`Learning to Control' an Unknown System

Rahul Jain

University of Southern California

Joint work with ***Mukul Gagrani (USC)****,* ***Ashutosh Nayyar (USC)*** *and* ***Yi Ouyang (UC Berkeley)***

Simons Institute RTDM Program - Societal Networks Workshop
March 26, 2018

Outline

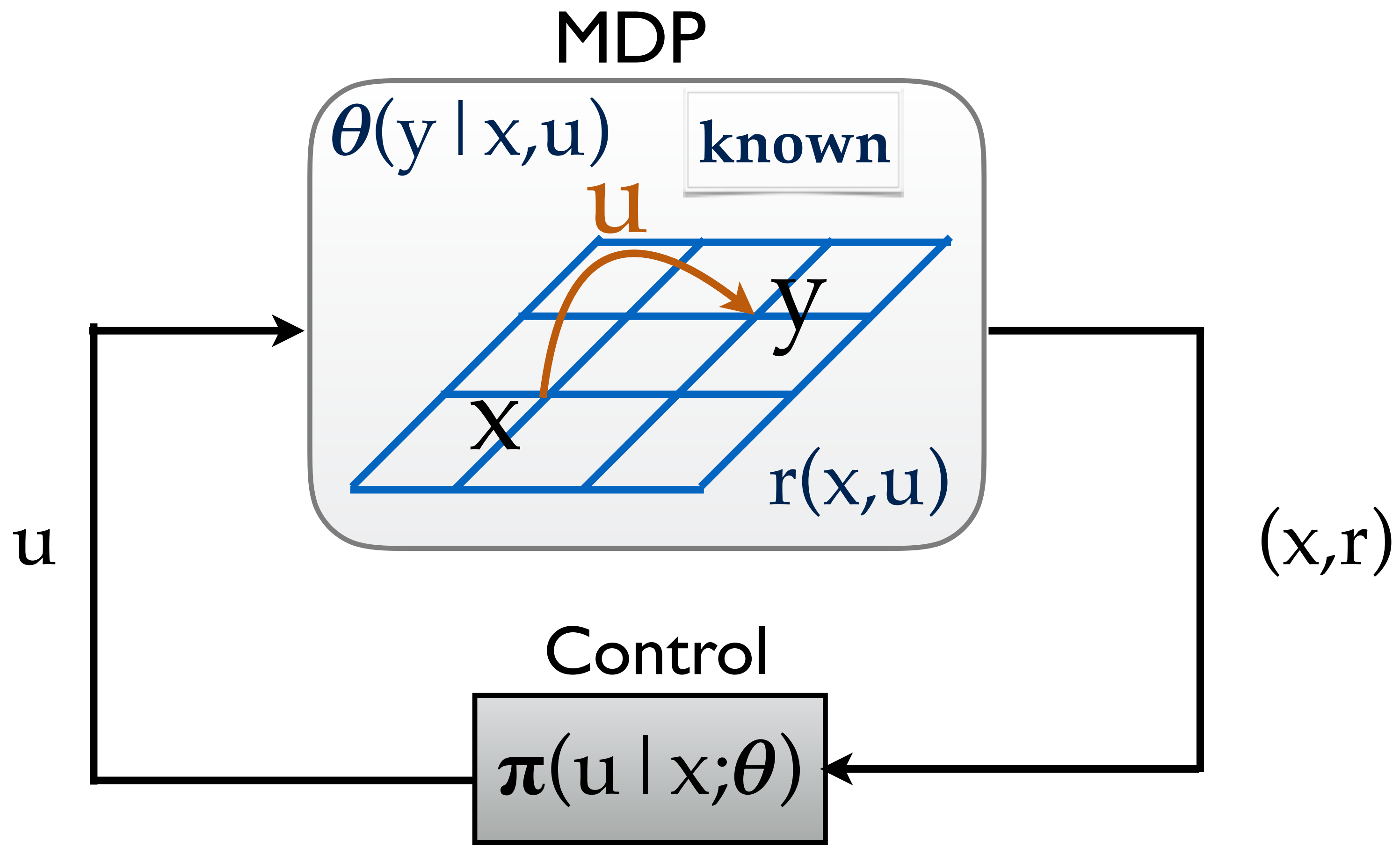
I. MDPs, Dynamic Programming

II. Bandit Models, Online Learning

III. PSDE: An RL Algorithm for Unknown MDPs

IV. PSDE Algorithm for Unknown Linear Stochastic Systems

A Markov Decision Process



Finite State space **X** *Finite* Action space **U**

$$V_\pi(\theta) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} r(x_t, u_t)]$$

Dynamic Programming

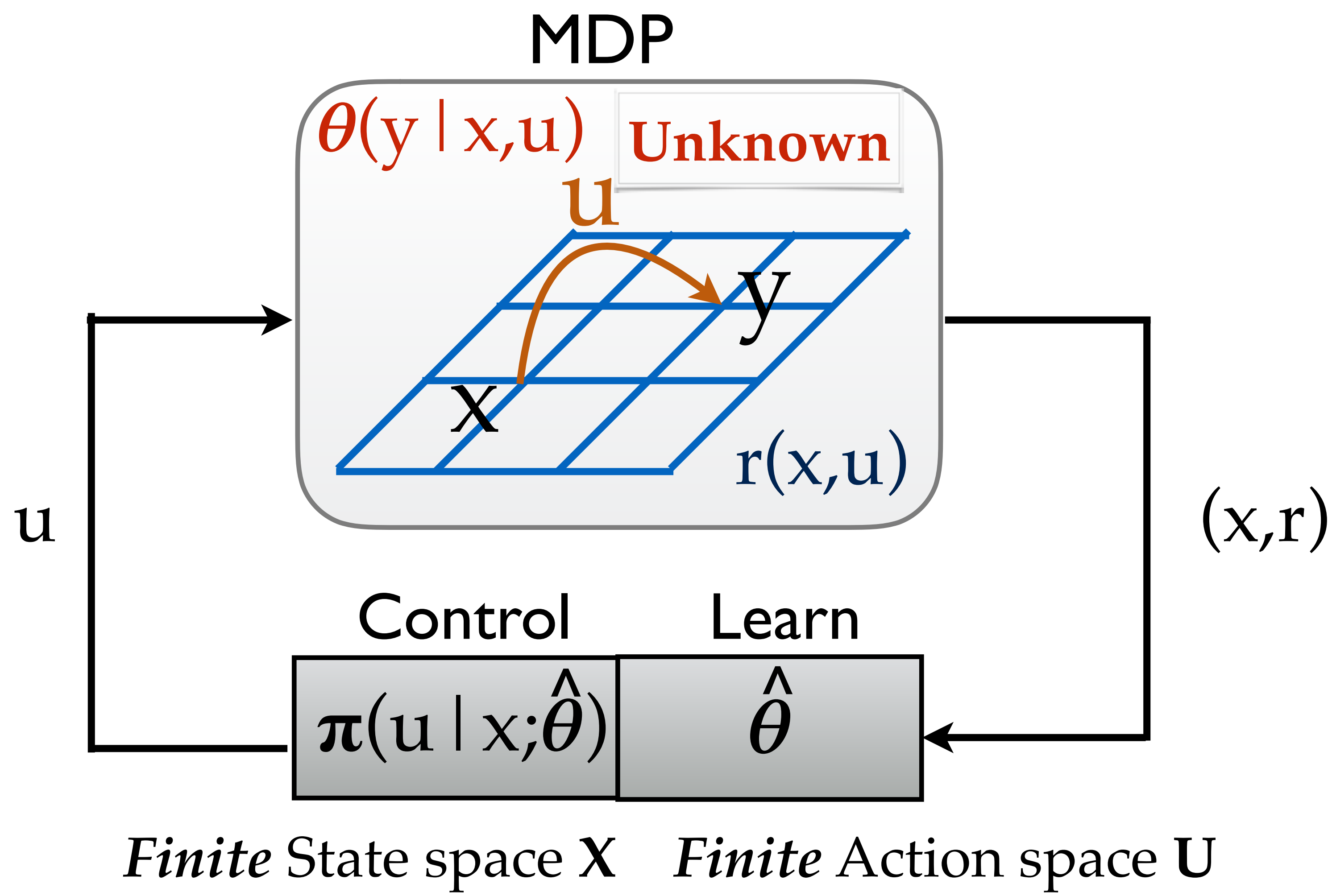
- Weakly communicating finite MDP
- Optimal average reward $V^*(\theta) = \sup_\pi V_\pi(\theta)$
- Bellman equation

$$V^*(\theta) + w^*(x,\theta) = \sup_u \{ r(x,u) + \sum_y \theta(y|x,u) w^*(x,\theta) \}$$

$E[w^*(y) \mid x,u]$

  - $w^*(x,\boldsymbol{\theta})$ is relative value function
- Solve by average-reward DP algorithms

Unknown Model



[Borkar-Varaiya'82]

- ★ True θ_o, unknown ~ prior μ
- ★ Learning policy $\phi_t(h_t)$, history h_t=(states,actions)
- ★ <u>Objective of Learning</u>: To find a nearly optimal policy at the fastest possible rate?

Bandit Models and Online Learning

Unknown $\boldsymbol{\theta}_1$

Unknown $\boldsymbol{\theta}_2$

- Reward on Heads = \$1, on Tails =0
- *Objective*: "max expected long-term total reward"

≡

- *min* (expected) Regret

$$\mathcal{R}_T(\phi) = T\theta_{max} - \mathbb{E}[\sum_{t=1}^{T} r_t]$$

- Lai & Robbins (1985) lower bound *O(log T)*
- UCB_1 algorithm achieves *O(log T)* [Agrawal'95, Auer, et al'02]
  - $$g_i(t, t_i) = \underline{X}_i + \sqrt{2\log t/t_i}$$

Optimism in the Face of Uncertainty (OFU)

The (Thompson) Posterior Sampling Algorithm

Unknown $\boldsymbol{\theta}_1$

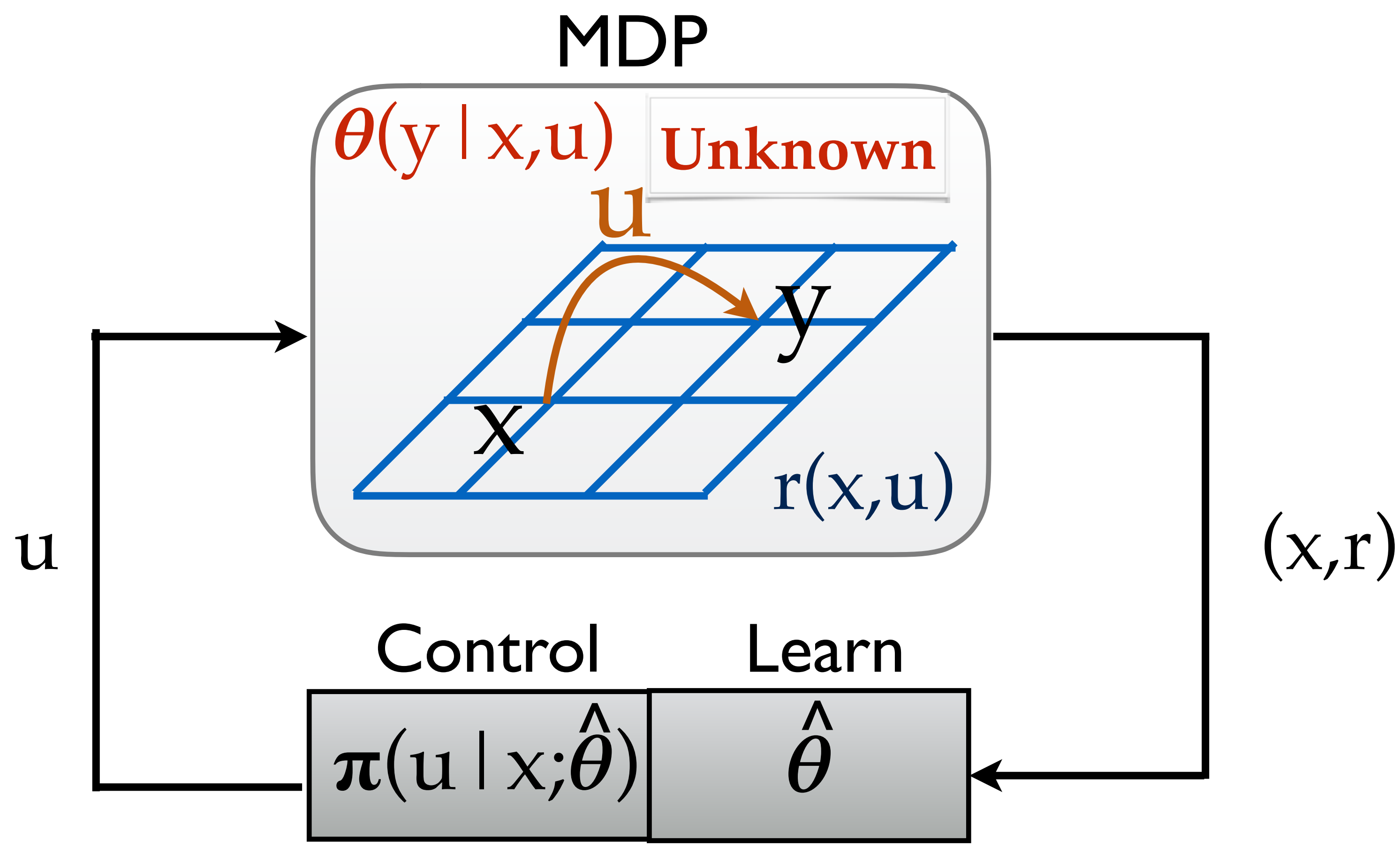
Unknown $\boldsymbol{\theta}_2$

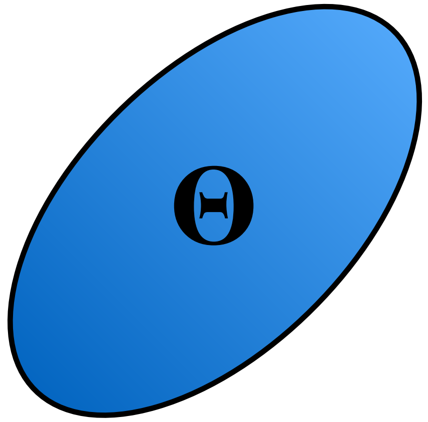
…

Unknown $\boldsymbol{\theta}_m$

- ★ Maintain a belief (posterior distribution), μ_i over $\boldsymbol{\theta}_i$
- ★ Sample $\hat{\boldsymbol{\theta}}_i$ from μ_i
- ★ Choose $i^* = arg\ max_i\ \hat{\boldsymbol{\theta}}_i$
- ★ Achieves (exp) regret $R_T(\phi) = O(log\ T)$
  - *Advantage: superior numerical performance, computationally simpler*
  - Thompson'33, Chapelle-Li'11, Agrawal-Goyal'12

Learning an Unknown MDP





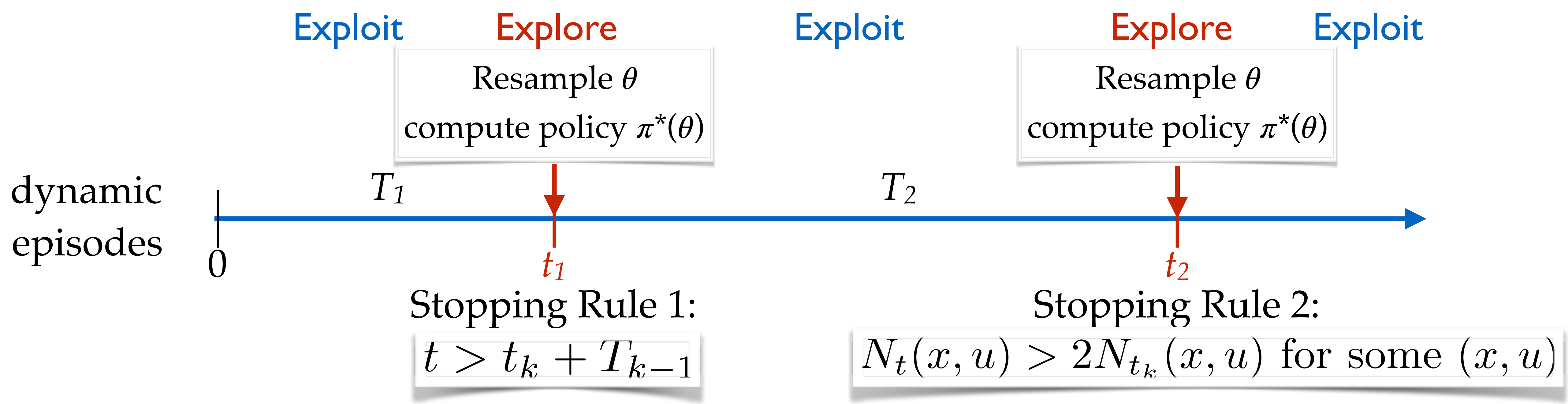
Finite State space **X** ***Finite*** Action space **U**

- Learning policy $\phi_t(h_t)$ to search over space $\boldsymbol{\Theta}$
- Objective of Learning:

$$\mathcal{R}_T(\phi) = TV^*(\theta_o) - \mathbb{E}[\sum_{t=1}^{T} r(x_t, u_t)]$$

- *Lower Bound* $=\Omega(\sqrt{T})$ [Tsitsiklis, et al (2010)]
- *OFU v. PS*

The **PSDE** Algorithm: Posterior Sampling with Dynamic Episodes



The **PSDE** Algorithm:

- ★ *Resample* θ from posterior μ_t at *end of every episode*
  - *Compute* policy optimal for sampled θ
- ★ *At each* t, update posterior $\mu_t(\theta) = \mathbb{P}(\theta|h_t)$ using Bayes' rule

Non-asymptotic Regret bound for PSDE

Theorem.*

If the MDP is *weakly communicating* and its *span* ≤ H, then

$$\mathcal{R}_T(PSDE) \leq \tilde{O}(HX\sqrt{UT})$$

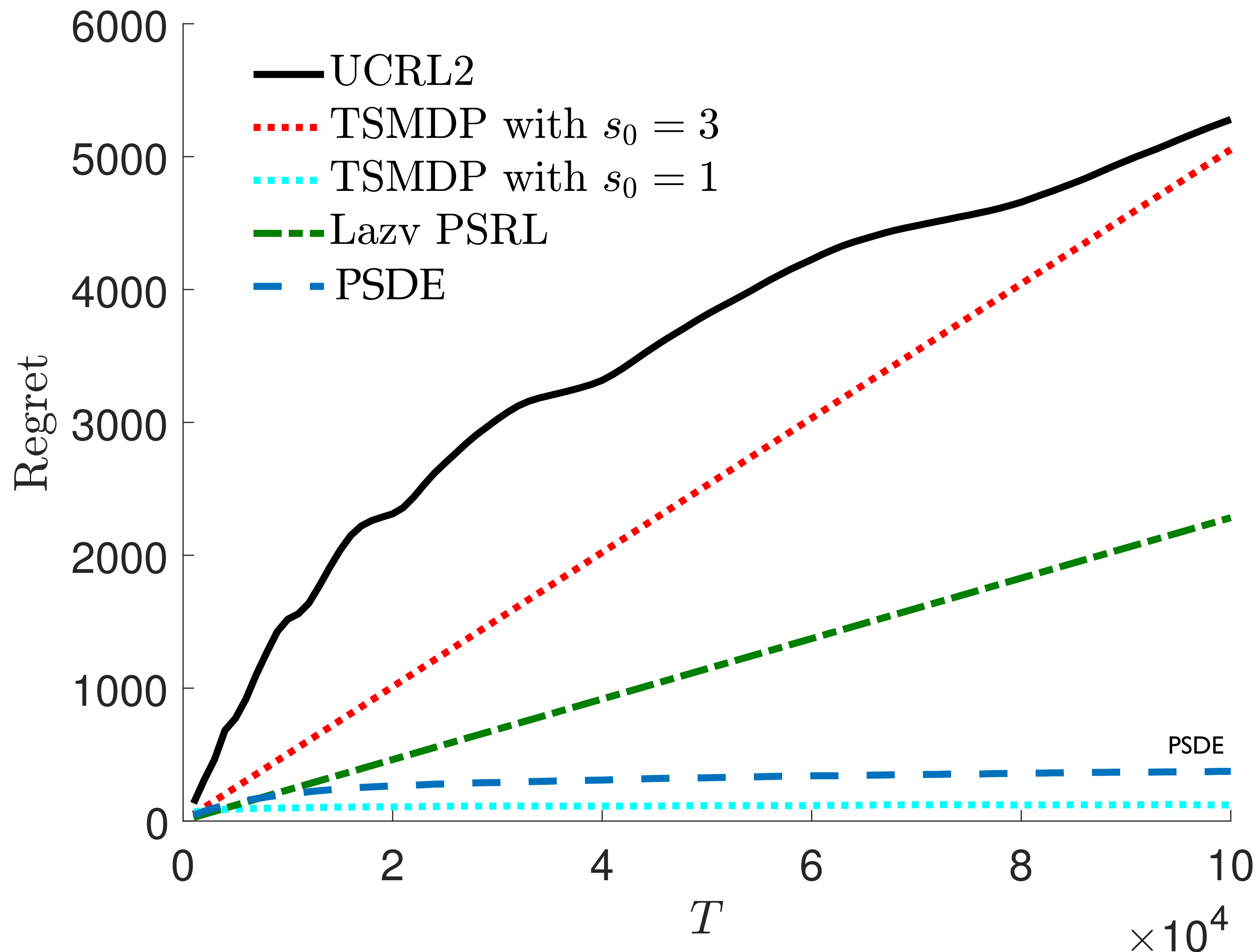
where X is state space size, and U is action space size.

- ★ Up to logarithmic factors, exact constants known
- ★ PSDE Algorithm works with approximately optimal policies in each episode also
- ★ Episode length can't increase faster

*Y. Ouyang, M. Gagrani, A. Nayyar and R. Jain, "Learning Unknown MDPs: A Thompson Sampling Approach", *NIPS, 2017*.

Numerical Performance

Riverswim Benchmark problem



UCRL2: [Jaksch, Ortner, Auer (2010)]

TSMDP: [Gopalan & Mannor (2015)]

Lazy-PSRL: [Yadkori & Szepesvari (2015)]

Proof Outline

- For any function f and RV X, algorithm must satisfy

$$E[f(\theta_k, X)] = E[f(\theta_o, X)]$$

- Upper bounds number of episodes

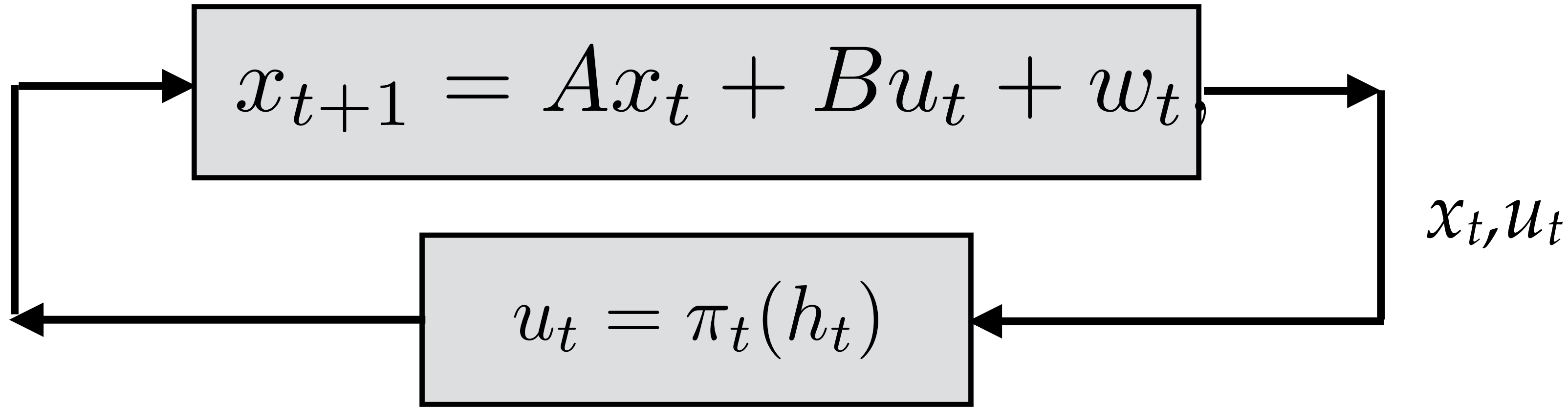
$$K_T \leq \sqrt{2XUT \log T}$$

- Upper bound between true and sampled parameters

$$T\mathbb{E}[V^*(\theta_o)] - \sum_{k=1}^{K_T} \mathbb{E}[T_k V^*(\theta_k)] \leq \mathbb{E}[K_T]$$

Unknown Stochastic Linear System

$\theta^\top = [A, B]$



$$c_t = x_t^\top Q x_t + u_t^\top R u_t$$

- Parameters θ unknown
- Regret $R_T(\pi) = \mathbb{E}[\sum_{t=1}^{T} c_t - TJ(\theta)]$
- Optimal control policy is linear:

$$u = G(\theta)x \quad \text{where} \quad G(\theta) = -(R + B^\top S(\theta B)^{-1} B^\top S(\theta) A.$$

Assumption 1: There is a set Θ such that for all $\theta \in \Theta$, there is a unique p.d. solution to the Ricatti equation

Stochastic Adaptive Control

- *Classical Adaptive Control…*
    - Certainty equivalence principle
    - Astrom-Wittenmark'94, Sastry'89, Narendra'89
- *Cost-biased Max Likelihood approach*
    - Campi and Kumar'98, Prandini-Campi'01,…
- *Optimism in the Face of Uncertainty (OFU)*
    - Yadkori-Szepesvari'11,'15, Van Roy, et al'12,'13,'16 (computation!)
    - Abeile-Lazaric'17 ~$O(T^{2/3})$

The Posterior Sampling with Dynamic Episodes (PSDE) Learning Algorithm

★ From data $z_t=[x_t,u_t]$, estimate parameters θ:

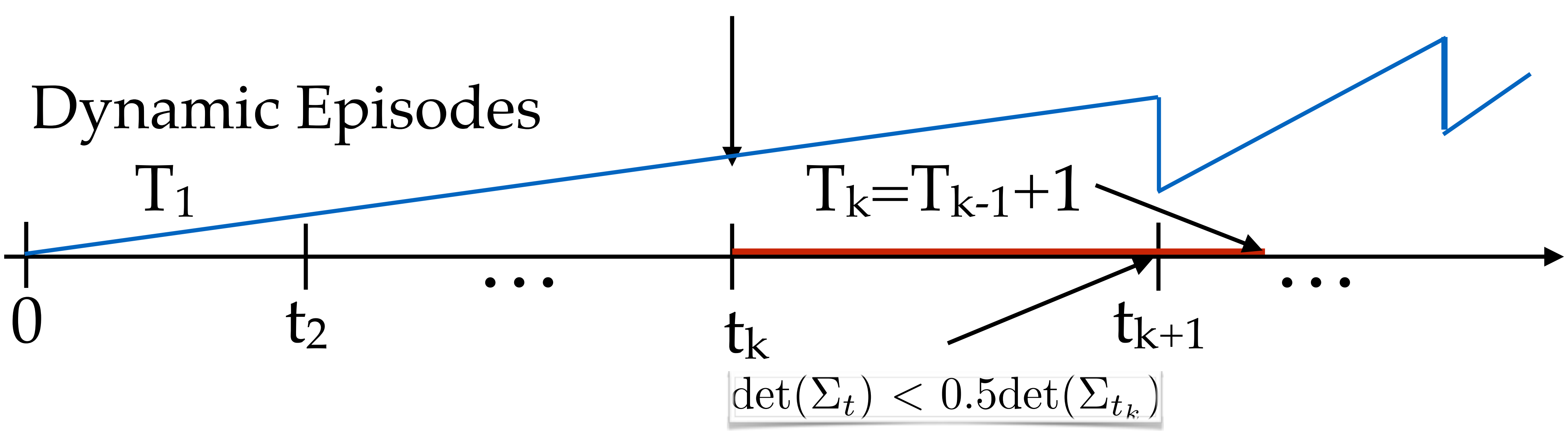
$$\hat{\theta}_{t+1}(i) = \hat{\theta}_t(i) + \frac{\Sigma_t z_t (x_{t+1}(i) - \hat{\theta}_t(i)^\top z_t)}{1 + z_t^\top \Sigma_t z_t}$$

$$\Sigma_{t+1} = \Sigma_t - \frac{\Sigma_t z_t z_t^\top \Sigma_t}{1 + z_t^\top \Sigma_t z_t}$$

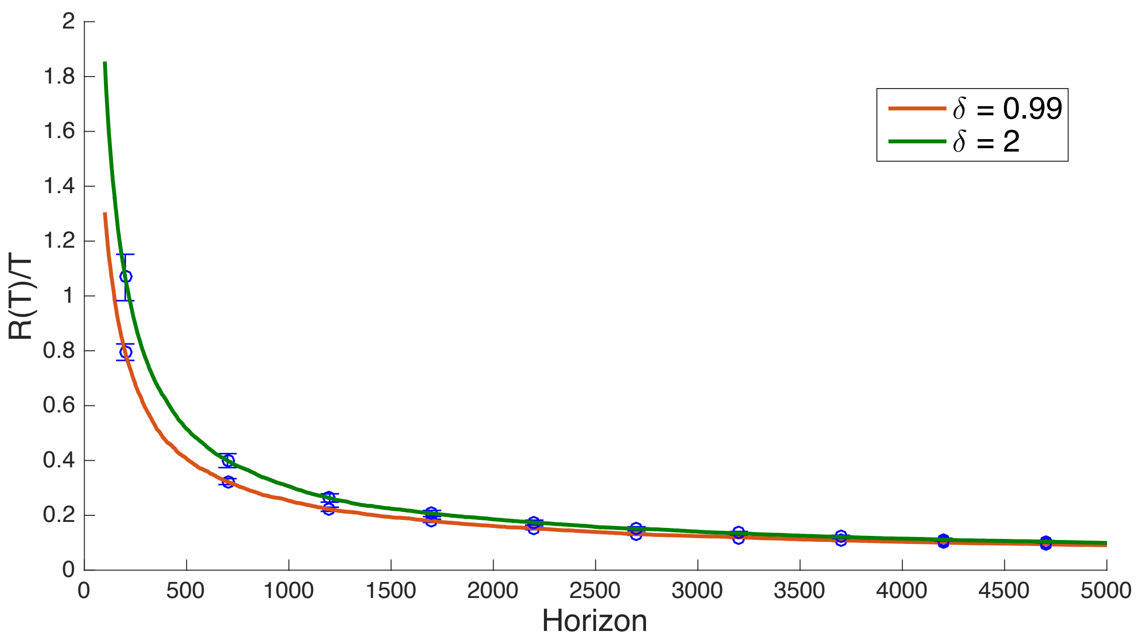
Posterior Sampling: Sample parameters $\tilde{\theta}_{t_k}$ from $\mu_{t_k}(\hat{\theta}_{t_k})$

Solve Ricatti equation

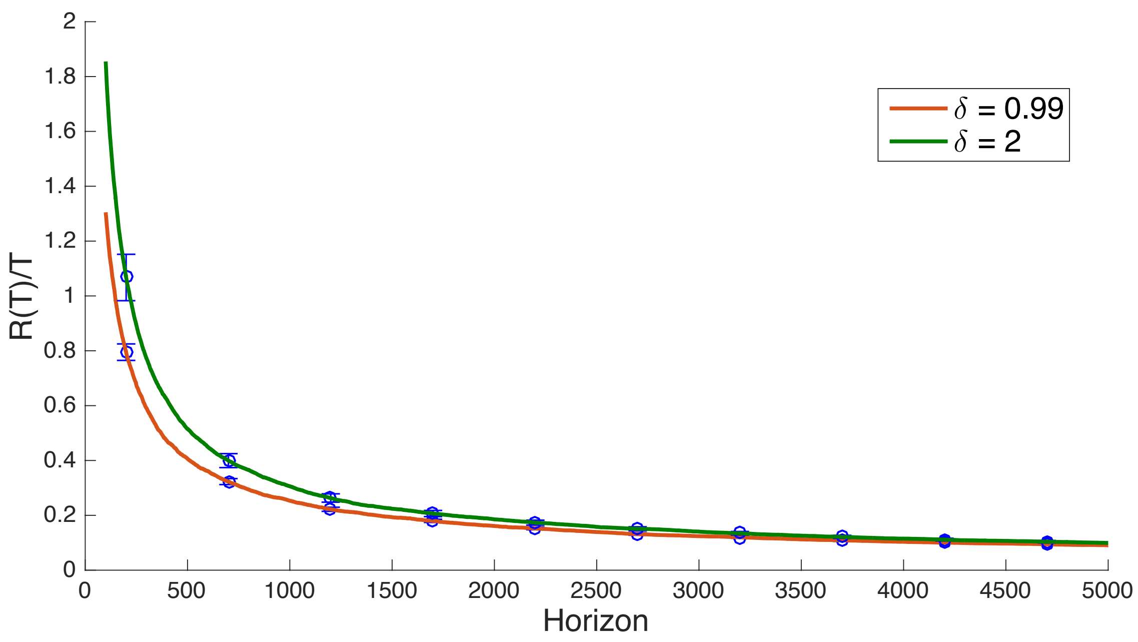
Compute Gain $G(\tilde{\theta}_{t_k})$



$\sqrt{T}$-Regret of PSDE Algorithm



Open Loop Stable System



Open Loop Unstable System

Assumption 2. State space X compact,

This implies for all $\theta \in \Theta$, spectral radius $\boldsymbol{\rho}(A_1+B_1G(\theta)) < \delta < 1$

Theorem.* Expected regret of PSDE, $\mathcal{R}_T(PSDE) \leq \tilde{O}(\sqrt{T})$

*Y. Ouyang, M. Gagrani and R. Jain, "Learning-based Control of Unknown Linear Systems with Thompson Sampling", arXiv:1709.04047

Conclusions

- Simple Posterior Sampling (PS)-based Learning-to-Control Algorithms
  - For MDPs and Linear Stochastic Systems
- Trades-off `*Exploration v. Exploitation*' nearly optimally to get O(√T) regret
  - Unlike OFU-type algorithms, computationally simple
  - A natural design
  - A deterministic schedule possible?
- Extensions
  - Continuous state space MDPs via function approximation
  - Time-varying systems

*Y. Ouyang, M. Gagrani, A. Nayyar and R. Jain, "*Learning Unknown MDPs: A Thompson Sampling Approach*", *NIPS, 2017*.
*Y. Ouyang, M. Gagrani and R. Jain, "*Learning-based Control of Unknown Linear Systems with Thompson Sampling*", arXiv:1709.04047, 2017