UCSB

Optimal Dynamic Regret in LQR Control

Dheeraj Baby and Yu-Xiang Wang
dheeraj@ucsb.edu and yuxiangw@cs.ucsb.edu

Problem Setting

◇ Online interaction protocol

We study an online LQR system:

1. At time $t \in [n]$, learner is at state $x_t \in R^{d_x}$ and plays a control signal $u_t \in R^{d_u}$.
2. The system transitions to next state as: $x_{t+1} = Ax_t + Bu_t + w_t$ with $\|w_t\|_2 \leq 1$. A and B are known.
3. Agent suffers loss $\ell(x_t, u_t) = \|x_t\|^2_{R_x} + \|u_t\|^2_{R_u}$ for known R_x, R_u

Definition 1 *(Disturbance action policies, Foster and Simchowitz, 2020). Let $M = (M^{[i]})_{i=1}^m$ denote a sequence of matrices $M^{[i]} \in \mathbb{R}^{d_u \times d_x}$. We define the corresponding disturbance action policies (DAP) π^M as:*

$$\pi_t^M(x_t) = -K_\infty x_t - q^M(w_{1:t-1}),$$

where $q^M(w_{1:t-1}) = \sum_{i=1}^m M^{[i]} w_{t-i}$. We are interested in DAPs for which the sequence M belongs to the set:

$$\mathcal{M}(m, R, \gamma) := \{M = (M^{[i]})_{i=1}^m : \|M^{[i]}\|_{op} \leq R\gamma^{i-1}\},$$

where m, R and γ are known parameters.

The performance of the learner is measured in terms of **dynamic policy regret**:

$$R(M_{1:n}) = \sum_{t=1}^n \ell(x_t^{\text{alg}}, u_t^{\text{alg}}) - \ell(x_t^{M_{1:n}}, u_t^{M_{1:n}})$$

◇ Responsible decision making

1. Physical constraints on the allowable control actions at any state.
2. Eg: Applying huge torque in a drone can burn the motor or drain the battery quickly.
3. We model safe control signals as: $\mathcal{F}_t := \{u_t | u_t = \pi_t^M(x_t) \text{ for some } M \in \mathcal{M}(m, R, \gamma)\}$.
4. Choosing parameters m, R and γ can constrain the magnitude of feasible control actions at any state.
5. To ensure safety, at each round the learner plays a control signal from the feasible set $\mathcal{F}_t$ thus necessitating the need for proper learning.

Reduction to OCO

Foster and Simchowitz, 2020 provides a reduction from the LQR problem to the problem of delayed online linear regression:

Proposition 1 *Suppose the learner plays policy of the form $\pi_t^{alg}(x) = -K_\infty x + q^{M_t^{alg}}(w_{1:t-1})$. Let the comparator policies take the form $\pi_t(x) = -K_\infty x + q^{M_t}(w_{1:t-1})$ for a sequence of matrices $M_{1:n}$ chosen in hindsight. Then the dynamic regret against the policies $\pi := (\pi_1, \ldots, \pi_n)$ satisfies:*

$$R_n(\pi) \leq O(1) + \sum_{t=1}^n \hat{A}_t(M_t^{alg}) - \hat{A}_t(M_t),$$

where the parameters involved in the inequality are defined as below: $\hat{A}_t(M) := \|q^M(w_{1:t-1}) - q_{\infty;h}(w_{t:t+h})\|^2_{\Sigma_\infty}$ and $h = O(\log n)$.

1. The losses $A_t(M_t)$ are linear regression losses which are exp-concave.
2. Need dynamic regret minimizing algorithms under exp-concave losses.
3. To ensure safety, we must play matrices $M_t^{\text{alg}} \in \mathcal{M}(m, R, \gamma)$.

The algorithm of Baby and Wang, 2022 minimises dynamic regret under exp-concave losses. But they can only support L_∞ constrained decision sets.

Central question: How to extend their algorithm for proper online linear regression?

Algorithm

ProDR.control: Inputs - Decision set $\mathcal{D}$, $G > 0$, a surrogate algorithm $\mathcal{A}$ which ensures low dynamic regret under general exp-concave losses against any comparator sequence in some $\mathcal{D}' \supset \mathcal{D}$. Here $\mathcal{D}'$ is a compact and convex set. Note that such an algorithm $\mathcal{A}$ may produce iterates outside $\mathcal{D}$. (See Theorem 1 for a specific choice of $\mathcal{A}$.)

1. At round t, receive w_t from $\mathcal{A}$.
2. Receive co-variate matrix $A_t := [a_{t,1}, \ldots, a_{t,p}]^T$.
3. Play $\hat{w}_t \in \text{argmin}_{x \in \mathcal{D}} \max_{i=1,\ldots,p} |a_{t,i}^T(x - w_t)|$.
4. Let $\ell_t(w) = f_t(w) + G \cdot S_t(w)$, where $f_t(w) = \|A_t w - b_t\|_2^2$ and $S_t(w) = \min_{x \in \mathcal{D}} \max_{i=1,\ldots,p} |a_{t,i}^T(x - w)|$.
5. Send $\ell_t(w)$ to $\mathcal{A}$.

Performance Guarantees

Theorem 1 (informal) *Set $\mathcal{D} = \mathcal{M}(m, R, \gamma)$ and $\mathcal{D}'$ as the tightest L_∞ ball that encloses $\mathcal{D}$. Choosing the surrogate algorithm $\mathcal{A}$ as the algorithm in Baby and Wang, 2022 and using the delayed to non-delayed reduction of Joulani et al, 2013 guarantees that $R(M_{1:n}) = \tilde{O}^*(n^{1/3}[\mathcal{TV}(M_{1:n})]^{2/3} \vee 1)$. Here $\mathcal{TV}(M_{1:n}) := \sum_{t=2}^n \sum_{i=1}^m \|M_t^{[i]} - M_{t-1}^{[i]}\|_1$. Further, the static regret against any DAP policy from $\mathcal{M}(m, R, \gamma)$ in any local time window is $O(\log n)$.*

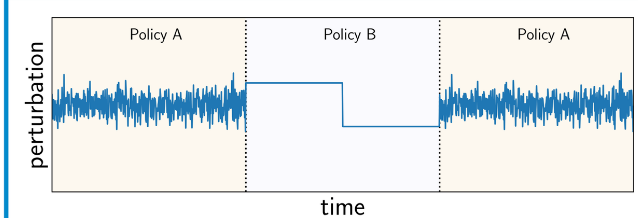
Design of ProDR.control is inspired by the improper to proper black-box reduction of Cutkosky and Orabona, 2018.

Theorem 2 *There exists an LQR system, a choice of the perturbations w_t and a DAP policy class such that:*

$$\sup_{M_{1:n} \text{ with } \mathcal{TV}(M_{1:n}) \leq C_n} \mathbb{E}[R(M_{1:n})] = \Omega(n^{1/3} C_n^{2/3} \vee 1),$$

where the expectation is taken wrt randomness in the strategies of the agent and adversary.

Example of Non-Stationarity



Depending on the perturbation process, different DAP policies are suitable across different sections of time.

References

Online learning under delayed feedback, Pooria Joulani, András György, and Csaba Szepesvari, ICML 2013

Black-box reductions for parameter-free online learning in banach spaces, Ashok Cutkosky and Francesco Orabona, COLT 2018.

Logarithmic regret for adversarial online control, Dylan J Foster and Max Simchowitz, ICML 2020.

Optimal dynamic regret in proper online learning with strongly convex losses and beyond, Dheeraj Baby and Yu-Xiang Wang, AISTATS, 2022.