A modified Thompson sampling-based learning algorithm for unknown linear systems

Mukul Gagrani[a], Sagar Sudhakara[b], Aditya Mahajan[c], Ashutosh Nayyar[b], Ouyang Yi[d]

[a]Qualcomm, [b]USC, [c]McGill, [d]Preferred Networks

IEEE Conference on Decision and Control
9 December 2022

Significant interest in RL for control

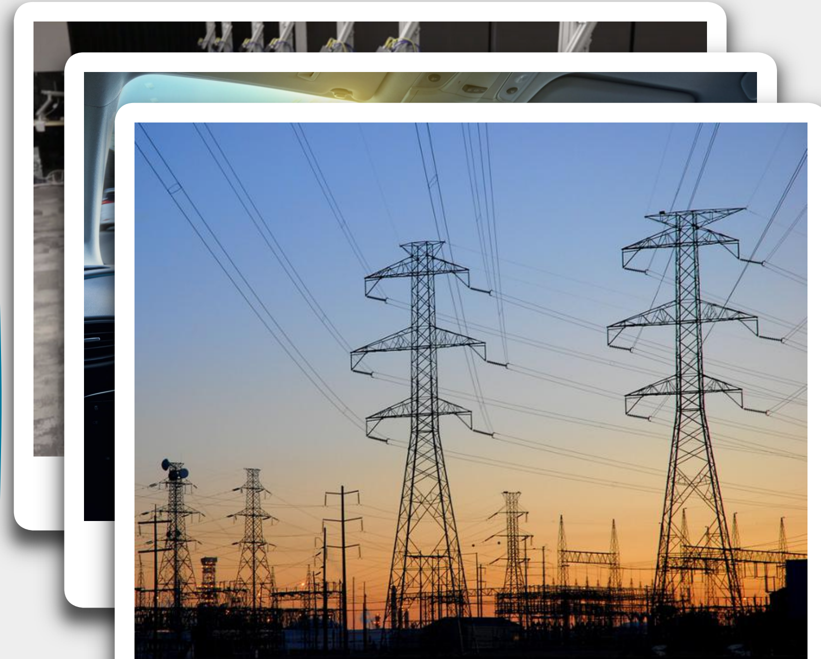
Significant interest in RL for control

Robotics

Significant interest in RL for control

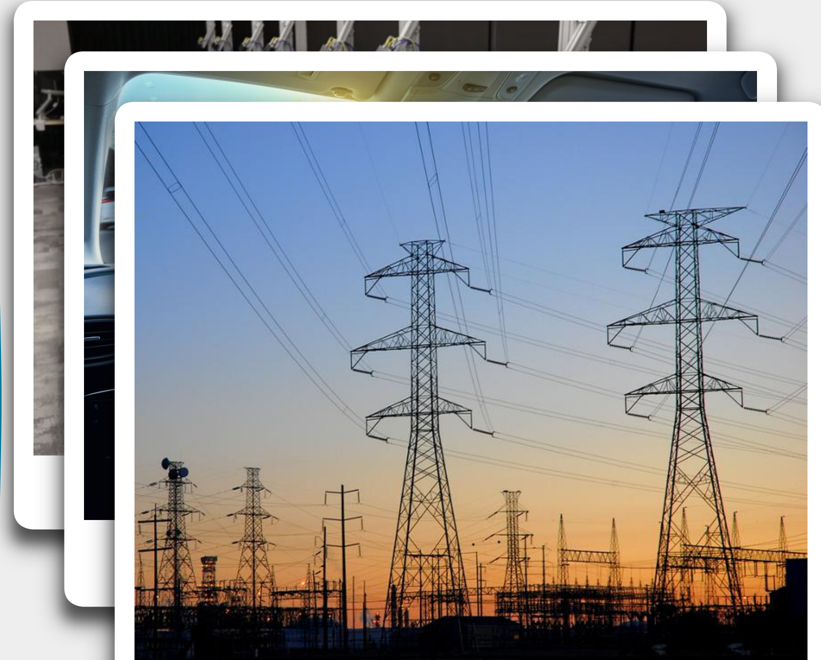
Self driving cars

Significant interest in RL for control



Smart Grids

Significant interest in RL for control

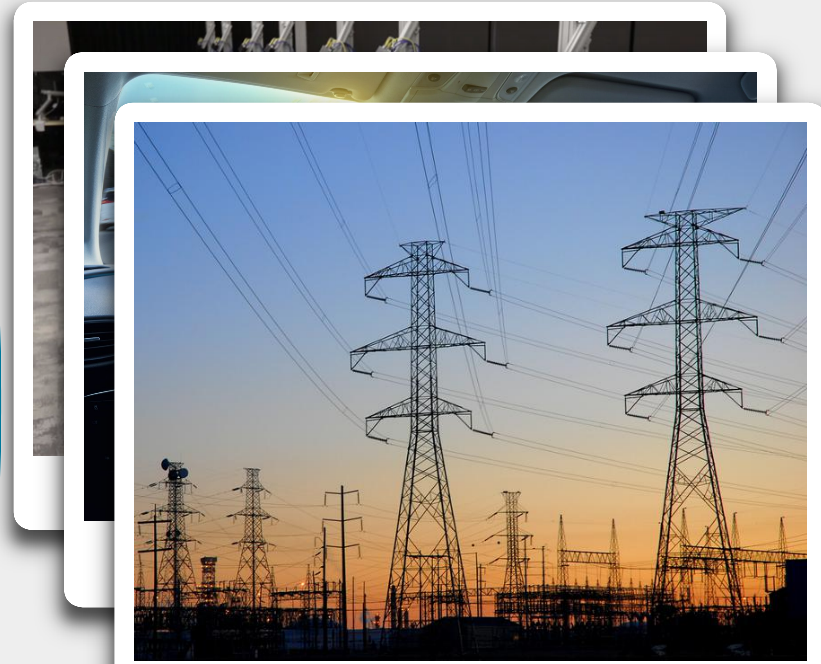


Smart Grids

Simplest setting: Linear quadratic regulation

- Different classes of RL algorithms
- Provide different performance guarantees under different assumptions on the uncertainty

Significant interest in RL for control



Simplest setting: Linear quadratic regulation

- Different classes of RL algorithms
- Provide different performance guarantees under different assumptions on the uncertainty

Relax the assumptions on uncertainty for a specific class of RL algorithms

Thompson sampling for LQ—(Gagrani et. al.)

Learning in unknown linear systems

Linear Quadratic Regulation

$$x_{t+1} = A_\theta x_t + B_\theta u_t + w_t, \quad w_t \sim N(0, \sigma_w^2 I)$$

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.$$

Given $\theta^\top = [A_\theta, B_\theta]$, choose a policy π to minimize

$$J(\pi; \theta) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} c(x_t, w_t) \right].$$

Learning in unknown linear systems

Linear Quadratic Regulation

$$x_{t+1} = A_\theta x_t + B_\theta u_t + w_t, \quad w_t \sim N(0, \sigma_w^2 I)$$
$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.$$

Given $\theta^\top = [A_\theta, B_\theta]$, choose a policy π to minimize

$$J(\pi; \theta) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} c(x_t, w_t)\right].$$

Optimal solution

When θ is known and (A_θ, B_θ) is stabilizable, optimal policy $\pi^\star$ is given by

$$u_t = G(\theta) x_t$$

where

- $G(\theta) = -(R + B_\theta^\top S_\theta B_\theta)^{-1} B_\theta^\top S_\theta A_\theta$
- S_θ is the solution of the algebraic Riccati eqn

Moreover: $J(\pi_\theta^\star; \theta) = \sigma_w^2 \operatorname{Tr}(S_\theta)$.

Learning in unknown linear systems

Linear Quadratic Regulation

$$x_{t+1} = A_\theta x_t + B_\theta u_t + w_t, \quad w_t \sim N(0, \sigma_w^2 I)$$
$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.$$

Given $\theta^\top = [A_\theta, B_\theta]$, choose a policy π to minimize

$$J(\pi; \theta) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} c(x_t, w_t) \right].$$

Learning setup

- True parameter $\theta_\star$ is unknown
- Regret of any learning-based policy π:

$$R(T; \pi) = \mathbb{E}^\pi \left[\sum_{t=1}^{T} c(x_t, u_t) - T J(\pi^\star_{\theta_\star}, \theta_\star) \right].$$

Optimal solution

When θ is known and (A_θ, B_θ) is stabilizable, optimal policy $\pi^\star$ is given by

$$u_t = G(\theta) x_t$$

where

- $G(\theta) = -(R + B_\theta^\top S_\theta B_\theta)^{-1} B_\theta^\top S_\theta A_\theta$
- S_θ is the solution of the algebraic Riccati eqn

Moreover: $J(\pi^\star_\theta; \theta) = \sigma_w^2 \mathrm{Tr}(S_\theta)$.

Learning in unknown linear systems

Linear Quadratic Regulation

$$x_{t+1} = A_\theta x_t + B_\theta u_t + w_t, \quad w_t \sim N(0, \sigma_w^2 I)$$

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t.$$

Given $\theta^\top = [A_\theta, B_\theta]$, choose a policy π to minimize

$$J(\pi; \theta) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} c(x_t, w_t)\right].$$

Learning setup

- True parameter $\theta_\star$ is unknown
- Regret of any learning-based policy π:

$$R(T; \pi) = \mathbb{E}^\pi\left[\sum_{t=1}^{T} c(x_t, u_t) - T J(\pi^\star_{\theta_\star}, \theta_\star)\right].$$

Optimal solution

When θ is known and (A_θ, B_θ) is stabilizable, optimal policy $\pi^\star$ is given by

$$u_t = G(\theta) x_t$$

where

- $G(\theta) = -(R + B_\theta^\top S_\theta B_\theta)^{-1} B_\theta^\top S_\theta A_\theta$
- S_θ is the solution of the algebraic Riccati eqn

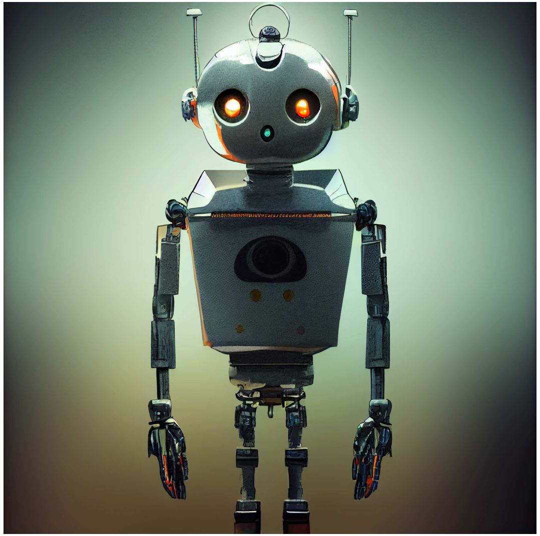
Moreover: $J(\pi^\star_\theta; \theta) = \sigma_w^2 \mathrm{Tr}(S_\theta)$.

Key research question

▷ How does regret scale with horizon T?

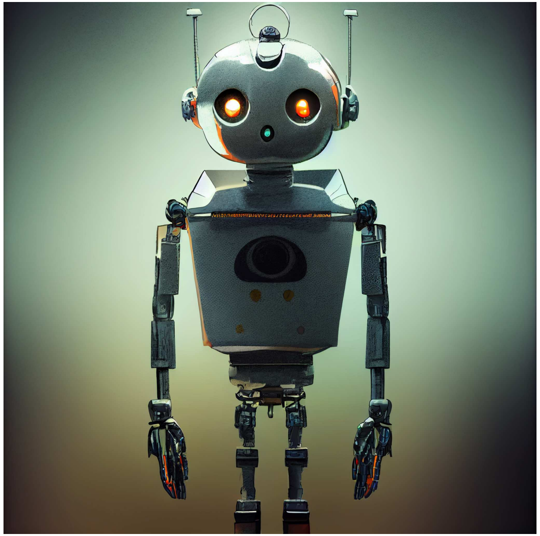
Different learning frameworks

Explore vs Exploit



Different learning frameworks

Explore vs Exploit

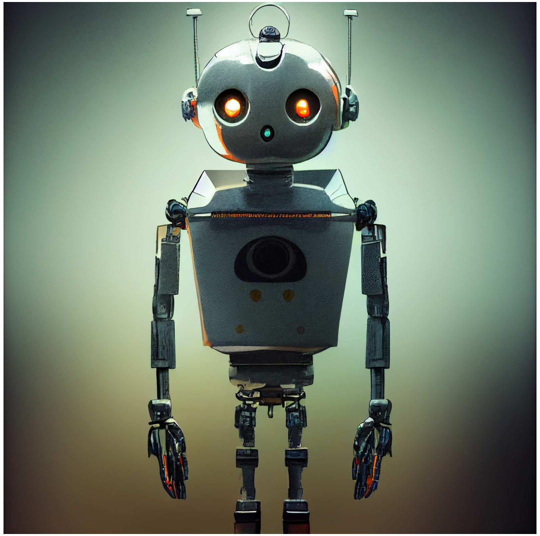


Certainty equivalence

- Generate estimate $\hat{\theta}_t$ based on past observations
- Use controller: $u_t = G(\hat{\theta}_t)\,x_t + \varepsilon_t$ (exploration noise)

Different learning frameworks

Explore vs Exploit



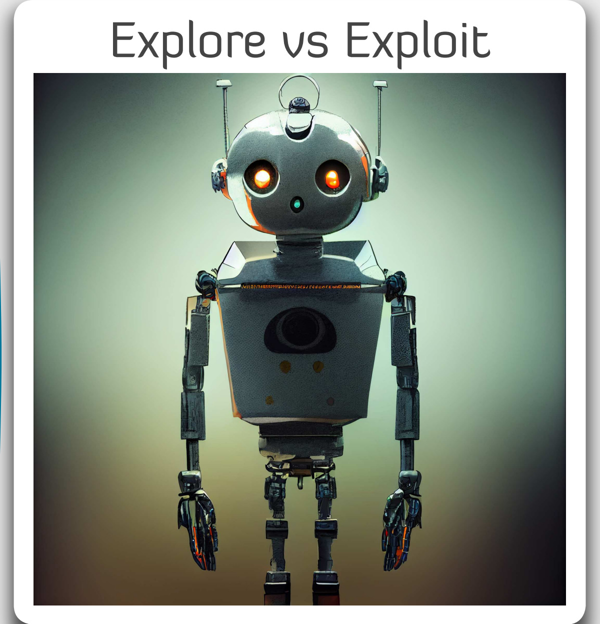
Certainty equivalence

- Generate estimate $\hat{\theta}_t$ based on past observations
- Use controller: $u_t = G(\hat{\theta}_t)x_t + \varepsilon_t$ (exploration noise)

Upper Confidence Bound (UCB)

- Generate UCB estimate $\bar{\theta}_t$ based on past observations.
- Use controller: $u_t = G(\bar{\theta}_t)x_t$

Different learning frameworks



Certainty equivalence

- Generate estimate $\hat{\theta}_t$ based on past observations
- Use controller: $u_t = G(\hat{\theta}_t)x_t + \varepsilon_t$ (exploration noise)

Upper Confidence Bound (UCB)

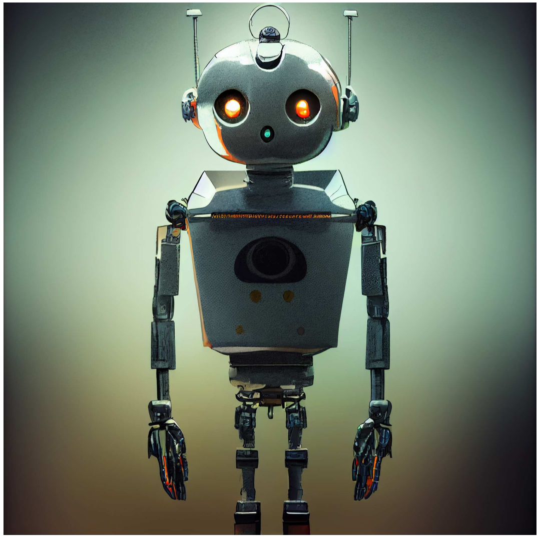
- Generate UCB estimate $\bar{\theta}_t$ based on past observations.
- Use controller: $u_t = G(\bar{\theta}_t)x_t$

Posterior/Thompson sampling

- Maintain posterior μ_t on $\theta_\star$
- Sample $\tilde{\theta}_t \sim \mu_t$
- Use controller: $u_t = G(\tilde{\theta}_t)x_t$

Different learning frameworks

Explore vs Exploit



Certainty equivalence

- Generate estimate $\hat{\theta}_t$ based on past observations
- Use controller: $u_t = G(\hat{\theta}_t)x_t + \varepsilon_t$ (exploration noise)

Upper Confidence Bound (UCB)

- Generate UCB estimate $\bar{\theta}_t$ based on past observations.
- Use controller: $u_t = G(\bar{\theta}_t)x_t$

Posterior/Thompson sampling

- Maintain posterior μ_t on $\theta_\star$
- Sample $\tilde{\theta}_t \sim \mu_t$
- Use controller: $u_t = G(\tilde{\theta}_t)x_t$

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

- Bayesian RL algorithm
- Generalization on Thompson sampling (or posterior sampling) for bandits

- Very simple algorithm which requires no hyper-parameter tuning and works well in practice

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

- Bayesian RL algorithm
- Generalization on Thompson sampling (or posterior sampling) for bandits
- Very simple algorithm which requires no hyper-parameter tuning and works well in practice

Assumptions on the true parameter

- $\theta_\star$ lies in a compact set.

$$\theta_\star^\top = \left[\begin{array}{c|c} A_\star & B_\star \end{array} \right] \in \underbrace{\Omega}_{\text{Compact set}}$$

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

- Bayesian RL algorithm
- Generalization on Thompson sampling (or posterior sampling) for bandits
- Very simple algorithm which requires no hyper-parameter tuning and works well in practice

Assumptions on the true parameter

- $\theta_\star$ lies in a compact set.
- Independent **truncated Gaussian prior** on each row of $\theta_\star^\top$:

$$\bar{\mu}_1(\theta) = \left[\prod_{i=1}^{n} N(\hat{\theta}_1(i), \Sigma_1) \right] \Bigg|_{\Omega}$$

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

Properties of the posterior

- Posterior μ_t is also truncated Gaussian with $\mu_t(\theta) = \left[\prod_{i=1}^{n} N(\hat{\theta}_t(i), \Sigma_t)\right]\Big|_{\Omega}$ where

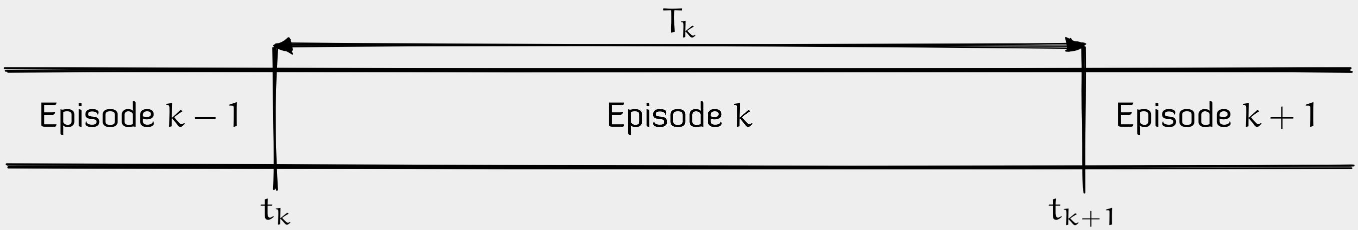
$$\hat{\theta}_{t+1}(i) = \hat{\theta}_t(i) + \frac{\Sigma_t z_t (x_{t+1}(i) - \hat{\theta}_t(i)^\top z_t)}{\sigma_w^2 + z_t^\top \Sigma_t z_t}$$

$$\Sigma_{t+1}^{-1} = \Sigma_t^{-1} + \frac{1}{\sigma_w^2} z_t z_t^\top .$$

where $z_t = \mathrm{vec}(x_t, u_t)$.

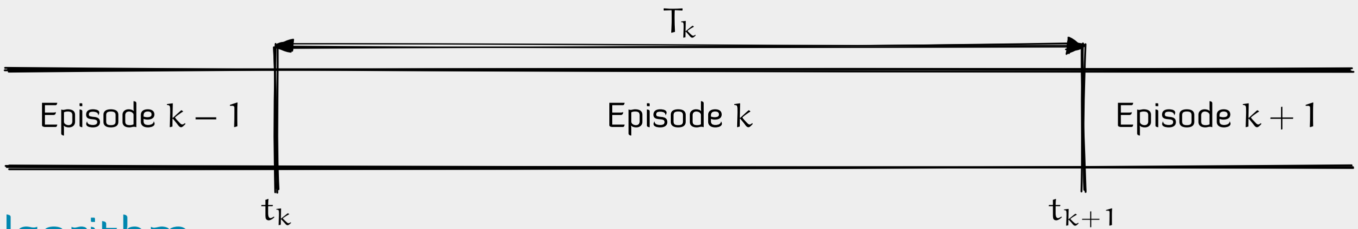
Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]



Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

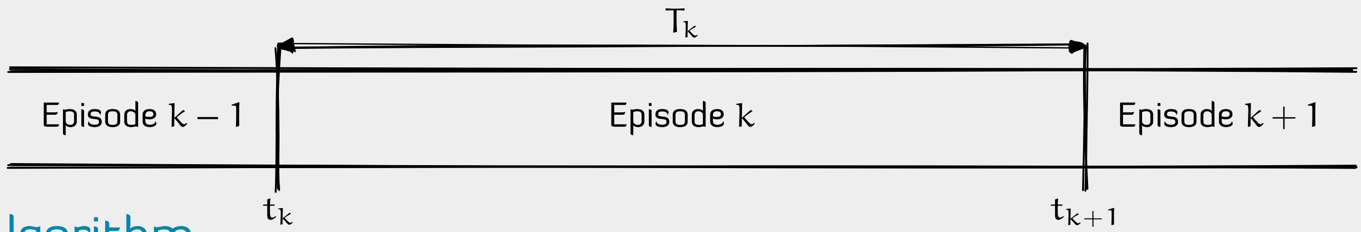


TSDE Algorithm

- At start of episode: Sample $\tilde{\theta}_k \sim \mu_{t_k}$
- During the episode: Use $u_t = G(\tilde{\theta}_k)\,x_t$
- Terminate episode if: $(t - t_k > T_{k-1})$ or $(\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k})$

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]



TSDE Algorithm

- At start of episode: Sample $\tilde{\theta}_k \sim \mu_{t_k}$
- During the episode: Use $u_t = G(\tilde{\theta}_k)\, x_t$
- Terminate episode if: $(t - t_k > T_{k-1})$ or $(\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k})$

Intuition: $\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k}$ implies that current posterior is much better than the posterior at the start of the episode. **Resample to exploit this knowledge**

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

Assumption A1 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\|A_\theta + B_\theta G(\phi)\| \leq \delta$.

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

Assumption A1 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\|A_\theta + B_\theta G(\phi)\| \le \delta$.

Discussion on assumptions

- A1 is a **strong assumption**.
- Requires that close loop system dynamics under any mismatched controller should have **spectral norm** less than one.

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

Assumption A1 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\|A_\theta + B_\theta G(\phi)\| \leq \delta$.

Theorem Under A1, $R(T; \text{TSDE}) \leq C\sqrt{T}\,(\log T)^q$

Discussion on assumptions

- A1 is a **strong assumption**.
- Requires that close loop system dynamics under any mismatched controller should have **spectral norm** less than one.

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

Assumption A1 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\|A_\theta + B_\theta G(\phi)\| \leq \delta$.

Theorem Under A1, $R(T; \text{TSDE}) \leq C\sqrt{T}\,(\log T)^q$

Discussion on assumptions

- A1 is a **strong assumption**.
- Requires that close loop system dynamics under any mismatched controller should have **spectral norm** less than one.

Discussion on the results

- The regret is **Bayesian regret**, i.e., includes an expectation over the prior.
- Different from frequentist regret, which provides a high-probability bound on regret for the true parameter.

Thompson sampling with dynamic episodes (TSDE)

[Ouyang, Gagrani, Jain 2020]

Why bother with TSDE

- Works very well in practice. Requires no parameter tuning.
- Continues to work well when A1 is violated.

Assumption ... $\le \delta$.

Theorem Under A1, $R(T; \text{TSDE}) \le C\sqrt{T}\,(\log T)^q$

Discussion on assumptions

- A1 is a **strong assumption**.
- Requires that close loop system dynamics under any mismatched controller should have **spectral norm** less than one.

Discussion on the results

- The regret is **Bayesian regret**, i.e., includes an expectation over the prior.
- Different from **frequentist regret**, which provides a high-probability bound on regret for the true parameter.

The strong assumption appears to be a limitation of the proof technique (and not the algorithm).

Can we relax it?

How should the stability assumption be relaxed?

Ideally, should only require the true $\theta_\star$ to be stabilizable

- Bayesian equivalent:

$$\mathbb{P}(\theta \in \Omega : \theta \text{ is stabilizable}) = 1$$

How should the stability assumption be relaxed?

Ideally, should only require the true $\theta_\star$ to be stabilizable

- Bayesian equivalent:

$$\mathbb{P}(\theta \in \Omega : \theta \text{ is stabilizable}) = 1$$

. . . and be able to construct a stabilizing controller in finite time

- Don't know how to do that in Bayesian setting
- Guaranteeing stability with high probability is not sufficient

How should the stability assumption be relaxed?

Ideally, should only require the true $\theta_\star$ to be stabilizable

- Bayesian equivalent:

$$\mathbb{P}(\theta \in \Omega : \theta \text{ is stabilizable}) = 1$$

. . . and be able to construct a stabilizing controller in finite time

- Don't know how to do that in Bayesian setting
- Guaranteeing stability with high probability is not sufficient

First step in weakening the stability assumption

- Assumption A1 is defined in terms of **spectral norm**
- A natural relaxation is to replace spectral norm by **spectral radius**.
- . . . which is what we do in this paper

This paper: Natural relaxation of Assumption A1

Assumption A2 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\rho(A_\theta + B_\theta G(\phi)) \leq \delta$.

- Controller for system ϕ stabilizes system θ
- Still a strong assumption, but weaker (and more natural) than A1.

This paper: Natural relaxation of Assumption A1

Assumption A2 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\rho(A_\theta + B_\theta G(\phi)) \leq \delta$.

- Controller for system ϕ stabilizes system θ
- Still a strong assumption, but weaker (and more natural) than A1.

Proof of regret bound of TSDE breaks down

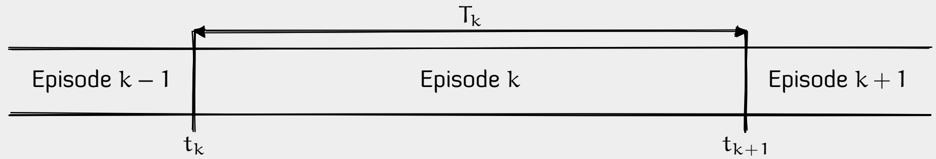
- Proof relies on showing that there is some constant α_0 such that

$$(\star) \qquad \mathbb{E}\Big[\max_{1 \leq t \leq T} \|x_t\|\Big] \leq \sigma_w + \alpha_0 \mathbb{E}\Big[\max_{1 \leq t \leq T} \|w_t\|\Big]$$

- Under (A1), $\mathbb{E}[\|x_{t+1}\|] \leq \delta \mathbb{E}[\|x_t\|] + \mathbb{E}[\|w_t\|]$, which implies $\alpha_0 = 1/(1-\delta)$.
- Such a bound does not work under (A2).

Need to modify the algorithm

Modified TSDE



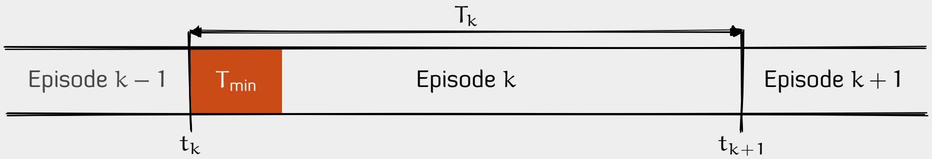
Intuition

- Under (A2), in each episode the system is asymptotically stable.
- Asymptotic stability implies exponential stability.
- So, if the episode is sufficiently large, we can show that

$$\mathbb{E}[\|x_{t_{k+1}}\|] \leq \beta \mathbb{E}[\|x_{t_k}\|] + \bar{\alpha}\mathbb{E}\Big[\max_{t_k \leq t \leq t_{k+1}} \|w_t\|\Big]$$

which implies $(\star)$.

Modified TSDE



Intuition

- Under (A2), in each episode the system is asymptotically stable.
- Asymptotic stability implies exponential stability.
- So, if the episode is sufficiently large, we can show that

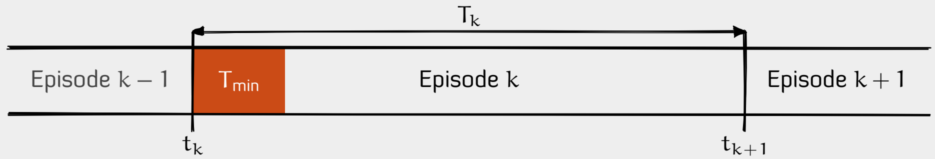
$$\mathbb{E}[\|x_{t_{k+1}}\|] \leq \beta \mathbb{E}[\|x_{t_k}\|] + \bar{\alpha} \mathbb{E}\Big[\max_{t_k \leq t \leq t_{k+1}} \|w_t\|\Big]$$

which implies $(\star)$.

Proposed modification

- To ensure that each episode is sufficiently large, do not stop in the first T_{min} steps of an episode
- See paper for choice of T_{min}.

Modified TSDE



Intuition

- Under (A2), in each episode the system is asymptotically stable.
- Asymptotic stability implies exponential stability.
- So, if the episode is sufficiently large, we can show that

$$\mathbb{E}[\|x_{t_{k+1}}\|] \leq \beta \mathbb{E}[\|x_{t_k}\|] + \bar{\alpha} \mathbb{E}\Big[\max_{t_k \leq t \leq t_{k+1}} \|w_t\|\Big]$$

which implies $(\star)$.

Proposed modification

- To ensure that each episode is sufficiently large, do not stop in the first T_{min} steps of an episode
- See paper for choice of T_{min}.

Implication

- The second stopping condition is not triggered for the T_{min} steps of each episode.
- Requires other changes in the proof argument. See paper for details.

Main results

Assumption A2 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\rho(A_\theta + B_\theta G(\phi)) \leq \delta$.

Theorem Under A2, $R(T; \text{m-TSDE}) \leq C\sqrt{T}\,(\log T)^q$

Main results

Assumption A2 There exists an $\delta \in (0, 1)$ such that for any $\theta, \phi \in \Omega$, $\rho(A_\theta + B_\theta G(\phi)) \leq \delta$.

Theorem Under A2, $R(T; \text{m-TSDE}) \leq C\sqrt{T}\,(\log T)^q$

Conclusion

- Relaxed a technical assumption for TSDE.
- Although A2 is weaker than A1, it still a **strong assumption**.

- Numerical experiments suggest that regret scales $\tilde{O}(\sqrt{T})$ even when A2 is not satisfied.
- Open question: How to further relax the stability assumption?

Thank you