Learning LQR via Thompson sampling

Yeoneung Kim
Seoul National University of Science and Technology (SeoulTech)

Joint with Gihun Kim and Insoon Yang (SNU)

1st Korean Bandit Workshop
June 20, 2024

Stochastic LQR

For random noise ω_t,

- The dynamics is given by

$$x_{t+1} = Ax_t + Bu_t + \omega_t$$

- The cost is given by

$$J = \mathbb{E}[\sum_{t=0}^{N-1}(x_t^\top Qx_t + u_t^\top Ru_t) + x_N^\top Q_f x_N]$$

  where the expectation is taken over all noises.

A goal is to find the control sequence $u_1, ..., u_{N-1}$ minimizing the cost.

Bellman's Equation for stochastic LQR

Let us begin by defining

$$V_t(z) := \min_{u_t,\ldots,u_{N-1}} \mathbf{E}\left[\sum_{k=t}^{N-1}\left(x_k^\top Q x_k + u_k^\top R u_k\right) + x_N^\top Q_f x_N \,\middle|\, x_t = z\right]$$

with $V_N(z) = z^\top Q_f z$ as before. Deduce that

$$\begin{aligned}
&V_{N-1}(z)\\
&= \min_u \left(z^\top Q z + u^\top R u + \mathbf{E}\left[(Az+Bu+w)^\top Q_f (Az+Bu+w)\right]\right)\\
&= \min_u \left(z^\top Q z + u^\top R u + (Az+Bu)^\top Q_f (Az+Bu) + \mathbf{E}\left[2w^\top Q_f (Az+Bu) + w^\top Q_f w\right]\right)\\
&= \min_u \left(z^\top Q z + u^\top R u + (Az+Bu)^\top Q_f (Az+Bu)\right) + \mathbf{E}\left[\mathrm{Tr}(w^\top Q_f w)\right]\\
&= \min_u \left(z^\top Q z + u^\top R u + (Az+Bu)^\top Q_f (Az+Bu)\right) + \mathrm{Tr}(Q_f \Sigma_w)\\
&= z^\top \left(A^\top Q_f A + Q - A^\top Q_f B (B^\top Q_f B + R)^{-1} B^\top Q_f A\right) z + \mathrm{Tr}(Q_f \Sigma_w),
\end{aligned}$$

One can infer that

$$V_t(z) = z^\top P_t z + r_t$$

.

Bellman's Equation for stochastic LQR

Substituting $V_t(z) = z^\top P_t z + r_t$,

$$
\begin{aligned}
V_{t-1}(z) &= \min_u \left(z^\top Q z + u^\top R u + \mathbf{E}\left[(Az + Bu + w)^\top P_t (Az + Bu + w) + r_t \right] \right) \\
&= \min_u \left(z^\top Q z + u^\top R u + (Az + Bu)^\top P_t (Az + Bu) \right) + \mathbf{E}\left[\mathrm{Tr}(w^\top P_t w) \right] + r_t \\
&= \min_u \left(z^\top Q z + u^\top R u + (Az + Bu)^\top P_t (Az + Bu) \right) + \mathrm{Tr}(P_t \Sigma_w) + r_t \\
&= z^\top \left(A^\top P_t A + Q - A^\top P_t B (B^\top P_t B + R)^{-1} B^\top P_t A \right) z + \mathrm{Tr}(P_t \Sigma_w) + r_t.
\end{aligned}
$$

As a result,

$$
\boxed{
\begin{aligned}
&P_N = Q_f \\
&r_N = 0 \\
&P_t = A^\top P_{t+1} A + Q - A^\top P_{t+1} B (B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A \quad && \text{for } t = N-1, \dots, 0 \\
&K_t = -(B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A && \text{for } t = N-1, \dots, 0 \\
&r_t = r_{t+1} + \mathrm{Tr}(P_{t+1} \Sigma_w) && \text{for } t = N-1, \dots, 0
\end{aligned}
}
$$

Infinite horizon LQR

We want to optimize

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\sum_{t=0}^{N-1}(x_t^\top Q x_t + u_t^\top R u_t) + x_N^\top Q_f x_N$$

subject to $x_{t+1} = Ax_t + Bu_t + \omega_t$, with $x(0) = x_0$ has a finite value if the system does not grow rapidly. Otherwise, the cost will be infinity.

Theorem

Assume (A, B) is controllable and $(A, \sqrt{Q})$ is observable. Then, there exists positive definite matrix P such that $\lim_{t\to\infty} P_t = P$ solving the Riccati equation:

$$P = A^\top PA + Q - A^\top PB(B^\top PB + R)^{-1}B^\top PA.$$

Moreover, the spectral radius of $A + BK$ is strictly less than 1 where $K = -(BPB + R)^{-1}B^\top PA$.

Learning infinite horizon LQR

What if A and B are unknown? our goal is to design an algorithm that can learn the unknown system parameters minimizing the regret.

Problem setup

Consider a linear stochastic system of the form

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t = 1, 2, \ldots,$$

with cost

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

Then of our interest is how we can minimize the regret:

$$R(T) = \sum_{t=0}^{T}(c_t - J_*)\,,$$

where J_* is the infimum over all policies.

Various approaches

- Force exploration : Regret can have strong worst-case regret
- OFU : Construct high-probability confidence set and optimize in the set. Frequentist regret $O(\sqrt{T})$ yet computationally unfavorable.
  1. Abbasi-Yadkori (2011)
  2. Abeille (2020) - Lagrange relaxation
- Beyesian : Only keep track posterior (with belief) and obtain expected regret. $O(\sqrt{T})$ is achieved.
  1. Ouyang (2019) - unverifiable set
  2. Abeille (2018, 2020) - 1D
  3. Kargin (2022) - extension to high dimensional space

- Let us define

$$\Theta := \begin{bmatrix}\Theta(1) & \cdots & \Theta(n)\end{bmatrix} := \begin{bmatrix}A & B\end{bmatrix}^\top,$$

with vectorization θ and $z_t := (x_t, u_t)$, hence,

$$x_{t+1} = \theta^\top z_t + w_t$$

- Subgaussian noise (Abbasi-Yadkori, 2011)

We know that optimal action is something like $u_t = Kx_t$. However if *bad* K is chosen, $x_t = (A + BK)^t x_0$ will blow up.

Feasible set

We assume that the unknown system parameter Θ_* is contained in

$$\mathcal{S} \subseteq \mathcal{S}_0 \cap \left\{ \Theta \in \mathbb{R}^{n\times(n+d)} \mid \operatorname{trace}(\Theta^\top \Theta) \leq S^2 \right\},$$

where

$$\mathcal{S}_0 = \Big\{ \Theta = (A, B) \in \mathbb{R}^{n\times(n+d)} \mid (A, B) \text{ is controllable,} \\ (A, M) \text{ is observable, where } Q = M^\top M \Big\}.$$

The condition implies (A, B) is stabilizable, i.e., there exists K such that

$$\rho(A + BK) < 1$$

Optimality and stability

When (A, B) is stabilizable,

- The Riccati equation has a unique positive semidefinite solution P, i.e.

$$P(\theta) = Q + A^\top P(\theta)A - A^\top P(\theta)B(R + B^\top P(\theta)B)^{-1}B^\top P(\theta)A.$$

- The gain matrix $K(\theta) := -(R + B^\top P(\theta)B)^{-1}B^\top P(\theta)A$ statbilizes the system parameter.
- The optimal cost is given by

$$J(\theta) = \mathrm{tr}(WP(\theta)),$$

whrere W is the covariance matrix for noise distribution

Construction of confidence sets

- Using the least square as before

$$e(\Theta) = \lambda \operatorname{trace}(\Theta^\top \Theta) + \sum_{s=0}^{t-1} \operatorname{trace}((x_{s+1} - \Theta^\top z_s)(x_{s+1} - \Theta^\top z_s)^\top).$$

  whose solution is given by

$$\hat{\Theta}_t = \operatorname*{argmin}_{\Theta} \; e(\Theta) = (Z^\top Z + \lambda I)^{-1} Z^\top X,$$

- Let $V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i^\top$ be the regularizaed design matrix underlying the covariates. Define

$$\beta_t(\delta) = \left(nL \sqrt{2 \log \left(\frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right)$$

Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\operatorname{trace}((\hat{\Theta}_t - \Theta_*)^\top V_t(\hat{\Theta}_t - \Theta_*)) \leq \beta_t(\delta).$$

In particular, $\mathbb{P}\left(\Theta_* \in \mathcal{C}_t(\delta),\, t = 1, 2, \dots\right) \geq 1 - \delta$, where

$$\mathcal{C}_t(\delta) = \left\{\Theta \in \mathbb{R}^{n \times (n+d)} : \operatorname{trace}\left\{(\Theta - \hat{\Theta}_t)^\top V_t(\Theta - \hat{\Theta}_t)\right\} \leq \beta_t(\delta)\right\}.$$

Now we choose optimal parameter as

$$J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) + \frac{1}{\sqrt{t}}$$

OFU algorithm

Inputs: $T, S > 0, \delta > 0, Q, L, \lambda > 0$.
Set $V_0 = \lambda I$ and $\hat{\Theta}_0 = 0$.
$(\tilde{A}_0, \tilde{B}_0) = \tilde{\Theta}_0 = \operatorname{argmin}_{\Theta \in \mathcal{C}_0(\delta) \cap S} J(\Theta)$.
for $t := 0, 1, 2, \ldots$ **do**
  if $\det(V_t) > 2 \det(V_0)$ **then**
    Calculate $\hat{\Theta}_t$ by (2).
    Find $\tilde{\Theta}_t$ such that $J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap S} J(\Theta) + \frac{1}{\sqrt{t}}$.
    Let $V_0 = V_t$.
  else
    $\tilde{\Theta}_t = \tilde{\Theta}_{t-1}$.
  end if
  Calculate u_t based on the current parameters, $u_t = K(\tilde{\Theta}_t) x_t$.
  Execute control, observe new state x_{t+1}.
  Save (z_t, x_{t+1}) into the dataset, where $z_t^\top = (x_t^\top, u_t^\top)$.
  $V_{t+1} := V_t + z_t z_t^\top$.
end for

Regret analysis

Theorem 2 *For any $0 < \delta < 1$, for any time T, with probability at least $1 - \delta$, the regret of Algorithm 1 is bounded as follows:*

$$R(T) = \tilde{O}\left(\sqrt{T \log(1/\delta)}\right) ,$$

where the constant hidden is a problem dependent constant.[2]

- Optimization is computationally unfavorable
- It is a frequentist regret (no expectation)
- $\log(1/\delta)$ is annoying !

- What is Thompson sampling : sample from the posterior distribution, choose an optimal action believing it is optimal
- Successful in many settings, bandit, MDP, ...
- Caveat is 'how to sample?'

Posterior update

Assume w_t follows Gaussian. Let $z_t := (x_t, u_t) \in \mathbb{R}^d$. Then, the system equation can be expressed as

$$x_{t+1} - \Theta^\top z_t = w_t \sim p_w,$$

which implies that

$$p(x_{t+1}|z_t, \theta) = p_w(x_{t+1} - \Theta^\top z_t | z_t, \theta),$$

The posterior at $(t+1)$-th time step is given by

$$\begin{aligned} p(\theta|h_{t+1}) &\propto p(x_{t+1}|z_t, \theta) p(\theta|h_t) \\ &= p_w(x_{t+1} - \Theta^\top z_t | z_t, \theta) p(\theta|h_t). \end{aligned}$$

Posterior update - Gaussian

'Posterior Sampling-based Reinforcement Learning for Control of Unknown Linear Systems' by Ouyang (2019)

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

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

Algorithm

Algorithm 1 PSRL-LQ

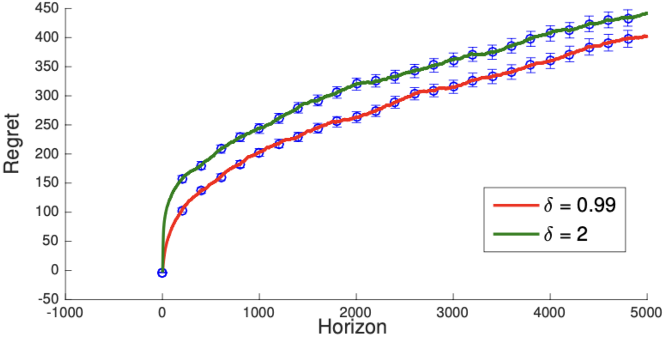
Input: $\Omega_1, \hat{\theta}_1, \Sigma_1$
Initialization: $t \leftarrow 1$, $t_k \leftarrow 0$
for episodes $k = 1, 2, \ldots$ **do**
  $T_{k-1} \leftarrow t - t_k$
  $t_k \leftarrow t$
  Generate $\tilde{\theta}_k \sim \mu_{t_k}$
  Compute $G_k = G(\tilde{\theta}_k)$ from (6)-(7)
  while $t \leq t_k + T_{k-1}$ and $\det(\Sigma_t) \geq 0.5 \det(\Sigma_{t_k})$ **do**
    Apply control $u_t = G_k x_t$
    Observe new state x_{t+1}
    Update μ_{t+1} according to (15)-(16)
    $t \leftarrow t + 1$

Regret

Theorem (Ouyang (2019))

The expected regret is upper bounded by

$$\sqrt{T}\log(T)$$



How do we choose the prior?

- Assume that there exists Ω_1 such that there exists $\delta < 1$ satisfying

$$\|A_* + B_* K(\theta)\| \leq \rho < 1$$

  for all $\theta \in \Omega$

- Stabilizatoin through random actions are discussed in two papers by M. Faradonbeh in series of works;
- Finite Time Adaptive Stabilization of Linear Systems (2019)
- On adaptive linear–quadratic regulators (2020)

More questions..

- Can we allow general class of admissible sets while obtaining the same or better regret?
- Can we deal with more general class of noises?

Langevin Monte Carlo

Consider the problem of sampling from a probability distribution with density $p(x) \propto e^{-U(x)}$, where the potential function $U : \mathbb{R}^{n_x} \to \mathbb{R}$ is continuously differentiable. The Langevin dynamics takes the form of

$$dX_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

Assumption

The potential U is m-convex and L-smooth, that is,

$$m \preceq \nabla^2 U \preceq L$$

In a continuous regime, the convergence is well-established.

- For a functional,
$$F : \rho \mapsto D_{KL}(\rho || e^{-U}),$$
$$\frac{\partial \rho_t}{\partial t} = -grad \quad F(\rho_t)$$
- Convergence is exponential.

Sampling via Unadjusted Langevin Algorithm

- For implementation, we need discretization in time.
- Apply the Euler-Maruyama discretization to the Langevin dynamics and obtain the following *unadjusted Langevin algorithm* (ULA):

$$X_{j+1} = X_j - \gamma_j \nabla U(X_j) + \sqrt{2\gamma_j} W_j,$$

  where $(W_j)_{j \geq 1}$ is an i.i.d. sequence of standard n_x-dimensional Gaussian random vectors, and $(\gamma_j)_{j \geq 1}$ is a sequence of step sizes.
- X_t can be used as a sample after enough iterations.

Convergence of ULA

Theorem

Suppose that pdf $p(x) \propto e^{-U(x)}$ is strongly log-concave and Lipschitz smooth with respect to x, i.e., $\lambda_{\min} \preceq \nabla^2 U(x) \preceq \lambda_{\max}$ for some $\lambda_{\max}, \lambda_{\min} > 0$. Let step size

$$\gamma_j \equiv \gamma = O\left(\frac{\lambda_{\min}(\nabla^2 U)}{\lambda_{\max}(\nabla^2 U)^2}\right),$$

and the number of iterations N

$$N = O\left(\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^2\right).$$

Given $X_0 = \arg\min U(x)$, let p_N denote the pdf of X_N. Then, the following inequality holds:

$$\mathbb{E}_{x\sim p, \tilde{x}\sim p_N}\left[|x - \tilde{x}|^2\right]^{\frac{1}{2}} \leq O\left(\sqrt{\frac{1}{\lambda_{\min}}}\right).$$

Bayesian update in our setting

- Relaxed assumptions on noises.

Assumption

For every $t = 1, 2, \ldots,$ the i.i.d. noise vector w_t satisfies the following properties:

1. The probability density function (pdf) of noise $p_w(\cdot)$ is known, smooth and twice differentiable. Additionally, the following inequalities hold:
$$\underline{m} I \preceq -\nabla^2_{w_t} \log p_w(w_t) \preceq \overline{m} I$$
$\underline{m}, \overline{m} > 0$;
2. $\mathbb{E}[w_t] = 0$ and $\mathbb{E}[w_t w_t^\top] = W$, where W is positive definite;

- Note the system equation can be expressed as
$$x_{t+1} - \Theta^\top z_t = w_t \sim p_w,$$
where $z_t := (x_t, u_t) \in \mathbb{R}^d$.
- Therefore,
$$\begin{aligned} p(\theta|h_{t+1}) &\propto p(x_{t+1}|z_t, \theta) p(\theta|h_t) \\ &= p_w(x_{t+1} - \Theta^\top z_t | z_t, \theta) p(\theta|h_t) \end{aligned}$$
preserves log-concavity.

Preconditioned ULA

By change of variable via

$$P_t := \lambda I_{dn} + \sum_{s=1}^{t-1} blkdiag\{z_s z_s^\top\}_{i=1}^n,$$

preconditioned ULA is defined as

$$\theta_{j+1} = \theta_j - \gamma_t P_t^{-1} \nabla U_t(\theta_j) + \sqrt{2\gamma P_t^{-1}} W_j,$$

for

$$\gamma_t := \frac{m\lambda_{\mathrm{min},t}}{16M^2 \max\{\lambda_{\mathrm{min},t}, t\}},$$

$$N_t := \frac{4\log_2(\max\{\lambda_{\mathrm{min},t}, t\}/\lambda_{\mathrm{min},t})}{m\gamma_t},$$

Preconditioned ULA vs Naive ULA

Lemma

For potential up to time t,

$$m \preceq P_t^{-\frac{1}{2}} \nabla^2 U_t(\theta) P_t^{-\frac{1}{2}} \preceq M,$$

where $m = \min\{\underline{m}, 1\}$, $M = \max\{\overline{m}, 1\}$, $P_t = \lambda I_{dn} + \sum_{s=1}^{t-1} \mathit{blkdiag}(\{z_s z_s^\top\}_{i=1}^n)$ *and the potential of the posterior* $U_t(\theta) = -\log p(\theta|h_t)$ *where* U_1 *satisfies* $\nabla_\theta^2 U_1(\cdot) = \lambda I_{dn}$ *for some* $\lambda > 0$.

- Stepsize

$$\frac{\lambda_{\min}}{\lambda_{\max}^2} \quad vs \quad \frac{m\lambda_{\min}}{16M^2 \max\{\lambda_{\min}, t\}}$$

- Step iteration

$$\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^2 \quad vs \quad \frac{4\log_2(\max\{\lambda_{\min,t}, t\}/\lambda_{\min,t})}{m\gamma},$$

Preconditioned ULA

By change of variable via

$$P_t := \lambda I_{dn} + \sum_{s=1}^{t-1} blkdiag\{z_s z_s^\top\}_{i=1}^n,$$

preconditioned ULA is defined as

$$\theta_{j+1} = \theta_j - \gamma P^{-1} \nabla U(\theta_j) + \sqrt{2\gamma P^{-1}} W_j,$$

Theorem

For any $t > 0$ and trajectory $(z_s)_{s \geq 1}$, the actual posterior μ_t and the approximate posterior $\tilde{\mu}_t$ obtained by preconditioned ULA satisfy

$$\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t} \left[|\theta_t - \tilde{\theta}_t|_{P_t}^p \mid h_t \right] \leq D_p,$$

where $D = 114 \frac{dn}{m}$ and $D_p = \left(\frac{pdn}{m} \right)^{\frac{p}{2}} \left(2^{2p+1} + 5^p \right)$ for $p \geq 2$. When $p = 2$, we further have

$$\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t} \left[|\theta_t - \tilde{\theta}_t|^2 \mid h_t \right]^{\frac{1}{2}} \leq \sqrt{\frac{D}{\max\{\lambda_{\min,t}, t\}}}.$$

Infusing noise for better exploration

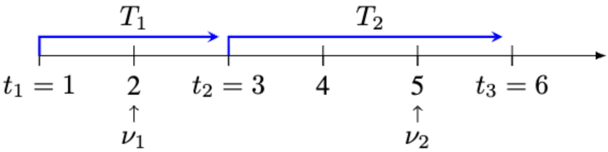
- Basically, we use

$$u_t = K_\theta x_t$$

- (Persistence of excitation)A key question how to we ensure that

$$\lambda_{\min}(U_t)$$

grows as t increases?
Our idea is to introduce noise injection.



- Noise injection

$$u_t = K_\theta x_t + \nu_t$$

Concentration of exact posterior μ_t

Proposition (Persistence of excitation)

Given $\lambda > 0$ and k sufficiently large,

$$\mathbb{E}\left[\frac{1}{\lambda_{\min,k+1}^p}\right] \le Ck^{-p}$$

for some global constant $C > 0$ where $\lambda_{\min,k+1}$ denotes the smallest eigenvalue of $\lambda I_d + \sum_{s=1}^{t_{k+1}-1} z_s z_s^\top$ where $(z_s)_{s\ge 1}$ is obtained via our main algorithm. In fact, $\lambda_{\min,k}$ is same as that of our preconditioner P_k.

Proposition

The true parameter θ_* and the exact posterior μ_t obtained by the main algorithm satisfies

$$\mathbb{E}[\mathbb{E}_{\theta_t\sim\mu_t}[|\theta_t - \theta_*|^p h_t]] \le C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p$$

for all $t \ge 1$ and $p > 0$.

Main result

We have the following result.

Theorem (K,Kim,Yang (2024))

The true parameter θ_* and the approximate posterior $\tilde{\mu}_t$ satisfy

$$\mathbb{E}\left[\mathbb{E}_{\tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\tilde{\theta}_t - \theta_*|^p | h_t\right]\right] \leq C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p$$

for any $p > 0$.

Skecth of proof

Assuming everything is nice.

Proof.

By Jensen's inequality,

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{\tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\tilde{\theta}_t - \theta_*|^p | h_t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\tilde{\theta}_t - \theta_*|^p | h_t\right]\right] \\
&\leq 2^{p-1}\mathbb{E}\left[\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\theta_t - \tilde{\theta}_t|^p | h_t\right]\right] + 2^{p-1}\mathbb{E}\left[\mathbb{E}_{\theta_t \sim \mu_t, \tilde{\theta}_t \sim \tilde{\mu}_t}\left[|\theta_t - \theta_*|^p | h_t\right]\right] \\
&\leq 2^{p-1}\mathbb{E}\left[\frac{D_p}{(\sqrt{\lambda_{\min,t}})^p}\right] + 2^{p-1}C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p \\
&\leq C\left(t^{-\frac{1}{4}}\sqrt{\log t}\right)^p.
\end{aligned}
$$

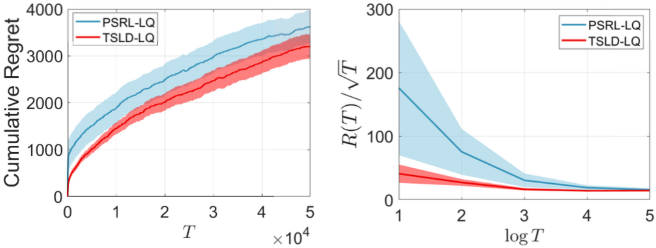
What we need is the concentration between exact posterior and true system parameter, μ_t and θ_*. □

Bayesian regret

The informal statement is..

Theorem (K, Kim, Yang (2024))

By applying fairly random action, we can construct tractable prior. Furthermore, the expected regret is given by $O(\sqrt{T})$



Bayesian regret

More results with different noises.

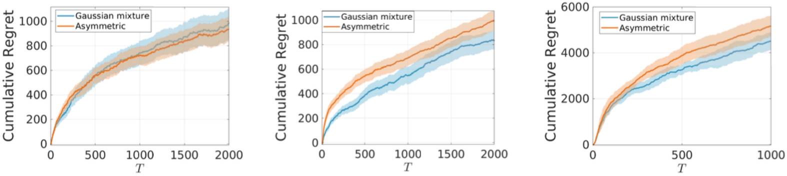


Figure: *3D (left) 5D (middle) 10D (right)*

Time horizon T	500	1000	1500	2000
Naive ULA	3.6×10^5	9.5×10^5	1.5×10^6	2.3×10^6
Preconditioned ULA	6.5×10^3	1.1×10^4	1.6×10^4	2.0×10^4

Figure: *Stepiterations*

Reference

- Abbasi-Yadkori, Yasin, and Csaba Szepesvári. "Regret bounds for the adaptive control of linear quadratic systems." Proceedings of the 24th Annual Conference on Learning Theory. JMLR Workshop and Conference Proceedings, 2011.
- Abeille, Marc, and Alessandro Lazaric. "Improved regret bounds for thompson sampling in linear quadratic control problems." International Conference on Machine Learning. PMLR, 2018.
- Faradonbeh, Mohamad Kazem Shirani, Ambuj Tewari, and George Michailidis. "On adaptive linear–quadratic regulators." Automatica 117 (2020): 108982.
- Faradonbeh, Mohamad Kazem Shirani, Ambuj Tewari, and George Michailidis. "Finite-time adaptive stabilization of linear systems." IEEE Transactions on Automatic Control 64.8 (2018): 3498-3505.