)$

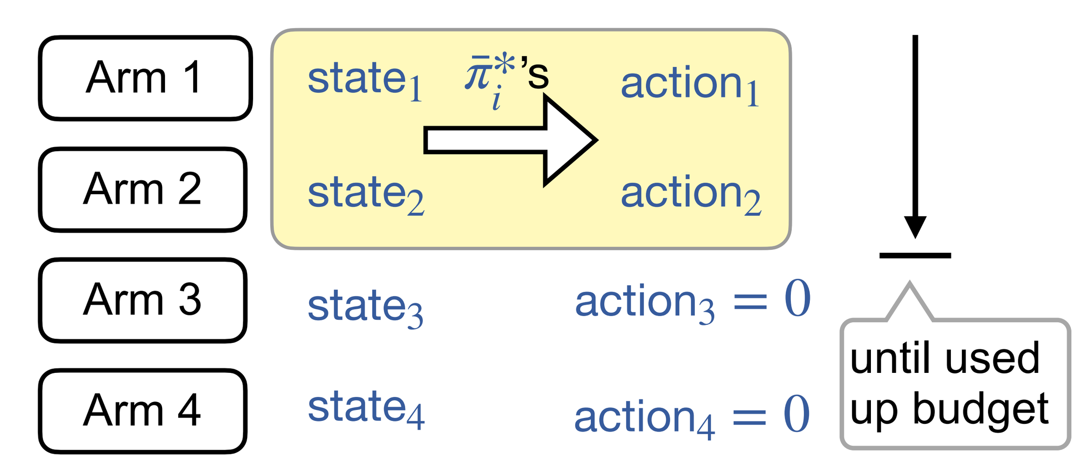
Policy is good if: as $t \to \infty$

i. State distribution approximates $\mu_i^*(s)$

ii. Action distribution approximates $\bar\pi_i^*(a \mid s)$

If arm i takes actions using $\bar\pi_i^*(a \mid s)$, its stationary state distribution is $\mu_i^*(s)$

4.3 ID policy



5 Main theorem

Assumption 1: For each arm i, let τ_i be its mixing time under the optimal single-armed policy $\bar\pi_i^*$. Assume that there exists a constant τ such that $\tau_i \le \tau$ for $i = 1,2,\ldots$.

Theorem 1: Let π be the ID policy. Under Assumption 1, there exists a constant C_{ID} s.t. $R_N^* - R_N^\pi \le C_{\mathrm{ID}} / \sqrt{N}$.

Remark: $C_{\mathrm{ID}} = O\left(K^5 \max\{r_{\max}, c_{\max}\}^7 \tau^4 / \alpha_{\min}^6\right)$

6 Analysis: a projection-based Lyapunov method

6.1 Lyapunov analysis overview

Define $V(\mathbf{X}_t)$ as "distance" between

$\mathbf{X}_t = (X_{i,t}(s))_{i,s}$ and $(\mu_i^*(s))_{i,s}$ s.t.

(C1) Drift condition: for some $\rho < 1$

$\mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] \le \rho V(\mathbf{X}_t) + O(1/\sqrt{N})$

(C2) Dominance condition:

$R_N^* - R_N^\pi \le \mathbb{E}[V(\mathbf{X}_t)] + O(1/\sqrt{N})$

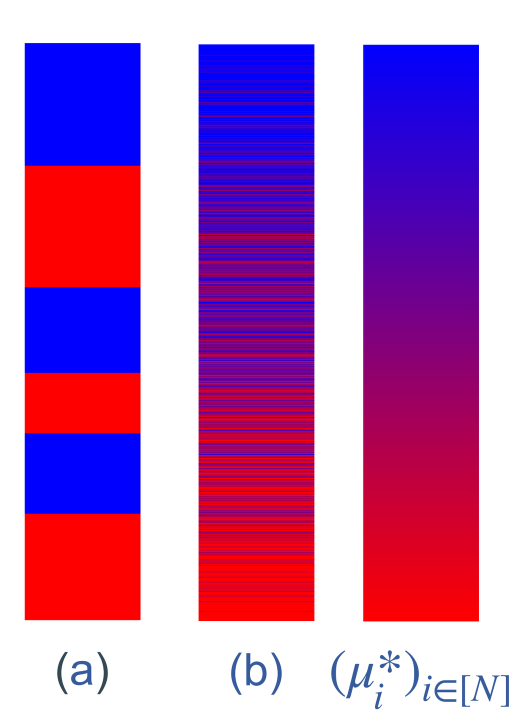
$\Longrightarrow$

$$R_N^* - R_N^\pi \le \mathbb{E}[V(\mathbf{X}_t)] + O(1/\sqrt{N}) \le O(1/\sqrt{N})$$

6.2 Defining "distance" to $(\mu_i^*)_{i\in[N]}$ assuming all arms independently follow $\bar\pi_i^*$

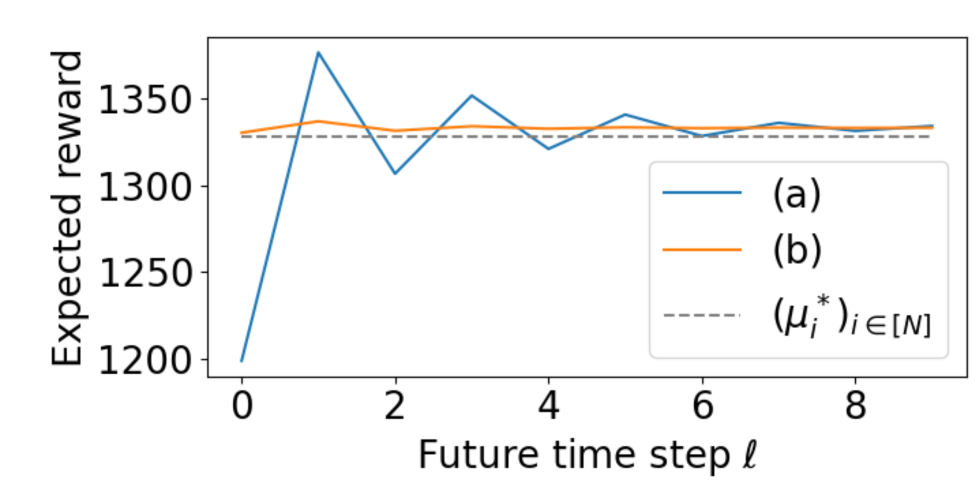
Example:

- Two states, blue and red
- Arm i = i-th row of pixels
- (a) and (b) represent two realizations of $\mathbf{X}_t$



- (a) and (b) have same fractions of arms in blue and red
- Which of (a) and (b) is closer to $(\mu_i^*)_{i\in[N]}$?

Future expected reward / cost as features:



- Lyapunov function based on feature projections:

$$h(\mathbf{X}_t) = \frac{1}{N} \max_{g\in\mathcal{G}} \sup_{\ell\in\mathbb{N}} e^{\ell/(2\tau)} \left| \sum_{i\in[N]} \left\langle (X_{i,t} - \mu_i^*) P_i^\ell,\, g_i \right\rangle \right|$$

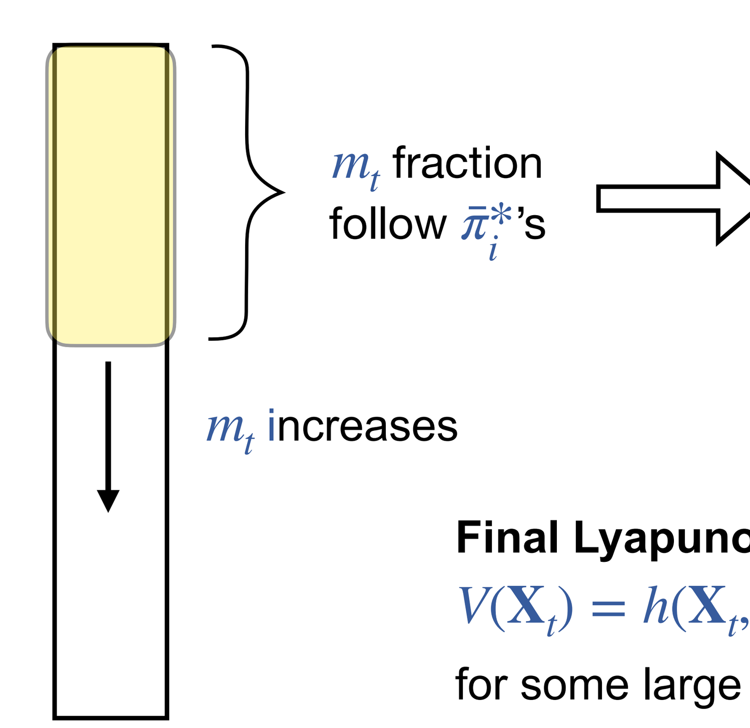
For each arm i, $g_i \in \mathbb{R}^{\mathbb{S}}$ is expected reward / cost function, P_i is transition matrix, $\langle X_{i,t} P_i^\ell, g_i\rangle$ is expected reward / cost ℓ steps later

- **(C1) (C2)** satisfied by $h(X_t)$ if all arms independently follow $\bar\pi_i^*$'s

6.3 One more hurdle: not all arms follow $\bar\pi_i^*$

Observation: a small enough subset of arms can follow $\bar\pi_i^*$

$h(\mathbf{X}_t, m) = m\, h(\mathbf{X}_t)$ evaluated on first m fraction of arms



Final Lyapunov function:

$V(\mathbf{X}_t) = h(\mathbf{X}_t, m_t) + L(1 - m_t)$

for some large constant L