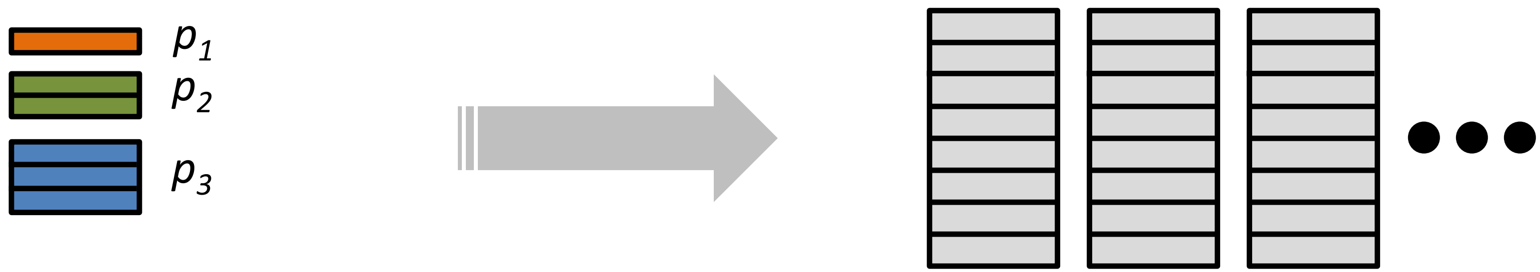
Online Control under Non-Stationarity: Dynamic Regret Minimization for the LQR system

Varun Gupta
University of Chicago

Joint with:
Yuwei Luo (Stanford GSB)
Mladen Kolar (Chicago Booth)

The Backstory – Online Stochastic Bin Packing



n i.i.d. items arrive from an unknown distribution

Infinite collection of bins of integer size B

Goal: Irrevocably pack items on arrival to minimize expected number of bins used

Main Result: A distribution-agnostic Primal-Dual (backpressure type) algorithm

+ Allows proving regret results for non-stationary arrivals

- Does not exploit nicer instances (e.g., *i.i.d.*) via intentional learning

Agenda: What is the "best" way to combine learning-based and agnostic control algorithms to get best-of-both-worlds guarantees?

* *Interior-Point-Based Online Stochastic Bin Packing*. Gupta, Radovanovic. OR 2020.

Outline

- Model and LQR preliminaries
- Stationary LQR learning and control algorithm
- The non-stationary LQR problem
  - Lower bound
  - Failure of static window heuristics
  - An adaptive restart algorithm
- A note on OLS estimator
- Next steps…

Stationary Linear Quadratic Regulator system

A discrete time, continuous space MDP

- State: $x_t \in \mathbb{R}^n$
- Action: $u_t \in \mathbb{R}^d$
- Parameter: $\Theta = [\mathrm{A\ B}]$
- Dynamics:

$$x_{t+1} = \mathrm{A} \cdot x_t + \mathrm{B} \cdot u_t + w_t$$

$$w_t \sim \mathcal{N}(0, W)$$
$$\psi \cdot I_d \preccurlyeq W \preccurlyeq \Psi \cdot I_d$$

- Cost:

$$c_t(x_t, u_t) = x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t$$

Goal: Minimize infinite horizon average cost

$$J = \lim_{T \to \infty} \frac{1}{T} \mathbf{E} \sum_{t=1}^{T} \left(x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t \right)$$

Stationary LQR preliminaries

Dynamics: $x_{t+1} = A \cdot x_t + B \cdot u_t + w_t$

Cost: $c_t(x_t, u_t) = x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t$

Linear feedback controllers: $u_t = K x_t$

Average cost $J(\Theta, K)$ and **bias function** $P(\Theta, K)$ satisfy the **Bellman recursion**:

$$x^T P x = x^T Q x + (Kx)^T R (Kx) - J + \mathbf{E}\left[x_1^T P x_1 | x_0 = x\right]$$

Which gives,

$$P = Q + K^T R K + (A + BK)^T P (A + BK),$$
$$J = \mathbf{E}[w^T P w] = Tr(P \cdot W).$$

Optimal controller: $K^*(\Theta) = \underset{K}{\operatorname{argmin}} J(\Theta, K)$

The non-stationary LQR learning and control problem

Finite horizon MDP

- Unknown parameter sequence: $\{\Theta_1, \Theta_2, \cdots, \Theta_T\}$; $\Theta_t = [\mathrm{A}_t \ \mathrm{B}_t]$
- Sublinear variation: V_T is $o(T)$,

$$V_T = \sum_{t=1}^{T-1} \|\Theta_{t+1} - \Theta_t\|_F$$

- Dynamics:

$$x_{t+1} = \mathrm{A}_t \cdot x_t + \mathrm{B}_t \cdot u_t + w_t$$

- Cost:

$$c_t(x_t, u_t) = x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t$$

Goal: Minimize finite horizon regret

$$R_T = \mathbf{E} \sum_{t=1}^{T} \left(x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t\right) - \min_{\pi} \mathbf{E} \sum_{t=1}^{T} \left(x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t\right)$$

Non-anticipative; knows $\{\Theta_t\}$

Remark: The optimal π is also a linear feedback controller $u_t = K_t^* x_t$

Short literature review

Stationary LQR with unknown dynamics -- $\mathcal{O}(\sqrt{T})$ regret is tight

Abbasi-Yadkori and Szepesvari (2011), Ibrahimi et al. (2012), Cohen et al. (2019), Faradonbeh et al. (2020), Mania et al. (2020), Simchowitz and Foster (2020), Cassel et al. (2020)

Learning and control of non-stationary MDPs with finite state and action spaces

Gajane et al. (2018), Cheung et al. (2020), Mao et al. (2021)

LQR with non-stationarity

- Hazan et al. (2020) : Known A, B but adversarial noise
- Simchowitz et a. (2020) : Disturbance feedback controller for adversarial noise
- Goel and Hassibi (2020) : Known A_t, B_t but adversarial noise
- Gradu et al. (2020) : Unknown A_t, B_t but observed after choosing u_t
- Lin et al. (2021) : Controller receives A_s, B_s, w_s for $s = t, \dots, t + k - 1$

Outline

- Model and LQR preliminaries
- Stationary LQR learning and control algorithm
- The non-stationary LQR problem
  - Lower bound
  - Failure of static window heuristics
  - An adaptive restart algorithm
- A note on OLS estimator
- Next steps…

A naïve exploration algorithm (Simchowitz and Foster, 2020)

Takeaway: LQR = bandit with linear feedback and quadratic loss

Linear feedback: Unknown $\Theta = \begin{bmatrix} A & B \end{bmatrix}$ but observe

$$x_{t+1} = \Theta \begin{bmatrix} x_t \\ u_t \end{bmatrix} + w_t$$

$$z_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}$$

A naïve exploration algorithm (Simchowitz and Foster, 2020)

Takeaway: LQR = bandit with linear feedback and quadratic loss

Linear feedback: Unknown $\Theta = [A \quad B]$ but observe

$$x_{t+1} = \Theta \cdot z_t + w_t$$

$$z_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}$$

Quadratic loss: True dynamics Θ, estimated dynamics $\widehat{\Theta}$, control $K = K^*(\widehat{\Theta})$

Theorem (Simchowitz, Foster): There exist constants C_1, C_2 such that

$$\left\|\Theta - \widehat{\Theta}\right\|_F \leq C_1 \implies J(\Theta, K) - J^*(\Theta) \leq C_2 \left\|\Theta - \widehat{\Theta}\right\|_F^2.$$

A naïve exploration algorithm (Simchowitz and Foster, 2020)

Idea: Create phases of doubling durations for exploration/exploitation

play K_i | play $K_{i+1} = K^*(\widehat{\Theta}_i)$ | play $K_{i+2} = K^*(\widehat{\Theta}_{i+1})$

phase i | phase $i+1$ | phase $i+2$ …

estimate $\widehat{\Theta}_i$ | estimate $\widehat{\Theta}_{i+1}$ | estimate $\widehat{\Theta}_{i+2}$

Estimate how? Ordinary Least Squares

$$\widehat{\Theta}_i = \underset{\widehat{\Theta}}{\operatorname{argmin}} \sum_{t \,\in\, \text{phase } i} \left\| x_{t+1} - \widehat{\Theta} \cdot z_t \right\|^2$$

Play how? With exploration noise. For $t \in$ phase i,

$$u_t = K_i \cdot x_t + \sigma_i \cdot \eta_t$$

$$\begin{cases} \eta_t \sim \mathcal{N}(0, I) \\ \sigma_i^2 \approx 1/\sqrt{2^i} \end{cases}$$

Intuition for $\boldsymbol{\sigma_i}$:

Total cost of exploration in phase $i \approx 2^i \cdot \sigma_i^2$

Variance of $\widehat{\Theta}_i \approx {}^{1}/_{2^i \cdot \sigma_i^2}$

Cost of estimation error $\approx {}^{2^i}/_{2^i \cdot \sigma_i^2} \approx 1/\sigma_i^2$

Balance these two

A randomized lower bound instance (Cassel et al., 2020)

- $n = d = 1$
- $a = 1/3$
- $b = \pm 1/T^{1/4}$ with equal probability

Theorem (Cassel et al., 2020): For T large enough, for any deterministic algorithm

$$E[R_T] = \Omega(\sqrt{T}).$$

Idea:

Cost of not learning $b = \Omega(\sqrt{T})$.

Cost of exploration noise needed to learn $b = \Omega(\sqrt{T})$.

Outline

- Model and LQR preliminaries
- Stationary LQR learning and control algorithm
- The non-stationary LQR problem
    - Lower bound
    - Failure of static window heuristics
    - An adaptive restart algorithm
- A note on OLS estimator
- Next steps…

A lower bound instance

Recall variation: $V_T = \sum_{t=1}^{T-1} \|\Theta_{t+1} - \Theta_t\|_F$

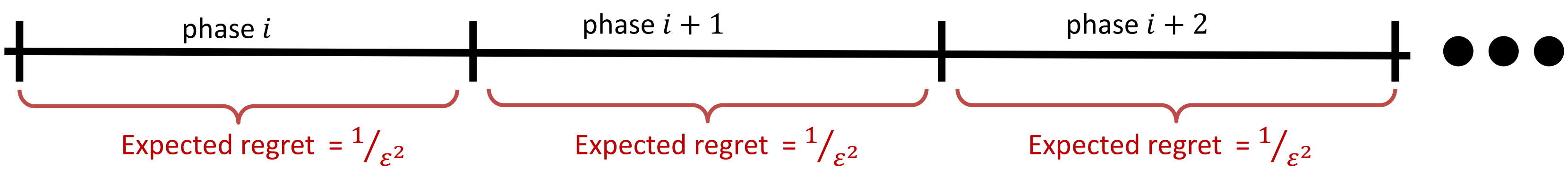
Theorem: For T large enough, $V_T = \widetilde{\Theta}(T^\alpha)$ for $\alpha \in (0,1)$, for any deterministic algorithm

$$E[R_T] = \Omega\left(V_T^{2/5} T^{3/5}\right).$$

Idea: Extend stationary LQR lower bound; define $\varepsilon = \left(\frac{V_T}{T}\right)^{1/5}$

$T\varepsilon^4$ phases of duration ${}^{1}/_{\varepsilon^4}$ each

b_t re-randomized at the beginning of each phase to be $\pm\varepsilon$



Total regret $\approx T\varepsilon^4 \cdot {}^{1}/_{\varepsilon^2} = T\varepsilon^2 = \Omega\left(V_T^{2/5} T^{3/5}\right).$

Towards an upper bound: Window-based algorithms

A common technique for non-stationary multi-armed bandits, linear bandits, and MDPs (e.g., *WindowUCB*, *WeightUCB*, *RestartUCB*,…)

- Fix a window size τ (with knowledge of V_T, or via "bandit-on-bandit")
- Restart the learning problem every τ time steps

Algorithm RestartLQR($\boldsymbol{\tau}, \boldsymbol{\sigma}$):

- Split horizon $[T]$ into non-overlapping phases of length τ
- Estimate $\widehat{\Theta}_i$ from phase i
- Action for $t \in \text{phase } (i+1)$: $u_t = K^*\left(\widehat{\Theta}_i\right)x_t + \sigma \cdot \eta_t$

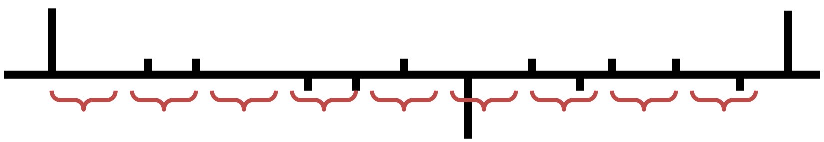
Theorem: There exists a randomized instance such that RestartLQR with optimally tuned τ, σ has $E[R_T] = \Omega\left(V_T^{1/3} T^{2/3}\right)$.

Towards an upper bound: Window-based algorithms

Theorem: There exists a randomized instance such that RestartLQR with optimally tuned τ, σ has $E[R_T] = \Omega\left(V_T^{1/3} T^{2/3}\right)$.

Instance: Again extend the 1-D LQR instance of Cassel et al. (2020)

$a = 1/3$; define $\varepsilon = \left(\frac{V_T}{T}\right)^{1/6}$



At time t:

- with probability $\frac{V_T}{2T}$: re-randomize $b_t = \pm 1$
- with probability $\left(\frac{V_T}{4T}\right)^{5/6}$: re-randomize $b_t = \pm \varepsilon$
- otherwise $b_t = b_{t-1}$

Idea: If there were only $\pm\varepsilon$ changes, algorithm would pick $\tau = \mathcal{O}\left(\frac{T}{V_T}\right)^{5/6}$

But, if a ± 1 change lands inside a τ-phase, we pay an $\Omega(\tau)$ regret

This forces the algorithm to pick $\tau = \mathcal{O}\left(\frac{T}{V_T}\right)^{2/3}$

An Adaptive restart algorithm

Intuition: Large changes in dynamics should be easy to detect

Instead of committing to a window size, we should restart when we detect a large change

Since we do not know how "large" the change might be, we simultaneously explore to detect changes at multiple scales*

Assumption: The learner/controller is given a sequence of (potentially suboptimal) sequentially-strong stabilizing controllers $\{K_t^{\text{stab}}\}$.

Implies that for some $0 < \gamma < 1$, for any interval $[\tau_1, \tau_2]$:

$$\left\| \prod_{t=\tau_1}^{\tau_2} (A_t + B_t K_t^{\text{stab}}) \right\| \sim \gamma^{\tau_2 - \tau_1}.$$

* *A new algorithm for non-stationary contextual bandits*. Chen et al. COLT 2019.

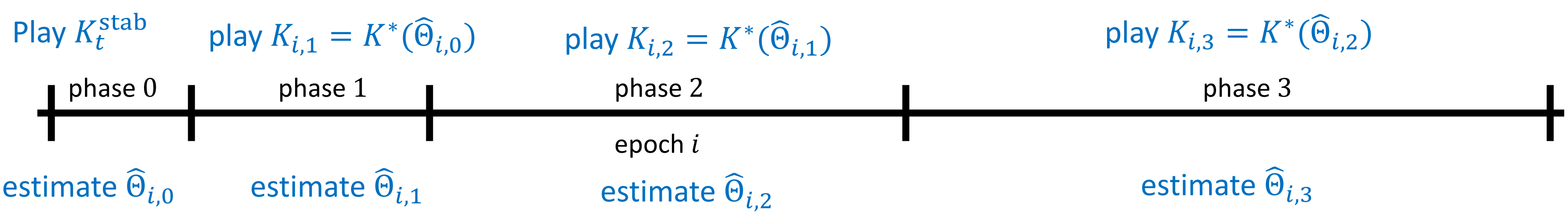
An Adaptive restart algorithm

Idea 1: Split the horizon into *epochs*

$$(\text{variation within epoch})^2 \approx 1/\sqrt{\text{duration of epoch}}$$

Idea 2: Since we apriori do not know the length of the epoch, we use phases of doubling durations

- phase 0: Play $u_t = K_t^{\text{stab}} \cdot x_t + \sigma_0 \cdot \eta_t$ $\qquad \sigma_j^2 \approx 1/\sqrt{2^j}$
- use OLS to estimate $\widehat{\Theta}_{i,0}$, $K_{i,1} = K^*\big(\widehat{\Theta}_{i,0}\big)$
- phase 1: Play $u_t = K_{i,1} \cdot x_t + \sigma_1 \cdot \eta_t$
- …
- phase j: Play $u_t = K_{i,j} \cdot x_t + \sigma_j \cdot \eta_t$



An Adaptive restart algorithm (contd.)

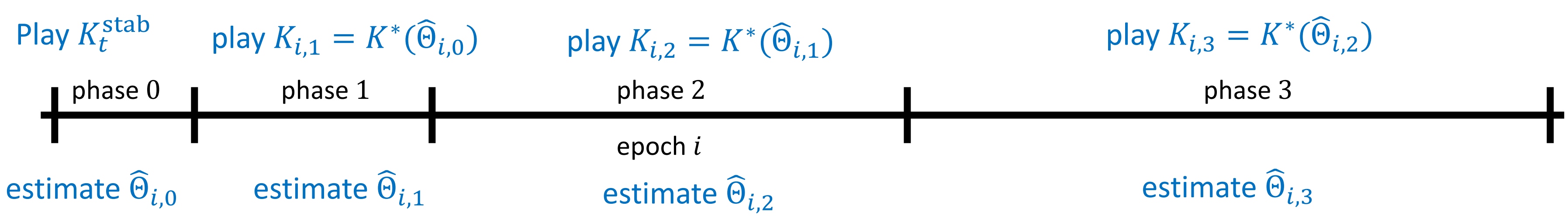
Idea 3: End the epoch at the end of phase j if

$$\left\|\widehat{\Theta}_{i,j} - \widehat{\Theta}_{i,j-1}\right\|_F^2 \gtrsim {}^1\!/\!_{\sqrt{2^j}}$$

Idea 4: At each t in phase j begin a scale m detection test for $m \in \{0,1,..,j-1\}$ with probability ${}^1\!/\!_{\sqrt{2^{j+m}}}$

- For the next 2^m time steps increase the exploration noise to σ_m
- Estimate $\widehat{\Theta}_{i,j,m}$
- End the epoch at the end of detection test if

$$\left\|\widehat{\Theta}_{i,j,m} - \widehat{\Theta}_{i,j-1}\right\|_F^2 \gtrsim {}^1\!/\!_{\sqrt{2^m}}$$



Proof sketch

1. It suffices to compare with $\sum_t J^*(\Theta_t)$

$$\min_\pi \mathbf{E} \sum_{t=1}^{T} \left(x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t\right) \ \geq \sum_t J^*(\Theta_t) - \tilde{O}(V_T + 1)$$

$$\approx \sum_t \mathbf{E}\left\|\Theta_t - \widehat{\Theta}_t\right\|_F^2$$

2. Regret decomposition, for policy $K_t = K^*\left(\widehat{\Theta}_t\right)$

$$\mathbf{E} \sum_{t=1}^{T} \left(x_t^T \mathrm{Q} x_t + u_t^T \mathrm{R} u_t\right) - \sum_t J^*(\Theta_t) = \sum_t \mathbf{E}\left[J(\Theta_t, K_t) - J^*(\Theta_t)\right]$$

$$+ \ (\text{Exploration cost})$$

$$+ \sum_t \mathbf{E}\left[x_t^T \left(P(\Theta_t, K_t) - P(\Theta_{t-1}, K_{t-1})\right) x_t\right]$$

$$\lesssim V_T + (\text{\# policy changes})$$

Proof sketch (contd.)

3. Regret for epoch i of duration E_i

Square variation: $\Delta_i^2 \approx {}^{1}/_{\sqrt{E_i}}$

Regret $\approx E_i \cdot \Delta_i^2$

$(\sum_i \Delta_i = V_T)$ + $(\sum_i E_i = T)$ + Hölder's inequality $\Longrightarrow E[R_T] = \tilde{O}\left(V_T^{2/5} T^{3/5}\right)$

Outline

- Model and LQR preliminaries
- Stationary LQR learning and control algorithm
- The non-stationary LQR problem
  - Lower bound
  - Failure of static window heuristics
  - An adaptive restart algorithm
- A note on OLS estimator
- Next steps…

OLS estimator – stationary case

$$\widehat{\Theta}^* = \underset{\widehat{\Theta}}{\operatorname{argmin}}\, \mathcal{L}(\widehat{\Theta})$$

Where: $\mathcal{L}(\widehat{\Theta}) = \sum_t \left\| x_{t+1} - \widehat{\Theta} \cdot z_t \right\|^2 = \sum_t \left\| \Theta \cdot z_t + w_t - \widehat{\Theta} \cdot z_t \right\|^2$

Solution: $$\widehat{\Theta}\left(\sum z_t z_t^T\right) = \underbrace{\left(\sum \Theta z_t z_t^T\right)}_{\text{mean}} + \underbrace{\sum w_t z_t^T}_{\text{variance}}$$

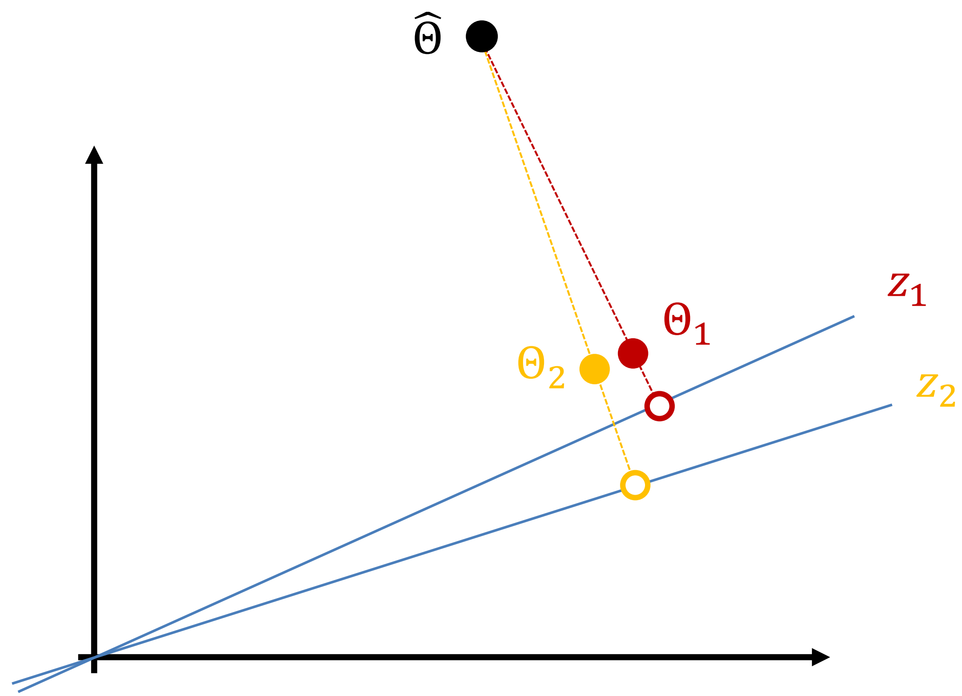
The OLS estimator is unbiased under mild conditions.

OLS estimator – non-stationary case

Solution: $$\widehat{\Theta}\left(\sum z_t z_t^T\right) = \underbrace{\left(\sum \Theta_t z_t z_t^T\right)}_{\text{mean}} + \underbrace{\sum w_t z_t^T}_{\text{variance}}$$

The OLS estimator could have a large bias even if all Θ_t are close

Pictorially:



OLS estimator – non-stationary case

We prove that the OLS estimator has small bias "from scratch"

Given the OLS loss function

$$\mathcal{L}(\widehat{\Theta}) = \sum_t \left\|\Theta_t \cdot z_t + w_t - \widehat{\Theta} \cdot z_t\right\|^2,$$

fix a representative $\overline{\Theta}$ and direction v, and construct the one dimensional loss function

$$\mathcal{L}_v(\lambda) = \mathcal{L}\left(\overline{\Theta} + \lambda \cdot v\right).$$

Finally, we show that for enough directions v, the minimizer $|\lambda_v^*|$ is small with high probability $\Longrightarrow$ $\widehat{\Theta}^*$ is close to $\overline{\Theta}$

Key idea: The function $\mathcal{L}_v(\lambda)$ looks very different for v lying in the space spanned by $\begin{bmatrix} I \\ K \end{bmatrix} x$ for $x \in \mathbb{R}^n$ and in its orthogonal subspace

Outline

- Model and LQR preliminaries
- Stationary LQR learning and control algorithm
- The non-stationary LQR problem
  - Lower bound
  - Failure of static window heuristics
  - An adaptive restart algorithm
- A note on OLS estimator
- Next steps…

For non-stationary LQR

The assumption that dynamics are non-stationary but the noise covariance is known and stationary seems unrealistic

Q1: Learning and control of unknown non-stationary dynamics in the present of non-stochastic noise?

What does $V_T = o(T)$ mean in practice?

Q2: With $V_T = \varepsilon T$, for $\varepsilon \leq \varepsilon_0$, $\mathbf{E}[R_T] \sim \varepsilon^{2/5} T$?

Summarizing the hardness of the instance in a single number V_T seems unsatisfactory.

Q3: An instance-optimal notion of regret?

Q4. Model free non-stationary LQR? (Gradu et al. do this but under the assumption that dynamics are observed after each time step)

Q5. A sliding adaptive-window size algorithm which avoids hard restarts?

For non-stationary control more broadly..

Q6: How should we combine learning and agnostic/back-pressure type policies?

- Results of Neely, Huang
- See forthcoming survey/tutorial by Neil Walton and Kuang Xu

Q7: Combining noisy forecasts with learning and robust policies?

Q8: Combining NN policy approximation with non-stationarity – a meta-Reinforcement Learning approach

References

- Y. Abbasi-Yadkori and C. Szepesv́ ari. Regret bounds for the adaptive control of linear quadratic systems. ICML, 2011.
- A. Cassel, A. Cohen, and T. Koren. Logarithmic regret for learning linear quadratic regulators efficiently. ICML, 2020.
- Y. Chen, C.-W. Lee, H. Luo, and C.-Y. Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. COLT, 2019.
- W. C. Cheung, D. Simchi-Levi, and R. Zhu. Non-stationary reinforcement learning: The blessing of (more) optimism. 2020.
- A. Cohen, T. Koren, and Y. Mansour. Learning linear-quadratic regulators efficiently with only √T regret, 2019.
- M. K. S. Faradonbeh, A. Tewari, and G. Michailidis. Input perturbations for adaptive control and learning. Automatica, 2020.
- P. Gajane, R. Ortner, and P. Auer. A sliding-window algorithm for Markov decision processes with arbitrarily changing rewards and transitions. arXiv:1805.10066, 2018
- G. Goel and B. Hassibi. Regret-optimal control in dynamic environments. arXiv:2010.10473, 2020
- P. Gradu, E. Hazan, and E. Minasyan. Adaptive regret for control of time-varying dynamics. arXiv:2007.04393, 2020
- E. Hazan, S. Kakade, and K. Singh. The nonstochastic control problem. ALT, 2020.
- M. Ibrahimi, A. Javanmard, and B. V. Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. NeurIPS, 2012
- H. Mania, S. Tu, and B. Recht. Certainty equivalence is efficient for linear quadratic control. arXiv:1902.07826, 2019.
- M. Simchowitz and D. Foster. Naive exploration is optimal for online LQR. ICML, 2020
- M. Simchowitz, K. Singh, and E. Hazan. Improper learning for non-stochastic control. ICML, 2020.