Optimal Dynamic Regret in LQR Control

Dheeraj Baby [1] **Yu-Xiang Wang** [1]

Abstract

We consider the problem of nonstochastic control with a sequence of quadratic losses, i.e., LQR control. We provide an efficient online algorithm that achieves an optimal dynamic (policy) regret of $\tilde{O}(n^{1/3}\mathcal{TV}(M_{1:n}^{2/3} \vee 1)$, where $\mathcal{TV}(M_{1:n})$ is the total variation of any oracle sequence of *Disturbance Action* policies parameterized by $M_1, ..., M_n$ — chosen in hindsight to cater to unknown nonstationarity. The rate improves the best known rate of $\tilde{O}(\sqrt{n(\mathcal{TV}(M_{1:n})+1)})$ for general convex losses and is information-theoretically optimal for LQR. Main technical components include the reduction of LQR to online linear regression with delayed feedback due to Foster and Simchowitz (2020), as well as a new *proper* learning algorithm with an optimal $\tilde{O}(n^{1/3})$ dynamic regret on a family of "minibatched" quadratic losses, which could be of independent interest.

1. Introduction

This paper studies the linear quadratic regulator (LQR) control problem which is a specific instantiation of the more general RL framework where the evolution of states follows a predefined linear dynamics. At each round $t \in [n] := \{1, \ldots, n\}$, the agent is at state $x_t \in \mathbb{R}^{d_x}$. Based on the state, the agent select a control input $u_t \in \mathbb{R}^{d_u}$. The next state evolves according to the law:

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

where A and B are system matrices known to the agent. $w_t \in \mathbb{R}^{d_x}$ is a disturbance term that can be selected by a potentially adaptive adversary. We assume that $\|w_t\|_2 \leq 1$. This disturbance term reflects the perturbation from the ideal linear state transition arising due to environmental factors that could be difficult to model. The loss suffered by playing the control u at state x is given by $\ell(x, u) := x^T R_x x + u^T R_u u$, where $R_x, R_u \succcurlyeq 0$, that are apriori fixed and known.

Recently there has been a surge of interest in viewing this classical LQR problem under the lens of online learning (Hazan, 2016). The work of Agarwal et al. (2019) places regret of the agent against a set of benchmark policies as the central notion to evaluate learner's performance. Following Agarwal et al. (2019); Foster and Simchowitz (2020) we adopt the class of disturbance action policies (DAP) as our benchmark class:

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:*

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

where $q^M(w_{1:t-1}) = \sum_{i=1}^m M^{[i]} w_{t-i}$ and K_∞ as in Eq.(5). We are interested in DAPs for which the sequence M belongs to the set:

$$\mathcal{M}(m, R, \gamma) := \{M = (M^{[i]})_{i=1}^m : \|M^{[i]}\|_{op} \leq R\gamma^{i-1}\}, \tag{2}$$

where m, R and γ are algorithm parameters.

This class is known to be sufficiently rich to approximate many linear controllers. A policy takes in the past history and current state as input and produces a control signal as output. Let's denote $M_{1:n} := (M_1, \ldots, M_n)$ to be a sequence of DAP policies such that at time t, the control signal is selected using the policy parameterized by M_t (see Eq.(1)). We denote $x_t^{M_{1:n}}$ to be the state reached at round t by playing the sequence of policies defined by parameters $M_{1:t-1}$ in the past. Similarly $u_t^{M_{1:n}}$ is used to denote the control signal produced by the policy M_t. The *universal dynamic (policy) regret* of the learner against the policy sequence $M_{1:n}$ is defined as:

$$R(M_{1:n}) = \sum_{t=1}^n \ell(x_t^{\text{alg}}, u_t^{\text{alg}}) - \ell(x_t^{M_{1:n}}, u_t^{M_{1:n}}), \tag{3}$$

where $(x_t^{\text{alg}}, u_t^{\text{alg}})$ denotes the state and control signal of the learner at round t. Note that the policy sequence $M_{1:n}$ can

[*] Equal contribution [1] Department of Computer Science, University of California, Santa Barbara. Correspondence to: Dheeraj Baby <dheeraj@ucsb.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

be *any* valid sequence of DAP polices. The main focus of this paper is to design algorithms that can control the dynamic regret against a sequence of reference policies as a function of the time horizon n and the a path variation of the DAP parameters of the comparator $M_{1:n}$. We remark that the comparator polices $M_{1:n}$ can be chosen in hindsight and potentially unknown to the learner.

Whenever $M_{1:n} = (M, \dots, M)$ for a fixed parameter M, we recover the notion of static regret. However the notion of static regret is not befitting for non-stationary environments. For example consider the scenario of controlling a drone. Suppose during the initial half of the trajectory there is heavy wind eastwards and in the second half, wind blows westwards. For best performance, a controller has to choose different policies that can counter-act the wind and guide the motion properly in each half. Hence, we aim to control the dynamic regret which allows us to be competent against a sequence of potentially time-varying polices chosen in hindsight. We remark that our algorithm automatically adapts to the level of non-stationarity in the hindsight sequence of policies.

Next, we take a digression and discuss a desirable property for the design of algorithms for LQR control.

Proper learning in LQR control. Proper learning is an online learning paradigm where the decisions of the learner are required to obey some user specified physical limits. On the other hand, improper learning framework allows the learner to disregard such constraints. The paradigm of improper learning may not be attractive in certain applications where safety is a paramount concern. Improper algorithms can possibly take the system through trajectories that are deemed to be risky. It is desirable to avoid such behaviours in physical systems such as self driving cars, control of medical ventilators, robotic control (Levine et al., 2016) and cooling data centers (Cohen et al., 2018). A policy selects a control signal u_t depending on the current state x_t. Given the value of current state, there can be physical constraints on the allowable control actions. For example, imagine the situation where we want to maintain the velocity of a drone. Depending on the current position and other system and environmental factors (the state), one can only apply a range of allowable torque (the control action) to the blades. Not respecting this torque range can drain the battery quickly or can lead to catastrophic damages such as burnt rotors. In our framework, we model this set of allowable control actions as $\mathcal{F}_t := \{u_t | u_t = \pi_t^M(x_t) \text{ for some } M \in \mathcal{M}(m, R, \gamma)\}$ (see Definition 1). By appropriately choosing the values of m, R and γ, one can ensure desirable norm constraints on the state and control signal. To ensure safety (and responsibility), at each round the learner plays a control signal from the feasible set $\mathcal{F}_t$ thus motivating the need of proper learning.

Below are our contributions:

- We develop an optimal universal dynamic regret minimization algorithm for the general mini-batch linear regression problem (see Theorem 2).
- Applying the reduction of Foster and Simchowitz (2020) from LQR problem to online linear regression, the above result lends itself to an algorithm for controlling the dynamic regret of the LQR problem (Eq.(3)) to be $\tilde{O}^*(n^{1/3}[\mathcal{TV}(M_{1:n})]^{2/3})$, where $\mathcal{TV}$ denotes the total variation incurred by the sequence of DAP policy parameters in hindsight (see Corollary 4). O^* hides the dependencies in dimensions and system parameters.
- We show that the aforementioned dynamic regret guarantee is minimax optimal modulo dimensions and factors of $\log n$ (see Theorem 5).
- The resulting algorithm is also strongly adaptive, in the sense that the static regret against a DAP policy in any local time window is $O^*(\log n)$.

Notes on novelty and impact. As discussed before, the reduction of Foster and Simchowitz (2020) casts LQR problem to an instance of proper online linear regression. In the context of regression, proper learning means that the decisions of the learner belongs to a user specified convex domain. The main challenge in developing aforementioned contributions rests on the design of an optimal universal dynamic regret minimization algorithm for online linear regression under the setting of *proper learning*. We are not aware of any such algorithms in the literature to-date and the problem remains open. However, there exists an improper algorithm from Baby and Wang (2021) for controlling the desired dynamic regret. Given this fact, the design of our algorithm is facilitated by coming up with *new* black-box reductions (see Section 2) that can convert an improper algorithm for non-stationary online linear regression to a proper one. There are improper to proper black-box reduction schemes given in the influential work of Cutkosky and Orabona (2018). However, they are developed to support general convex or strongly convex (see Definition 11) losses. The linear regression losses arising in our setting are exp-concave (see Definition 10) which enjoy strong curvature only in the direction of the gradients as opposed to uniformly curved strongly convex losses. Hence the reduction scheme of Cutkosky and Orabona (2018) is inadequate to provide fast regret rates in our setting. In contrast, we develop novel reduction schemes that carefully take the non-uniform curvature of the linear regression losses into account so as to facilitate fast dynamic regret rates (see Section 2.2). The construction of this new reduction scheme requires non-trivial adaptation of the ideas in Cutkosky and Orabona

(2018). We remark that the algorithm ProDR.control developed in Section 2 can be impactful in general online learning literature. That the non-stationary LQR problem can be *optimally* solved using ProDR.control is a testament to this fact. Further our algorithm is out-of-the-box applicable to more general settings such as non-stationary multi-task linear regression, which is beyond the current scope.

2. Non-stationary "mini-batch" linear regression

We refer the reader to Appendix B for a brief overview of the results of Foster and Simchowitz (2020) who reduces the LQR control problem to proper online linear regression. So we take a digression in this section and study the problem of controlling dynamic regret in a general linear regression setting.

2.1. Linear regression framework

Consider the following linear regression protocol.

- At round t, nature reveals a co-variate matrix $A_t \in \mathbb{R}^{p \times d}$.
- Learner plays $z_t \in \mathcal{D} \subset \mathbb{R}^d$.
- Nature reveals the loss $f_t(z) = \|A_t z - b_t\|_2^2$.

Under the above regression framework, we are interested in controlling the universal dynamic regret against an arbitrary sequence of predictors $u_1, \dots, u_n \in \mathcal{D}$ (abbreviated as $u_{1:n}$) :

$$R_n(u_{1:n}) = \sum_{t=1}^{n} f_t(z_t) - f_t(u_t). \tag{4}$$

Dynamic regret is usually expressed as a function of n and a path variational that captures the smoothness of the comparator sequence. We will focus on the path variational defined by:

$$\mathcal{TV}(u_{1:n}) = \sum_{t=2}^{n} \|u_t - u_{t-1}\|_1.$$

Below are the list of assumptions made:

Assumption 1. Let $a_{t,i} \in \mathbb{R}^d$ be the i^{th} row vector of A_t. We assume that $\|a_{t,i}\|_1 \leq \alpha$ for all $t \in [n]$ and $i \in [p]$. Further $\|b_t\|_1 \leq \sigma$ for all t.

Assumption 2. For any $x \in \mathcal{D}$, $\|x\|_1 \leq \chi$ and $\|x\|_\infty \leq \tilde{R}$.

We refer this setting as mini-batch linear regression since the loss at round t can be written as a sum of a batch of quadratic losses: $f_t(z) = \sum_{i=1}^{p} \left(z^T a_{t,i} - b_t[i]\right)^2$.

```
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 2 for a specific choice of $\mathcal{A}$.)

  1. At round $t$, receive $w_t$ from $\mathcal{A}$.
  2. Receive co-variate matrix $A_t := [a_{t,1}, \dots, a_{t,p}]^T$.
  3. Play $\hat{w}_t \in \mathrm{argmin}_{x \in \mathcal{D}} \max_{i=1,\dots,p} |a_{t,i}^T(x - w_t)|$.
  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 $S_t(w) = \min_{x \in \mathcal{D}} \max_{i=1,\dots,p} |a_{t,i}^T(x - w)|$.
  5. Send $\ell_t(w)$ to $\mathcal{A}$.
```

Figure 1: ProDR.control: An algorithm for non-stationary and proper linear regression.

2.2. The Algorithm

Due to space constraints, we explain the intuition behind our algorithm ProDR.control (Fig.1) in Appendix E.

2.3. Main Results

We have the following guarantee for ProDR.control:

Theorem 2. *Let $u_{1:n} \in \mathcal{D}$ be any comparator sequence. In Fig.1, choose G such that $\sup_{w_1, w_2 \in \mathcal{D}_\infty(\tilde{R}), t \in [n]} \|A_t(w_1 + w_2) - 2b_t\|_1 \leq G$. Let α be as in Assumption 2. Let L be such that $\sup_{w \in \mathcal{D}_\infty(\tilde{R}), j \in [p]} 2\|A_t w - b_t\|_2^2 + 2G^2 \leq L$ for all $t \in [n]$. Choose $\mathcal{A}$ as the algorithm from Baby and Wang (2022) (see Appendix C) with parameters $\gamma = 2G\alpha\tilde{R}\sqrt{d/8L} + \sqrt{2L}$ and $\zeta = \min\{\frac{1}{16G\alpha\tilde{R}\sqrt{d}}, 1/(4\gamma^2)\}$ and decision set $\mathcal{D}_\infty(\tilde{R})$. Under Assumptions 1 and 2, a valid of assignment of G and L are $2p\chi + 2\sigma$ and $6(p\chi + \sigma)^2$ respectively.*

Then the algorithm ProDR.control yields a dynamic regret rate of

$$\sum_{t=1}^{n} f_t(\hat{w}_t) - f_t(u_t) = \tilde{O}(d^3 n^{1/3} [\mathcal{TV}(u_{1:n})]^{2/3} \vee 1),$$

where $(a \vee b) := \max\{a, b\}$. Further for any interval $[a, b] \subseteq [n]$:

$$\sum_{t=a}^{b} f_t(x_t) - f_t(u) = O(d^{1.5} \tau \log n).$$

2.4. Linear regression with delayed feedback

Foster and Simchowitz (2020) reduces the LQR control problem to linear regression problem with *delayed* feedback.

In this setup, the loss at round t is revealed only at round $t+\tau$. This delayed setting can be handled by the framework developed in Joulani et al. (2013).The entire algorithm is as shown in Fig.2 for completeness.

```
ProDR.control.delayed: Inputs- delay $\tau > 0$

  • Maintain $\tau$ separate instances of ProDR.control (Fig.1). Enumerate them by $0, 1, \ldots, \tau - 1$.

  • At time $t$:

      1. Update instance $(t-1) \mod \tau$ with loss $f_{t-\tau}$.
      2. Predict using instance $(t-1) \mod \tau$.
```

Figure 2: ProDR.control.delayed: An instance of delayed to non-delayed reduction from Joulani et al. (2013)

We have the following regret guarantee for Algorithm ProDR.control.delayed.

Theorem 3. *Let x_t be the prediction of the algorithm in Fig. 2 at time t. Instantiating each ProDR.control instance by the parameter setting described in Theorem 2. Let τ be the feedback delay. We have that*

$$\sum_{t=1}^{n} f_t(x_t) - f_t(u_t) = \tilde{O}(d^3 \tau^{2/3} n^{1/3} [\mathcal{TV}(u_{1:n})]^{2/3} \vee \tau).$$

Further for any interval $[a,b] \subseteq [n]$ we have $\sum_{t=a}^{b} f_t(x_t) - f_t(u) = O(d^{1.5} \tau \log n)$.

3. Instantiation for the LQR problem

In view of Proposition 7, the LQR problem is reduced to a mini-batch linear regression problem with delayed feedback, where the delay is given by $h = O(\log n)$ in Proposition 7. In this section, we provide explicit form of the linear regression losses arising in the LQR problem and instantiate Algorithm ProDR.control.delayed (Fig.2). First we need to define certain quantities:

For a sequence of matrices $(M^{[i]})_{i=1}^m$ define $\texttt{flatten}((M^{[i]})_{i=1}^m)$ as follows: Let $M_k^{[i]}$ be the k^{th} column of $M^{[i]}$.

Let's define

$$z^k = \left[(M_1^k)^T, \ldots, (M_{d_x}^k)^T\right]^T \in \mathbb{R}^{d_u d_x},$$

and

$$\texttt{flatten}((M^{[i]})_{i=1}^m) := \left[(z^1)^T, \ldots, (z^m)^T\right]^T \in \mathbb{R}^{m d_u d_x}.$$

For a sequence of DAP parameters $M_{1:n}$, let $\mathcal{TV}(M_{1:n}) := \sum_{t=2}^{n} \sum_{i=1}^{m} \|M_t^{[i]} - M_{t-1}^{[i]}\|_1$. We define $\texttt{deflatten}$ as the natural inverse operation of $\texttt{flatten}$. We have the following Corollary of Theorem 3 and Proposition 7.

Corollary 4. *Assume the notations in Fig.1 and Section B. Let $\Sigma_\infty = U_\infty^T \Lambda_\infty U_\infty$ be the spectral decomposition of the positive semi definite (PSD) matrix $\Sigma_\infty \in \mathbb{R}^{d_u \times d_u}$. . Let the covariate matrix $A_t := [w_{t-1}^T \ldots w_{t-m}^T] \otimes \Lambda_\infty^{1/2} U_\infty \in \mathbb{R}^{d_u \times m d_u d_x}$, where $\otimes$ denotes the Kronecker product. Let the bias vector $b_t := \Lambda_\infty^{1/2} U_\infty q^*_{\infty;h}(w_{t:t+h})$. Let the delay factor of ProDR.control.delayed (Fig.2) be $\tau = h$ as defined in Proposition 7 and let the decision set given to the ProDR.control instances in Fig.2 be the DAP space defined in Eq.(2). Let z_t be the prediction at round t made by the ProDR.control.delayed algorithm and let $M_t^{alg} := \texttt{deflatten}(z_t)$. At round t, we play the control signal $u_t^{alg}(x_t) = \pi_t^{M_t^{alg}}(x_t)$ according to Eq.(1). There exists a choice of input parameters for the ProDR.control instances in Fig.2 such that*

$$\begin{aligned} R(M_{1:n}) &= \sum_{t=1}^{n} \ell(x_t^{alg}, u_t^{alg}) - \ell(x_t^{M_{1:n}}, u_t^{M_{1:n}}) \\ &= \tilde{O}\left(m^3 d^4 d_x^5 (d_u \wedge d_x)(n^{1/3} [\mathcal{TV}(M_{1:n})]^{2/3} \vee 1)\right), \end{aligned}$$

where $M_{1:n}$ is a sequence of DAP policies where each $M_t \in \mathcal{M}$ (eq.(2)). Further the algorithm ProDR.control.delayed also enjoys a strongly adaptive regret guarantee for any interval $[a,b] \subseteq [n]$:

$$\sum_{t=a}^{b} \ell(x_t^{alg}, u_t^{alg}) - \ell(x_t^{M}, u_t^{M}) = \tilde{O}((m d_u d_x)^{1.5} \log n),$$

for any fixed DAP policy $M \in \mathcal{M}$.

The following theorem provides a nearly matching lower bound.

Theorem 5. *There exists an LQR system, a choice of the perturbations w_t and a DAP policy class such that:*

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

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

Remark 6. *The covariate matrix $A_t \in \mathbb{R}^{d_u \times m d_u d_x}$ that arise in Corollary 4 is rank deficient whenever $m d_x > 1$. In such cases, the linear regression losses $f_t(w)$ as in Fig.1 cannot be strongly convex. So the proper universal dynamic regret minimizing algorithm for strongly convex losses from Baby and Wang (2022) is inapplicable in general except potentially for the particular setting of $m = d_x = 1$. Moreover, in the setting of $m = d_x = 1$ a non-zero strong convexity parameter can exist only if the magnitude of the perturbations $|w_t|$ are bounded away from zero which is restrictive in its scope.*

4. Conclusion

In this paper, we proposed a new algorithm for minimizing dynamic regret of the non-stationary linear regression problem. We applied this algorithm to obtain a non-stationary and safe LQR controller. The techniques developed in this work can be of independent interest in the broader literature of online learning under safety constraints. We defer the task of deriving similar dynamic regret rates for general strongly convex losses in the LQR problem as a future work.

References

Dmitry Adamskiy, Wouter M. Koolen, Alexey Chernov, and Vladimir Vovk. A closer look at adaptive regret. *Journal of Machine Learning Research*, 2016.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: Price of past mistakes. In *Advances in Neural Information Processing Systems*, 2015.

Dheeraj Baby and Yu-Xiang Wang. Online forecasting of total-variation-bounded sequences. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In *COLT*, 2021.

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

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 3rd edition, 2005.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Asaf Cassel and Tomer Koren. Bandit linear control. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Ting-Jui Chang and Shahin Shahrampour. On online optimization: Dynamic regret analysis of strongly convex and smooth problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Xi Chen, Yining Wang, and Yu-Xiang Wang. Non-stationary stochastic optimization under lp, q-variation measures. 2018.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Y. Mansour, and Kunal Talwar. Online linear quadratic control. In *ICML*, 2018.

Ashok Cutkosky. Parameter-free, dynamic, and strongly-adaptive online learning. In *ICML*, 2020.

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

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, pages 1405–1411, 2015.

Ronald A. DeVore and George G. Lorentz. Constructive approximation. In *Grundlehren der mathematischen Wissenschaften*, 1993.

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

Gautam Goel and Babak Hassibi. Regret-optimal estimation and control. *Transactions of Automatic Control (TAC)*, 2021.

Gautam Goel and Adam Wierman. An online algorithm for smoothed regression and lqr control. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Paula Gradu, John Hallman, and Elad Hazan. Non-stochastic control with bandit feedback. In *Advances in Neural Information Processing Systems*, 2020a.

Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. *ArXiv*, abs/2007.04393, 2020b.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. In *Electronic colloquium on computational complexity (ECCC)*, volume 14, 2007.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Elad Hazan, Sham Kakade, and Karan Singh. The non-stochastic control problem. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

Andrew Jacobsen and Ashok Cutkosky. Parameter-free mirror descent. *COLT*, 2022.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pages 398–406, 2015.

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

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved Strongly Adaptive Online Learning using Coin Betting. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. 2016.

Yuwei Luo, Varun Gupta, and Mladen Kolar. Dynamic regret minimization for control of non-stationary linear dynamical systems. *Proc. ACM Meas. Anal. Comput. Syst.*, 2022.

N. Merhav, E. Ordentlich, G. Seroussi, and M.J. Weinberger. On sequential strategies for loss functions with memory. In *2000 IEEE International Symposium on Information Theory*, 2000.

Aryan Mokhtari, Shahin Shahrampour, A. Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7195–7201, 2016.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

Guanya Shi, Yiheng Lin, Soon-Jo Chung, Yisong Yue, and Adam Wierman. Online optimization with memory and competitive control. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic. In *Advances in Neural Information Processing Systems*, 2020.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *COLT*, 2020.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning (ICML-16)*, pages 449–457, 2016.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems (NeurIPS-18)*, pages 1323–1333, 2018a.

Lijun Zhang, Tianbao Yang, Zhi-Hua Zhou, et al. Dynamic regret of strongly adaptive methods. In *International Conference on Machine Learning (ICML-18)*, pages 5877–5886, 2018b.

Lijun Zhang, G. Wang, Wei-Wei Tu, and Zhi-Hua Zhou. Dual adaptivity: A universal algorithm for minimizing the adaptive regret of convex functions. *NeurIPS*, 2021a.

Zhiyu Zhang, Ashok Cutkosky, and Ioannis Ch. Paschalidis. Strongly adaptive oco with memory. *ArXiv*, abs/2102.01623, 2021b.

Peng Zhao and Lijun Zhang. Improved analysis for dynamic regret of strongly convex and smooth functions. *L4DC*, 2021.

Peng Zhao, Y. Zhang, L. Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. *NeurIPS*, 2020.

Peng Zhao, Yu-Xiang Wang, and Zhi-Hua Zhou. Nonstationary online learning with memory and nonstochastic control. *AISTATS*, 2022.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

A. Related work

In this section, we review recent progress at the intersection of control and online convex optimization (OCO) that are most relevant to our work.

Online control. The idea of using tools from OCO for general control problem was proposed in Agarwal et al. (2019). They place the notion of regret against the class of DAP policies as the central performance measure. The DAP class is also shown to be sufficiently rich to approximate a wide class of linear state-feedback controllers. Under general convex losses, they propose a reduction to OCO with memory (Merhav et al., 2000; Anava et al., 2015) and derives $O(\sqrt{n})$ regret when the system matrices (A, B) are known. For the case of unknown system, Hazan et al. (2020) provides $O(n^{2/3})$ regret via system identification techniques. When the losses are strongly convex and sub-quadratic, Simchowitz (2020) strengthens these results to attain $\tilde{O}(n)$ regret for known systems and $\tilde{O}(\sqrt{n})$ when the system is unknown. For partially observable systems strong regret guarantees are provided in Simchowitz et al. (2020). Luo et al. (2022) provides an $O(n^{3/5})$ dynamic regret bound for the case when the system matrices (A_t, B_t) can change over time. Their results are incompatible to ours in that they consider unknown dynamics, stochastic disturbances and the dynamic regret compete with controllers that are pointwise optimal (restricted dynamic regret), while we assume known dynamics, adversarial disturbances and compete with an arbitrary sequence of controllers (i.e., universal dynamic regret). There are also a series of recent works such as Gradu et al. (2020a;b); Cassel and Koren (2020); Zhang et al. (2021b); Shi et al. (2020); Goel and Hassibi (2021); Zhao et al. (2022) which explore various other aspects of the control problem. We defer further discussion to the appendix.

Gradu et al. (2020a); Cassel and Koren (2020) studied online control in the partially observed cases with Bandit feedback, but did not consider the problem of non-stationarity with dynamic regret. Gradu et al. (2020b); Zhang et al. (2021b) studied the adaptive regret in nonstochastic control problems, which is an alternative metric to capture the performance of the learning controller in non-stationary environments. Our algorithm uses a reduction to adaptive regret too, but it is highly nontrivial to show that one can tweak adaptive regret minimizing algorithms into ones that achieve optimal dynamic regret. Moreover, our algorithm is the first that achieves logarithmic adaptive regret for nonstochastic LQR control problems too. In contrast, Gradu et al. (2020b); Zhang et al. (2021b) focused on the slower adaptive regret in the general convex loss cases.

To the best of our knowledge, Goel and Hassibi (2021) and Zhao et al. (2022) are the only existing work that considered dynamic regret in non-stochastic control. Goel and Hassibi (2021) used tools from H_∞ control and derived a controller with exact minimax optimal dynamic regret against an oracle controller that sees the whole sequence of disturbances and chooses an optimal sequence of control actions. But the optimal dynamic regret against the sequence of control actions given by the unrealizable oracle controller is linear in n in general. It is unclear whether this oracle controller can be realized by a sequence of time-varying DAP controllers. Zhao et al. (2022) studied the universal dynamic (policy) regret problem similar to ours, but works for a broader family convex loss functions. Their regret bound $O(\sqrt{n(1+C_n)})$ is optimal for the convex loss family. Our results show that the optimal regret improves to $\tilde{O}(n^{1/3}C_n^{2/3} \vee 1)$ when specializing to the LQR problem where the losses are quadratic. On the technical level, Zhao et al. (2022) used a reduction to the dynamic regret of OCO with memory, while we reduced to the dynamic regret of OCO with delayed feedback.

We emphasize that the regret in Eq.(3) is dynamic *policy* regret (Anava et al., 2015; Zhao et al., 2022). The states visited by the reference policy is counterfactual and is different from that of the learner's trajectory which we observe. This is very different from the standard OCO framework where the state of both the learner and adversary are same. So bounding the policy regret seems qualitatively harder than bounding the regret in an OCO setting. Nevertheless, for the LQR problem, the fact that there exists a reduction (Foster and Simchowitz, 2020) from the problem of controlling policy regret to the problem of controlling the standard OCO regret is remarkable

Dynamic regret minimization in online learning. There is a rich body of literature on dynamic regret (Eq.(4)) minimization. As discussed in Section B, the non-staionary LQR problem can be reduced to an instance of linear regression losses which are exp-concave on compact domains. There is a recent line of research (Baby and Wang, 2021; 2022) that provides optimal universal dynamic regret rates under exp-concave losses. However, the algorithm of Baby and Wang (2021) is improper, in the sense that the iterates of the learner can lie outside the feasibility set. The work of Baby and Wang (2022) ameliorates this issue to some extend by providing proper algorithms for the particular case of L_∞ constrained (box) decisions sets. The DAP policy space in Definition 1 is indeed not an L_∞ ball. We note that if improper learning is allowed in the LQR problem, one can run the algorithms of Baby and Wang (2021; 2022) to attain optimal dynamic regret rates. The proper learning algorithms such as Zinkevich (2003); Zhang et al. (2018a); Cutkosky (2020); Jacobsen and Cutkosky (2022) control dynamic regret for general convex losses. However, they are not adequate to optimally minimize dynamic regret under curved losses that are strongly convex or exp-concave. The notion of restrictive dynamic regret introduced in Besbes et al.

(2015) competes with a sequence of minimizers of the losses. This notion of regret can sometimes be overly pessimistic as noted in Zhang et al. (2018a). There is a series of work in the direction of dynamic regret minimization in OCO such as Jadbabaie et al. (2015); Yang et al. (2016); Mokhtari et al. (2016); Chen et al. (2018); Zhang et al. (2018b); Goel and Wierman (2019); Baby and Wang (2019); Zhao et al. (2020); Zhao and Zhang (2021); Zhao et al. (2022); Chang and Shahrampour (2021). However, to the best of our knowledge none of these works are known to attain the optimal universal dynamic regret rate for the setting of online linear regression.

Strongly adaptive regret minimization. There is also a complementary body of literature on strongly adaptive algorithms that focus on controlling the static regret in any local time window. For example, the algorithm of Daniely et al. (2015); Jun et al. (2017) can lead to $\tilde{O}(\sqrt{|I|})$ static regret in any interval of $I \subseteq [n]$ under convex losses. When the losses are exp-concave the algorithm of Hazan and Seshadhri (2007); Adamskiy et al. (2016); Zhang et al. (2021a) can lead to $O(\log n)$ static regret in any interval.

B. Preliminaries

We start with a brief overview of the LQR problem for the sake of completeness. The material of this section closely follows Foster and Simchowitz (2020). The definitions and notations introduced in this section will be used throughout the paper.

A linear control law is given by $u_t = -Kx_t$ for a controller $K \in \mathbb{R}^{d_u \times d_x}$. A linear controller K is said to be stabilizing if $\rho(A - BK) < 1$ where $\rho(A - BK)$ is the maximum of the absolute values of the eigenvalues of $A - BK$. We assume that there exists a stabilizing controller for the system (A, B). For such systems, there exists a unique matrix P_∞ which is the solution to the equation:

$$P = A^T PA + R_x - A^T PB(R_u + B^T PB)^{-1} B^T PA.$$

The solution P_∞ is called the infinite horizon Lyapunov matrix. It is an intrinsic property of the system (A, B) and characterizes the optimal infinite horizon cost for control in the absence of noise (Bertsekas, 2005). We also define the optimal state feedback controller

$$K_\infty := (R_u + B^T P_\infty B)^{-1} B^T P_\infty A, \tag{5}$$

the steady state covariance matrix:

$$\Sigma_\infty := R_u + B^T P_\infty B,$$

and the closed loop dynamics matrix: $A_{\text{cl},\infty} := A - BK_\infty$.

Foster and Simchowitz (2020) shows that the problem of controlling the regret in the LQR problem can be reduced to online linear regression problem with delays. Specifically we have the following fundamental result due to Foster and Simchowitz (2020).

Proposition 7. *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:*

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

where the parameters involved in the inequality are defined as below: $\hat{A}_t(M, w_{t:t+h}) := \|q^M(w_{1:t-1}) - q_{\infty;h}(w_{t:t+h})\|^2_{\Sigma_\infty}$. $q_{\infty;h}(w_{t:h+t}) := \sum_{i=t+1}^{t+h} \Sigma_\infty^{-1} B^T (A_{cl,\infty})^{i-1-t} P_\infty w_i$. $h := 2(1 - \gamma_\infty)^{-1} \log(\kappa_\infty^2 \beta_*^2 \Psi_* \Gamma_*^2 n^2)$. $\gamma_\infty := \|I - P + \infty^{-1/2} R_x P_\infty^{1/2}\|_{op}^{1/2}$. $\kappa_\infty := \|P_\infty^{1/2}\|_{op} \|P_\infty^{-1/2}\|_{op}$. $\beta_* := \max\{1, \lambda_{min}^{-1}(R_u), \lambda_{min}^{-1}(r_x)\}$. $\Psi_* = \max\{1, \|A\|_{op}, \|B\|_{op}, \|R_x\|_{op}, \|R_u\|_{op}\}$. $\Gamma_* := \max\{1, \|P_\infty\|_{op}\}$

Observe that the losses $\hat{A}_t(M, w_{t:t+h}) := \|q^M(w_{1:t-1}) - q_{\infty;h}(w_{t:t+h})\|^2_{\Sigma_\infty} = \hat{A}_t(M, w_{t:t+h}) := \|\Sigma_\infty^{1/2} q^M(w_{1:t-1}) - \Sigma_\infty^{1/2} q_{\infty;h}(w_{t:t+h})\|_2^2$ are essentially linear regression losses. The quantity $\Sigma_\infty^{1/2} q^M(w_{1:t-1})$ is a linear map from the matrix sequence M to $\mathbb{R}^{d_u}$. However, there is one caveat in that the bias vector at round t given by $\Sigma_\infty^{1/2} q_{\infty;h}(w_{t:t+h})$ is only available at round $t + h = t + O(\log n)$. This issue of delayed feedback can be directly handled using the delayed to non-delayed online learning reduction from Joulani et al. (2013).

C. Brief overview of results from Baby and Wang (2022)

For the sake of completeness, we dedicate this session for a short discussion about the results of Baby and Wang (2021).

First, we recall the description of Follow-the-Leading-History (FLH) algorithm from (Hazan and Seshadhri, 2007).

```
FLH: inputs - Learning rate $\zeta$ and $n$ base learners $E^1, \dots, E^n$
  1. For each $t$, $v_t = (v_t^{(1)}, \dots, v_t^{(t)})$ is a probability vector in $\mathbb{R}^t$. Initialize $v_1^{(1)} = 1$.
  2. In round $t$, set $\forall j \leq t$, $x_t^j \leftarrow E^j(t)$ (the prediction of the $j^{th}$ bas learner at time $t$). Play $x_t = \sum_{j=1}^t v_t^{(j)} x_t^{(j)}$.
  3. After receiving $f_t$, set $\hat{v}_{t+1}^{(t+1)} = 0$ and perform update for $1 \leq i \leq t$:
       $$\hat{v}_{t+1}^{(i)} = \frac{v_t^{(i)} e^{-\zeta f_t(x_t^{(i)})}}{\sum_{j=1}^t v_t^{(j)} e^{-\zeta f_t(x_t^{(j)})}}$$
  4. Addition step - Set $v_{t+1}^{(t+1)}$ to $1/(t+1)$ and for $i \neq t+1$:
       $$v_{t+1}^{(i)} = (1 - (t+1)^{-1}) \hat{v}_{t+1}^{(i)}$$
```

Figure 3: FLH algorithm

Next, we describe Online Newton Step (ONS) algorithm from Hazan et al. (2007).

```
ONS: inputs - $\zeta$. Decision set $\mathcal{D}$.
  1. At round 1, predict 0.
  2. At iteration $t > 1$ predict:
       $$w_t \in \underset{x \in \mathcal{D}}{\operatorname{argmin}} \|w_{t-1} - \frac{1}{\beta} A_{t-1}^{-1} \nabla_{t-1} - x\|_{A_{t-1}},$$
     where $\nabla_\tau = \nabla f_\tau(x_\tau),\ A_t = \zeta I_d + \sum_{i=1}^t \nabla_i \nabla_i^\top$.
```

Figure 4: ONS algorithm

Assumption A1: The loss functions ℓ_t are α exp-concave in the box decision set $\mathcal{D} = \{x \in \mathbb{R}^d : \|x\|_\infty \leq B\}$ .ie, $\ell_t(y) \geq \ell_t(x) + \nabla \ell_t(x)^T (y - x) + \frac{\alpha}{2} \left(\nabla \ell_t(x)^T (y - x)\right)^2$ for all $x, y \in \mathcal{D}$.

Assumption A2: The loss functions ℓ_t satisfy $\|\nabla \ell_t(x)\|_2 \leq G$ and $\|\nabla \ell_t(x)\|_\infty \leq G_\infty$ for all $x \in \mathcal{D}$. Without loss of generality, we let $G \wedge G_\infty \wedge B \geq 1$, where $a \wedge b := \min\{a, b\}$.

We consider the following protocol:

- At time $t \in [n]$ learner predicts $x_t \in \mathbb{R}^d$ with $\|x_t\|_\infty \leq B$.
- Adversary reveals the loss function ℓ_t.

In view of Assumption A1, following (Hazan et al., 2007), one can define the surrogate losses:

$$f_t(x) = \left(\sqrt{\alpha/2} \nabla \ell_t(x_t)^T (x - x_t) + 1/\sqrt{2\alpha}\right)^2.$$

The surrogate losses satisfy the following property:

$$\sum_{t=1}^{n} \ell_t(x_t) - \ell_t(w_t) \leq \sum_{t=1}^{n} f_t(x_t) - f_t(w_t),$$

where $x_t, w_t \in \mathcal{D}$.

We have the following dynamic regret guarantee from Baby and Wang (2021).

Theorem 8. *Suppose Assumptions A1-A2 are satisfied. Define $\gamma := 2GB\sqrt{\alpha d/2} + 1/\sqrt{2\alpha}$. By using the base learner as ONS with parameter $\zeta = \min\left\{\frac{1}{16GB\sqrt{d}}, 1/(4\gamma^2)\right\}$, decision set $\mathcal{D}$, loss at time t to be f_t and choosing learning rate of FLH as $\eta = 1/(2\gamma^2)$, FLH-ONS obeys*

$$\sum_{t=1}^{n} \ell_t(x_t) - \ell_t(w_t) \leq \tilde{O}\left(140d^2(8G^2B^2\alpha d + G^2B^2 + 1/\alpha)(n^{1/3}[\mathcal{TV}(u_{1:n})]^{2/3} \vee 1)\right)\mathbb{I}\{\mathcal{TV}(u_{1:n}) > 1/n\}$$
$$+ \tilde{O}\left(d(8G^2B^2\alpha d + 1/\alpha\right)\mathbb{I}\{\mathcal{TV}(u_{1:n}) \leq 1/n\},$$

where x_t is the decision of the algorithm at time t and $\tilde{O}(\cdot)$ hides polynomial factors of $\log n$. $\mathbb{I}\{\cdot\}$ is the boolean indicator function assuming values in $\{0, 1\}$.

We also have the following strongly adaptive regret guarantee:

Theorem 9. *Consider the setting of FLH-ONS in Theorem 8. Then for any interval $[a, b] \subseteq [n]$ we have that*

$$\sum_{t=a}^{b} \ell_t(x_t) - \ell_t(w_t) = O(d^{1.5}\log n).$$

D. Some definitions and terminologies

Terminology. For a convex loss function f, we abuse the notation and take $\nabla f(x)$ to be a sub-gradient of f at x. We denote $\mathcal{D}_\infty(\tilde{R}) := \{x \in \mathbb{R}^d : \|x\|_\infty \leq \tilde{R}\}$.

Linear regression losses belong to a broad family of convex loss functions called exp-concave losses:

Definition 10. *A convex function f is α exp-concave in a domain $\mathcal{D}$ if for all $x, y \in \mathcal{D}$ we have $f(y) \geq f(x) + \nabla f(x)^T(x - y) + \frac{\alpha}{2}(\nabla f(x)^T(x-y))^2$.*

The losses $f_t(z) = \|A_t z - b_t\|_2^2$ are $(2R)^{-1}$ exp-concave if $f(z) \leq R$ for all $z \in \mathcal{D}$ (see Lemma 2.3 in Foster and Simchowitz (2020)).

Definition 11. *A convex function f is σ strongly convex wrt $\|\cdot\|_2$ norm in a domain $\mathcal{D}$ if for all $x, y \in \mathcal{D}$ we have $f(y) \geq f(x) + \nabla f(x)^T(x - y) + \frac{\sigma}{2}\|x - y\|_2^2$.*

We note that if the matrix A_t is rank deficient, then the losses $f_t(z)$ cannot be strongly convex. Moving forward we do not impose any restrictive assumptions on the rank of A_t. As mentioned in Remark 6, the covariate matrix that arise in the reduction of the LQR problem to linear regression is not in general full rank. So we target a solution that can handle general covariate matrices irrespective of their rank.

E. Intuition behind ProDR.control algorithm in Fig.1

Starting point of our algorithm design is the work of Baby and Wang (2022). They provide an algorithm that attains optimal dynamic regret when the losses are exp-concave. However, their setting works only in a very restrictive setup where the decision set is an L_∞ constrained box. Consequently, we cannot directly apply their results to the linear regression problem of Section 2 whenever the decision set $\mathcal{D}$ is a general convex set.

An online learner is termed proper if the decisions of the learner are guaranteed to lie within the feasibility set $\mathcal{D}$. Otherwise it is called improper. A recent seminal work of Cutkosky and Orabona (2018) proposes neat reductions that can convert an improper online learner to a proper one, whenever the losses are convex. Following this line of research, we can aim to convert the algorithm of Baby and Wang (2022) that works exclusively on box decision set to one that can support arbitrary convex decision sets by coming up with suitable reduction schemes. However, the specific reduction scheme proposed in Cutkosky and Orabona (2018) is inadequate to yield fast dynamic rates for exp-concave losses. Our algorithm ProDR.control (Fig.1, **Pro**per **D**ynamic **R**egret.control) is a by-product of constructing new reduction schemes to circumvent the aforementioned problem for the case of linear regression losses. We expand upon these details below.

In ProDR.control, we maintain a surrogate algorithm $\mathcal{A}$, which is chosen to be the algorithm of Baby and Wang (2022) that produces iterates w_t in an L_∞ norm ball (box), $\mathcal{D}_\infty$, that encloses the actual decision set $\mathcal{D}$. Since w_t can be infeasible, we play $\hat{w}_t$ obtained via a special type of projection of w_t onto $\mathcal{D}$ which is formulated as a min-max problem in Line 3 of Fig.1. In Line 4, we construct surrogate losses ℓ_t to be passed to the algorithm $\mathcal{A}$. The surrogate loss penalises $\mathcal{A}$ for making predictions outside $\mathcal{D}$. We will show (see Lemma 12 in Appendix) that the instantaneous regret satisfies $f_t(\hat{w}_t) - f_t(u_t) \leq \ell_t(w_t) - \ell_t(u_t)$, where $u_t \in \mathcal{D}$ is the comparator at round t. Thus the dynamic regret of the proper iterates $\hat{w}_t$ wrt linear regression losses is upper bounded by the dynamic regret of the surrogate algorithm $\mathcal{A}$ on the losses ℓ_t and box decision set.

The design of the min-max barrier $S_t(w)$ is driven to ensure exp-concavity of the surrogate losses $\ell_t(w) = f_t(w) + G \cdot S_t(w)$. We capture its intuition as follows. We start by observing that since $\nabla^2 f_t(w) = 2A_t^T A_t$, the linear regression losses f_t exhibits strong curvature along the row-space of A_t, denoted by $\text{row}(A_t)$. Further we have $\nabla f_t(w) = 2A_t^T(A_t w - b_t) \in \text{row}(A_t)$. So the loss f_t exhibits strong curvature along the direction of its gradient too. This is the fundamental reason behind the exp-concavity of f_t. The min-max barrier $S_t(w)$ is designed such that its gradient is guaranteed to lie in the $\text{row}(A_t)$ (see Lemma 13 in Appendix for a formal statement). So the overall gradient $\nabla \ell_t(w)$ also lies in the $\text{row}(A_t)$. Since the function f_t already exhibits strong curvature along $\text{row}(A_t)$, we conclude that the sum $\ell_t(w) = f_t(w) + G \cdot S_t(w)$ exhibits strong curvature along its gradient $\nabla \ell_t(w)$. This maintains the exp-concavity of the losses ℓ_t over $\mathcal{D}_\infty$ (see Lemma 14 in Appendix). Such curvature considerations along with the fact that $S_t(w)$ has to be sufficiently large to facilitate the instantaneous regret bound $f_t(\hat{w}_t) - f_t(u_t) \leq \ell_t(w_t) - \ell_t(u_t)$ results in functional form for $S_t(w)$ displayed in Fig.1.

Consequently the fast dynamic regret rates derived in Baby and Wang (2022) becomes directly applicable. We remark that the design of appropriate surrogate losses and proving their exp-concavity is non-trivial in hindsight.

The reduction scheme used by Cutkosky and Orabona (2018) for producing proper iterates $\hat{w}_t$ and their accompanying surrogate loss design ℓ_t also allows one to upper bound the regret wrt linear regression losses f_t by the regret of the algorithm $\mathcal{A}$ wrt surrogate losses ℓ_t. However, the surrogate loss ℓ_t they construct is not guaranteed to be exp-concave and consequently not amenable to fast dynamic regret rates.

F. Omitted Proofs

In this section we use the notations defined in Fig.1.

The lemma below shows how the surrogate losses ℓ_t can be used to upper bound the regression losses f_t.

Lemma 12. *Assume the notations in Fig.1. Let G be such that* $\sup_{w_1, w_2 \in \mathcal{D}_\infty(\tilde{R})} \|A_t(w_1 + w_2) - 2b_t\|_1 \leq G$ *for all* $t \in [n]$. *We have that:*

- $f_t(\hat{w}_t) \leq \ell_t(w_t)$,
- $f_t(u) = \ell_t(u)$ *for all* $u \in \mathcal{D}$

Proof. For any $w_1, w_2 \in \mathcal{D}_\infty(\tilde{R})$

$$\begin{aligned} f_t(w_1) - f_t(w_2) &= \|A_t w_1 - b_t\|_2^2 - \|A_t w_2 - b_t\|_2^2 \\ &= \left(A_t(w_1 + w_2) - 2b_t\right)^T \left(A_t(w_1 - w_2)\right) \\ &\leq \|A_t(w_1 + w_2) - 2b_t\|_1 \|A_t(w_1 - w_2)\|_\infty \\ &\leq G \max_{i=1,\ldots,p} |a_{t,i}^T (w_1 - w_2)|, \end{aligned} \tag{6}$$

for a G such that $\sup_{w_1, w_2 \in \mathcal{D}_\infty(\tilde{R})} \|A_t(w_1 + w_2) - 2b_t\|_1 \leq G$ holds true.

In particular we have that:

$$f_t(\hat{w}_t) \leq f_t(w_t) + G \max_{i=1,\dots,p} |a_{t,i}^T(\hat{w}_t - w_t)| := \ell_t(w_t)$$

For any $u \in \mathcal{D}$, we have that $S_t(u) = 0$. Hence $f_t(u) = \ell_t(u)$.

□

The lemma below establishes certain useful properties of the barrier function $S_t(w)$.

Lemma 13. *The function $S_t(w)$ satisfies the following properties:*

1. $S_t(w) = \max_{i=1,\dots,p} \min_{x \in \mathcal{D}} |a_{i,t}^T(x-w)|$.
2. *$S_t(w)$ is convex over $\mathbb{R}^d$.*
3. *Let i^* be such that $S_t(w) = \min_{x \in \mathcal{D}} |a_{i^*,t}^T(x-w)|$. Let $\Pi(w) \in \operatorname{argmin}_{x \in \mathcal{D}} |a_{i^*,t}^T(x-w)|$. Let $g_t \in \partial S_t(w)$, When $a_{i^*,t}^T(\Pi(w)-w) \neq 0$ we have:*

$$g_t = \begin{cases} a_{i^*,t}, & \text{if} \quad a_{i^*,t}^\top(\Pi(w)-w) < 0 \\ -a_{i^*,t}, & \text{if} \quad a_{i^*,t}^\top(\Pi(w)-w) > 0. \end{cases}$$

If $a_{i^*,t}^T(\Pi(w)-w) = 0$ then we take $g_t = 0$.

Proof. We set out to prove the first statement. Let Δ_p be the p dimensional simplex. We have that

$$\begin{aligned} S_t(w) &= \min_{x \in \mathcal{D}} \max_{i=1,\dots,p} |a_{i,t}^T(x-w)| \\ &=_{(a)} \min_{x \in \mathcal{D}} \max_{v \in \Delta_p} \sum_{i=1}^{p} v_i |a_{i,t}^T(x-w)| \\ &=_{(b)} \max_{v \in \Delta_p} \min_{x \in \mathcal{D}} \sum_{i=1}^{p} v_i |a_{i,t}^T(x-w)|. \end{aligned}$$

For line (a) we observed that for a given x $\max_{v \in \Delta_p} \sum_{i=1}^p v_i |a_{i,t}^T(x-w)|$ is attained by putting all the weights of v to an $i^* \in \operatorname{argmax}_{i=1,\dots,p} |a_{i,t}^T(x-w)|$.

For line (b) we observe that the function $r(x,v) = \sum_{i=1}^p v_i |a_{i,t}^T(x-w)|$ is a convex function of x and concave function of p. So by applying Sion's minimax theorem we arrive at line (b).

Next we set out to prove that:

$$\max_{v \in \Delta_p} \min_{x \in \mathcal{D}} r(x,v) = \max_{i=1,\dots,p} \min_{x \in \mathcal{D}} |a_{i,t}^T(x-w)| \tag{7}$$

Let (x^*, v^*) be a solution that attains $\max_{v \in \Delta_p} \min_{x \in \mathcal{D}} r(x,v)$. Further, for the sake of contradiction, let's assume that $v^* \neq e_k$ for any $k \in [p]$. (e_k is the unit vector with 1 at entry k). Let the index j be such that $|a_{j,t}^T(x^*-w)| > |a_{i,t}^T(x^*-w)|$ for all $i \in [p] \setminus \{j\}$. Then we can find a solution e_j such that $r(x^*, e_j) > r(x^*, v^*)$. This contradicts the fact that (x^*, v^*) is a valid solution.

In the alternate case let j be an index in $[p]$ such that $|a_{j,t}^T(x^*-w)| \geq |a_{i,t}^T(x^*-w)|$ for all $i \in [p] \setminus \{j\}$. Suppose for all $i \in Q \subseteq [p] \setminus \{j\}$ we have $|a_{j,t}^T(x^*-w)| = |a_{i,t}^T(x^*-w)|$. By earlier arguments, we must have $v^*[k]$ must be equal to zero for all $k \in [p] \setminus (Q \cup \{j\})$. Then putting all the weight to j produces an equally valid solution in the sense that $r(x^*, e_j) = r(x^*, v^*)$

Combining the above two cases, we conclude that there exists maximizers v^* such that $v^* = e_k$ for some $k \in [p]$. This leads to Eq.(7).

Next we prove statement 2. For any given i we have that $|a_{i,t}^T(x-w)|$ is a convex function of both x and w. Hence the point-wise maximum $\max_{i=1,\dots,p}|a_{i,t}^T(x-w)|$ is also convex in both x and w. Since partial minimisation preserves convexity, we have that $\min_{x\in\mathcal{D}}\max_{i=1,\dots,p}|a_{i,t}^T(x-w)|$ remains convex in $w\in\mathbb{R}^d$.

Next we prove statement 3. We know that sub-gradient set of point-wise maximum of convex functions is the convex hull of sub-gradients of the active functions. Applying this result along with the sub-gradient characterization of the function $\min_{x\in\mathcal{D}}|a_{i,t}^T(x-w)|$ in Lemma 15 leads to the third statement.

□

The next lemma establishes the exp-concavity of the surrogate losses ℓ_t over the decision domain of the surrogate algorithm $\mathcal{A}$.

Lemma 14. *Assume the notations in Fig.1. Let L be such that $\sup_{w\in\mathcal{D}_\infty(\tilde{R}),j\in[p]}2\|A_tw-b_t\|_2^2+2G^2\leq L$ for all $t\in[n]$. Then the losses ℓ_t are exp-concave over $\mathcal{D}_\infty(\tilde{R})$ with parameter $1/4L$.*

Proof. Observe that $\nabla f_t(w)=2A_t^T(A_tw-b_t)$ and $\nabla^2 f_t(w)=2A_t^TA_t$.

We have that for any $w_1,w_2\in\mathbb{R}^d$

$$f_t(w_2)=f_t(w_1)+\langle\nabla f_t(w_1),w_2-w_1\rangle+\frac{1}{2}\|w_2-w_1\|_{2A_t^TA_t}^2. \tag{8}$$

Due to the convexity of $S_t(w)$ over $\mathbb{R}^d$ from Lemma 13, we have that

$$S_t(w_2)\geq S(w_1)+\langle\nabla S_t(w_1),w_2-w_1\rangle. \tag{9}$$

Combining Eq.(8) and (9) we have that

$$\ell_t(w_2)\geq\ell_t(w_1)+\langle\nabla\ell_t(w_1),w_2-w_1\rangle+\frac{1}{2}\|w_2-w_1\|_{2A_t^TA_t}^2$$

Observe that $\nabla\ell_t(w_1)=2A_t^T(A_tw_t-b_t)+GhA_t^Te_j$, for some $h\in\{-1,0,1\}$ and $j\in[p]$ due to Lemma 13. Now, let's focus on points $w_1,w_2\in\mathcal{D}_\infty(\tilde{R})$. We have

$$\begin{aligned}\nabla\ell_t(w_1)\nabla\ell_t(w_1)^T&=4A_t^T(A_tw_1-b_t+Ghe_j)(A_tw_1-b_t+Ghe_j)^TA_t\\&\preccurlyeq 4LA_t^TA_t,\end{aligned}$$

L is such that:

$$\sup_{w\in\mathcal{D}_\infty(\tilde{R}),j\in[p]}\|(A_tw-b_t+Ghe_j)\|_2^2\leq L.$$

Hence for all $w_1,w_2\in\mathcal{D}_\infty(\tilde{R})$, we have the relation

$$\ell_t(w_2)\geq\ell_t(w_1)+\langle\nabla\ell_t(w_1),w_2-w_1\rangle+\frac{1}{4L}\|w_2-w_1\|_{\nabla\ell_t(w_1)\nabla\ell_t(w_1)^T}^2.$$

Thus the losses ℓ_t remains exp-concave over $\mathcal{D}_\infty(\tilde{R})$ with parameter $1/4L$.

□

We are now ready to prove Theorem 2.

Theorem 2. *Let $u_{1:n} \in \mathcal{D}$ be any comparator sequence. In Fig.1, choose G such that $\sup_{w_1,w_2 \in \mathcal{D}_\infty(\tilde{R}), t\in[n]} \|A_t(w_1 + w_2) - 2b_t\|_1 \leq G$. Let α be as in Assumption 2. Let L be such that $\sup_{w \in \mathcal{D}_\infty(\tilde{R}), j \in [p]} 2\|A_t w - b_t\|_2^2 + 2G^2 \leq L$ for all $t \in [n]$. Choose $\mathcal{A}$ as the algorithm from Baby and Wang (2022) (see Appendix C) with parameters $\gamma = 2G\alpha\tilde{R}\sqrt{d/8L} + \sqrt{2L}$ and $\zeta = \min\{\frac{1}{16G\alpha\tilde{R}\sqrt{d}}, 1/(4\gamma^2)\}$ and decision set $\mathcal{D}_\infty(\tilde{R})$. Under Assumptions 1 and 2, a valid of assignment of G and L are $2p\chi + 2\sigma$ and $6(p\chi + \sigma)^2$ respectively.*

Then the algorithm ProDR.control yields a dynamic regret rate of

$$\sum_{t=1}^{n} f_t(\hat{w}_t) - f_t(u_t) = \tilde{O}(d^3 n^{1/3} [\mathcal{TV}(u_{1:n})]^{2/3} \vee 1),$$

where $(a \vee b) := \max\{a, b\}$. Further for any interval $[a, b] \subseteq [n]$:

$$\sum_{t=a}^{b} f_t(x_t) - f_t(u) = O(d^{1.5} \tau \log n).$$

Proof. From Eq.(6) we have that for any $w_1, w_2 \in \mathcal{D}_\infty(\tilde{R})$

$$f_t(w_1) - f_t(w_2) \leq G\alpha \|w_1 - w_2\|_2,$$

for a G such that $\sup_{w_1,w_2 \in \mathcal{D}_\infty(\tilde{R})} \|A_t(w_1 + w_2) - 2b_t\|_1 \leq G$ holds true.

From Lemma 13 we have for any subgradient $\|\nabla S_t(w)\|_2 \leq \alpha$ (where α is as in Assumption 1). Thus the losses ℓ_t are $2G\alpha$-Lipschitz in L2 norm over $\mathcal{D}_\infty(\tilde{R})$. Now combining Lemma 14 and Theorem 10 in Baby and Wang (2022) (or see Appendix C) we have that

$$\begin{aligned}\sum_{t=1}^{n} \ell_t(w_t) - \ell_t(u_t) &= \tilde{O}\left((d^3 G^2 \alpha^2 \tilde{R}^2 / L + d^2 G^2 \alpha^2 \tilde{R}^2 + d^2 L)(n^{1/3} [\mathcal{TV}(u_{1:n})]^{2/3} \vee 1) \right)) \\ &= \tilde{O}(d^3 n^{1/3} [\mathcal{TV}(u_{1:n})]^{2/3} \vee 1).\end{aligned}$$

Applying Lemma 12 now concludes the proof. □

Lemma 15. *Let $f(x) = min_{u \in \mathcal{D}} |a^T(u - x)|$ for a compact and convex set $\mathcal{D}$. Let $0 \in \mathcal{D}$. f(x) is convex. Let $s \in argmin_{u \in \mathcal{D}} |a^T(u - x)|$.*

$$\nabla f(x) = \begin{cases} -a & a^T(s - x) > 0 \\ a & a^T(s - x) < 0 \\ 0 & o.w \end{cases}$$

Proof. First we argue the convexity of f. Observe that

$$\begin{aligned} f(x) &= \min_{u \in \mathcal{D}} |a^T(u - x)| \\ &= \min_{u \in \mathcal{D}} \|u - x\|_{aa^T}. \end{aligned}$$

The norm $\|u - x\|_{aa^T}$ is convex in both u and x across $\mathbb{R}^d$. So we have that $f(x)$ which is obtained by partial minimization of a convex function across a convex domain remains convex over $\mathbb{R}^d$. It follows that for any $x, y \in \mathbb{R}^d$,

Now let x be such that $\nabla f(x) = 0$. Existence of such a point is guaranteed since $\mathcal{D}$ in the definition of f is compact.

$$f(y) \geq f(x) + \nabla f(x)^T (y - x). \tag{10}$$

We proceed to show the Lipschitzness of f. Let $w \in argmin_{u \in \mathcal{D}} |a^T(u-x)|$. We have

$$\begin{aligned} f(y) - f(x) &= \min_{u \in \mathcal{D}} |a^T(u-y)| - \min_{u \in \mathcal{D}} |a^T(u-x)| \\ &\leq |a^T(w-x)| - |a^T(w-y)| \\ &\leq |a^T(x-y)| \\ &\leq \|a\|_2 \|x-y\|_2. \end{aligned} \tag{11}$$

Since $\|a\|_2 \leq \kappa$, we conclude that the function f is κ Lipschitz.

We argue that $\nabla f(x) = \lambda a$ for some scalar λ. Let b be a such that $a^T b = 0$. Let $z = y + \sigma b$. Notice that by the definition of f, we have that $f(y) = f(z)$. So,

$$\begin{aligned} f(z) &= f(y) \\ &\geq f(x) + \nabla f(x)^T (z-x) \\ &= f(x) + \nabla f(x)^T (y-x) + \sigma \nabla f(x)^T b. \end{aligned}$$

The above inequality must hold for any σ. Note that both $f(y)$ and $f(x)$ is bounded for any two points in $x, y \mathbb{R}^d$. Further, $\nabla f(x)^T (y-x)$ is also bounded due to the Lipschitzness of f. So if $\nabla f(x)^T b$ is not zero, we can choose a σ such that inequality is violated, leading to a contradiction in the convexity of f across $\mathbb{R}^d$.

So $\nabla f(x)^T b = 0$. This implies that $\nabla f(x) = \lambda(x) a$ for some scalar $\lambda(x)$ and for any $x \in \mathbb{R}^d$.

Next, we argue that $\lambda(x) \in [-1, 1]$. Combining Eq.(10) and (11) we have

$$|a^T(x-y)| \geq \nabla f(x)^T (y-x),$$

for all $x, y \in \mathbb{R}^d$. So taking $y = 0$ followed by $y = 2x$ leads to

$$|a^T x| \geq \pm \lambda(x) a^T x.$$

Suppose x is chosen such that $a^T x \neq 0$. Then the above inequality implies that $\lambda(x) \in [-1, 1]$.

Let $w \in argmin_{u \in \mathcal{D}} |a^T(u-x)|$. Let $s = (x+w)/2$. We have that

$$f(s) \geq f(x) + \lambda(x) a^T (s-x). \tag{12}$$

Moroever,

$$\begin{aligned} f(s) &\leq |a^T(w-s)| \\ &= \frac{1}{2} |a^T(x-w)| \\ &= f(x) - |a^T(x-s)|. \end{aligned} \tag{13}$$

Combining Eq.(14) and (15), we obtain

$$-|a^T(s-x)| \geq \lambda(x) a^T (s-x).$$

Recall that when $a^T x \neq 0$, $\lambda(x) \in [-1, 1]$.

So we conclude that if $a^T x \neq 0$ and $a^T(s-x) > 0$, then $\lambda(x) \leq -1$. This implies that $\lambda(x) = -1$ as $\lambda(x) \in [-1, 1]$ holds true.

Similarly if $a^T x \neq 0$ and $a^T(s-x) < 0$, then $\lambda(x) \geq 1$. This implies that $\lambda(x) = 1$ as $\lambda(x) \in [-1, 1]$ holds true.

Now if $a^T x \neq 0$ and $a^T(s-x) = 0$, we can choose $\lambda(x) = 0$ as $f(z) \geq f(x) + \lambda(x) a^T(z-x) = 0$ holds true for any z.

If $a^T x = 0$, $0 \in argmin_{u \in \mathcal{D}} |a^T(u-x)|$ as $0 \in \mathcal{D}$ is assumed to be true. So by using the previous line of arguments we conclude that $\lambda(x) = 0$. □

Theorem 3. *Let x_t be the prediction of the algorithm in Fig. 2 at time t. Instantiating each ProDR.control instance by the parameter setting described in Theorem 2. Let τ be the feedback delay. We have that*

$$\sum_{t=1}^{n} f_t(x_t) - f_t(u_t) = \tilde{O}(d^3 \tau^{2/3} n^{1/3} [\mathcal{TV}(u_{1:n})]^{2/3} \vee \tau).$$

Further for any interval $[a, b] \subseteq [n]$ we have $\sum_{t=a}^{b} f_t(x_t) - f_t(u) = O(d^{1.5} \tau \log n)$.

Proof. By following the arguments in Joulani et al. (2013), we have that

$$\sum_{t=1}^{n} f_t(x_t) - f_t(u_t) = \sum_{i=1}^{\tau} \sum_{k=1}^{\lfloor 1 + \frac{n-i}{\tau} \rfloor} f_t(x_{i+(k-1)\tau}) - f_t(u_{i+(k-1)\tau}).$$

The second summation in the above expression is the dynamic regret of instance i wrt comparator sequence $\{u_{i+(k-1)\tau}\}$ with k ranging from 1 to $\lfloor 1 + \frac{n-i}{\tau} \rfloor$. Now by triangle inequality we have that

$$\sum_{k=2}^{\lfloor 1 + \frac{n-i}{\tau} \rfloor} \|u_{i+(k-1)\tau} - i + (k-2)\tau\|_1 \leq \sum_{t=2}^{n} \|u_t - u_{t-1}\|_1 = \mathcal{TV}(u_{1:n}).$$

Thus by Theorem 2 we have

$$\begin{aligned} \sum_{t=1}^{n} f_t(x_t) - f_t(u_t) &\leq \sum_{i=1}^{\tau} \tilde{O}(d^3 (n/\tau)^{1/3} \vee 1) \\ &\leq \tilde{O}(d^3 \tau^{2/3} n^{1/3} [\mathcal{TV}(u_{1:n})]^{2/3} \vee \tau). \end{aligned}$$

□

Next, we provide the version of Corollary 4 indicating the closed form expression for all the algorithm parameters. We fo

Corollary 16. *Let $\Sigma_\infty = U_\infty^T \Lambda_\infty U_\infty$ be the spectral decomposition of the positive semi definite (PSD) matrix $\Sigma_\infty \in \mathbb{R}^{d_u \times d_u}$. Assume the notations in Fig.1. Let the covariate matrix $A_t := [w_{t-1}^T \dots w_{t-m}^T] \otimes \Lambda_\infty^{1/2} U_\infty$, where $\otimes$ denotes the Kronecker product. Let the bias vector $b_t := \Lambda_\infty^{1/2} U_\infty q_{\infty;h}^*(w_{t:t+h})$. For a sequence of DAP parameters $M_{1:n}$, let $\mathcal{TV}(M_{1:n}) := \sum_{t=2}^{n} \sum_{i=1}^{m} \|M_t^{[i]} - M_{t-1}^{[i]}\|_1$. For a sequence of matrices $(M^{[i]})_{i=1}^m$ define $\texttt{flatten}((M^{[i]})_{i=1}^m)$ as follows: Let $M_k^{[i]}$ be the k^{th} column of $M^{[i]}$.*

Let's define

$$z^k = \begin{bmatrix} M_1^k \\ \vdots \\ M_{d_x}^k \end{bmatrix} \in \mathbb{R}^{d_u d_x},$$

and

$$\texttt{flatten}((M^{[i]})_{i=1}^m) := \begin{bmatrix} z^1 \\ \vdots \\ z^m \end{bmatrix} \in \mathbb{R}^{m d_u d_x}.$$

Let the decision set given to the ProDR.control (Fig.1) algorithm be the DAP space defined in Eq.(2). Let $G = 2md_ud_xR\gamma\sqrt{d_x \wedge d_u}\|\Lambda^{1/2}U_\infty\|_1 + 2\frac{\|\Lambda^{-1/2}U_\infty B^T\|_2\|P_\infty\|_2\sqrt{d_u}}{1-\gamma}$*. Let the delay factor of ProDR.control.delayed (Fig.2) be* $\tau = h$ *as defined in Proposition 7. Choose* $\alpha = \sqrt{m\|\Sigma_\infty\|_{op}}$ *and* $L = 4G^2$*. Let* $\tilde{R}$ *in Theorem 2 be chosen as* $\tilde{R} = R\gamma\sqrt{d_u \wedge d_x}$*. Let* z_t *be the prediction at round* t *made by the ProDR.control.delayed algorithm. Let* $M_t^{alg} := \texttt{deflatten}(z_t)$*, where* `deflatten` *is the natural inverse operation of* `flatten` *defined above. Let* $\pi := (M_1, \ldots, M_n)$ *define a sequence of DAP policies. For a sequence of matrices* M*, define* $\|M\|_1 := \sum_{i=1}^m \|M^{[i]}\|_1$*. By playing a control* $u_t^{alg}(x_t) = \pi_t^{M_t^{alg}}(x_t)$ *according to Eq.(1), we have that*

$$R_n(M_{1:n}) = \sum_{t=1}^n \ell(x_t^{alg}, u_t^{alg}) - \ell(x_t^{M_{1:n}}, u_t^{M_{1:n}}) = \tilde{O}\left(m^3 d^4 d_x^5 (d_u \wedge d_x)(n^{1/3}[\mathcal{TV}(M_{1:n})]^{2/3} \vee 1)\right),$$

where $M_{1:n}$ *is a sequence of DAP policies where each* $M_t \in \mathcal{M}$ *(eq.(2)). Further the algorithm ProDR.control.delayed also enjoys a strongly adaptive regret guarantee for any interval* $[a, b] \subseteq [n]$*:*

$$\sum_{t=a}^b \ell(x_t^{alg}, u_t^{alg}) - \ell(x_t^M, u_t^M) = \tilde{O}((md_ud_x)^{1.5}\log n),$$

for any fixed DAP policy $M \in \mathcal{M}$*.*

Proof. Define

$$X_t = [w_{t-1}^T \ldots w_{t-m}^T] \otimes I_{d_u},$$

where $I_{d_u} \in \mathbb{R}^{d_u \times d_u}$ is the identity matrix and $\otimes$ denotes the Kronecker product. Clearly $X_t \in \mathbb{R}^{d_u \times md_ud_x}$.

With these definitions, it is easy to verify that

$$q^M(w_{t-1}) = X_t z.$$

Now we return back to losses $\hat{A}_t$ mentioned in Proposition 7. Let $\Sigma_\infty = U_\infty^T \Lambda_\infty U_\infty$ be the spectral decomposition of the positive semi definite (PSD) matrix $\Sigma_\infty \in \mathbb{R}^{d_u \times d_u}$. We have that

$$\begin{aligned}\hat{A}_t(M; w_{t+h}) &= \|\Lambda_\infty^{1/2}U_\infty q^M(w_{t-1}) - \Lambda_\infty^{1/2}U_\infty q_{\infty;h}^*(w_{t:t+h})\|_2^2\\ &= \|\Lambda_\infty^{1/2}U_\infty X_t z - \Lambda_\infty^{1/2}U_\infty q_{\infty;h}^*(w_{t:t+h})\|_2^2.\end{aligned}$$

Define

$$\begin{aligned}A_t &:= \Lambda_\infty^{1/2}U_\infty X_t\\ &= [w_{t-1}^T \ldots w_{t-m}^T] \otimes \Lambda_\infty^{1/2}U_\infty\end{aligned}$$

Next, we proceed to compute a box that encloses all DAP policies of interest. We have for each $i \in [m]$,

$$\begin{aligned}\|z^i\|_\infty^2 &\leq \|z^i\|_2^2\\ &= \|M^{[i]}\|_F^2\\ &\leq (d_u \wedge d_x)\|M^{[i]}\|_{op}^2\\ &\leq (d_u \wedge d_x)R^2\gamma^2,\end{aligned}$$

where the last line is due to the DAP policy set that we are interested in.

Thus the box $\mathcal{D}_\infty(R\gamma\sqrt{d_u \wedge d_x}) := \mathcal{D}_\infty(\tilde{R})$ encapsulates the DAP policy space that we are interested in.

We need to compute the parameters in Theorem 2. First, let's focus on computing G. We have for any $z_1, z_2 \in$

$$\|A_t(z_1 + z_2) - 2b_t\|_1 \leq 2\|A_t\|_1 md_ud_x\tilde{R} + 2\|b_t\|_1, \tag{14}$$

where $b_t = \Lambda_\infty^{1/2} U_\infty q^*_{\infty;h}(w_{t:t+h})$.

We have

$$\begin{aligned}\|A_t\|_1 &= \max_{i=1,\dots,m} \|w_{t-i}\|_\infty \|\Lambda^{1/2} U_\infty\|_1 \\ &\leq \|\Lambda^{1/2} U_\infty\|_1, \end{aligned} \tag{15}$$

as the disturbances obey $\|w_t\|_2 \leq 1$.

We have

$$\begin{aligned}\|b_t\|_2 &\leq \sum_{i=t}^{t+h} \|\Lambda^{-1/2} U_\infty B^T (A_{cl,\infty})^{i-t} P_\infty w_i\|_2 \\ &\leq \sum_{i=t}^{t+h} \|\Lambda^{-1/2} U_\infty B^T\|_2 \|(A_{cl,\infty})^{i-t}\|_2 \|P_\infty\|_2 \|w_i\|_2 \\ (a) &\leq \|\Lambda^{-1/2} U_\infty B^T\|_2 \|P_\infty\|_2 \sum_{i=1}^{h} \gamma^{i-1} \\ &\leq \|\Lambda^{-1/2} U_\infty B^T\|_2 \|P_\infty\|_2 \cdot \frac{1}{1-\gamma}, \end{aligned}$$

where in line (a) we used the strong stability criterion and the fact that $\|w_t\|_2 \leq 1$. Thus we have

$$\begin{aligned}\|b_t\|_1 &\leq \sqrt{d_u} \|b_t\|_2 \\ &\leq \frac{\|\Lambda^{-1/2} U_\infty B^T\|_2 \|P_\infty\|_2 \sqrt{d_u}}{1-\gamma}. \end{aligned} \tag{16}$$

Putting together Eq.(14).(15) and (16) we arrive at

$$\begin{aligned}\|A_t(z_1+z_2) - 2b_t\|_1 &\leq 2 m d_u d_x R \gamma \sqrt{d_x \wedge d_u} \|\Lambda^{1/2} U_\infty\|_1 + 2 \frac{\|\Lambda^{-1/2} U_\infty B^T\|_2 \|P_\infty\|_2 \sqrt{d_u}}{1-\gamma} \\ &:= G \end{aligned} \tag{17}$$

Next we proceed to calculate α in Theorem 2. Denote by U_j the j^{th} column of the matrix U_∞. The squared norm of the i^{th} row of the covariate matrix A_t is given by

$$\begin{aligned}\sum_{k=1}^{m} \|w_{t-k}\|_2^2 \sum_{j=1}^{d_u} \lambda_j u_j^2[i] &\leq \|\Sigma_\infty\|_{op} \sum_{k=1}^{m} \sum_{j=1}^{d_u} u_j^2[i] \\ &= m \|\Sigma_\infty\|_{op}, \end{aligned}$$

where we used the fact the matrix U_∞ is orthogonal. Thus we choose

$$\alpha = \sqrt{m \|\Sigma_\infty\|_{op}}.$$

By similar arguments used to reach Eq.(17), we choose

$$L = 4G^2$$

For a sequence of policies $M_1, \dots, M_n$, observe that $\sum_{t=2}^n \|\texttt{flatten}(M_t) - \texttt{flatten}(M_{t-1})\|_1 \leq d_x \sum_{t=2}^n \|M_t - M_{t-1}\|_1$. The last relation expresses the dynamic regret incurred by ProDR.control.delayed in terms of total variation of $\texttt{flatten}(M_t)$ to be bounded by total variation of the matrices themselves.

Putting all the constants together and applying Theorem 3 and Theorem 2 yields the Corollary.

□

Theorem 5. *There exists an LQR system, a choice of the perturbations w_t and a DAP policy class such that:*

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

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

Proof. Consider a system with matrices $A = 0 \in \mathbb{R}^{2\times 2}$, $B = \left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$, $R_x = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and $R_u = 0 \in \mathbb{R}^{2\times 2}$. In this setting $K_\infty = 0$ as per Eq.(5). We consider DAP polices (see Definition 1) with $m = 1$. Let the starting state be $x_1 = 0 \in \mathbb{R}^{2\times 2}$.

Let $y_t = \pm 1$ with probability half each. Let $w_t = [y_t, 1]^T$. For a policy that chooses a control signal u_t at time t, its next state is given by $x_{t+1} = w_t - u_t$ and $\ell_{t+1}(x_{t+1}, u_{t+1}) = (u_t[1] - y_t)^2$. Hence for any algorithm, the loss is given by:

$$\sum_{t=1}^{n} \ell_t(x_t, u_t) = \sum_{t=1}^{n-1} (u_t^{\text{alg}}[1] - y_t)^2. \tag{18}$$

Divide the time horizon into bins of width W. Let the number of bins be $M := n/W$. We assume that n/W is an integer for simplicity. Let the i^{th} be denoted by $[s_i, e_i]$ for $i \in [M]$. Define

$$a_i := \frac{1}{W} \sum_{t=s_i}^{e_i} y_t.$$

We will uniformly use the same DAP policy within a bin i as the comparator. This policy will be parameterized by the matrix $M_i := \left[\begin{smallmatrix} 0 & -a_i \\ 0 & 0 \end{smallmatrix}\right]$

By Hoeffding's inequality and a union bound across all M bins, we arrive at

$$a_i \in \left[-\sqrt{\frac{\log(nM/\delta)}{2W}}, \sqrt{\frac{\log(nM/\delta)}{2W}} \right],$$

with probability at-least $1 - \delta$. We will call this high probability event as $\mathcal{E}$. Due to symmetry we have that $P(y_t = 1|\mathcal{E}) = 1/2$. So under the event $\mathcal{E}$, the Bayes optimal online prediction of any algorithm as per Eq.(18) will be to set $u = [0, 0]^T$. So within a bin we have that

$$\sum_{t=1}^{n} E[\ell_t(x_t, u_t)|\mathcal{E}] \geq W.$$

Now we need to upper bound the cumulative loss of the comparator within a bin. Since the policy within a bin is parameterized by M_i, we have that $u_t = -M_t w_{t-1} = [a_i, 0]^T$ for all $t \in [s_i, e_i]$.

So we have:

$$\begin{aligned} E[(y_t - u_t)^2|\mathcal{E}] &= \frac{E[(y_t - u_t)^2] - E[(y_t - u_t)^2|\mathcal{E}^c]P(\mathcal{E}^c)}{P(\mathcal{E})} \\ &\leq \frac{E[(y_t - u_t)^2]}{1 - \delta}, \end{aligned}$$

where $\mathcal{E}^c$ denotes complement of event $\mathcal{E}$.

By bias variance decomposition, we have that

$$E[(y_t - u_t)^2] = 1 - 1/W.$$

So the overall regret is lower bounded by

$$\begin{aligned} \sum_{i=1}^{M} \sum_{t=s_i}^{e_i} E[(y_t - u_t^{\text{alg}}[1])^2|\mathcal{E}] - E[(y_t - a_i)^2|\mathcal{E}] &\geq \sum_{i=1}^{M} W(1 - \frac{1}{1-\delta}) + \frac{1}{1-\delta} \\ &\geq M/(1-\delta) - W\delta/(1-\delta) \\ &\geq M/2, \end{aligned} \tag{19}$$

where the last line is obtained by setting $\delta = 1/n^2$

Under the event $\mathcal{E}$ with $\delta = 1/n^2$, the total variation (TV) of the sequence $a_{1:n}$ is given by:

$$\mathcal{TV}(a_{1:n}) \leq \frac{n\sqrt{2\log(n^4)}}{W^{3/2}}.$$

Now setting $W = \frac{n^{2/3}(8\log n)^{1/3}}{C_n^{2/3}}$ we obtain $\mathcal{TV}(a_{1:n}) \leq C_n$ with probability at-least $1 - 1/n^2$.

Continuing from Eq.(19), we obtain that

$$\begin{aligned} E[R_n|\mathcal{E}] &:= \sum_{i=1}^{M}\sum_{t=s_i}^{e_i} E[(y_t - u_t^{\text{alg}}[1])^2|\mathcal{E}] - E[(y_t - a_i)^2|\mathcal{E}] \\ &\geq \frac{n^{1/3}C_n^{2/3}}{2(8\log n)^{1/3}}, \end{aligned} \tag{20}$$

where the event $\mathcal{E}$ occurs with probability at-least $1 - 1/n^2$.

When $C_n \leq 1/\sqrt{n}$, the static regret bound of $\Omega(\log n)$ (see Theorem 11.9 in Cesa-Bianchi and Lugosi (2006)). This completes the proof of the theorem.

□

Connections to online non-parametric regression framework of Rakhlin and Sridharan (2014). In the work of Rakhlin and Sridharan (2014), they study the following online regression framework (simplified here without affecting the information-theoretic rates):

- At each round t, learner plays a decision $x_t \in \mathbb{R}$.
- Nature reveals a label y_t such that $|y_t| \leq 1$.
- Learner suffers loss $(y_t - x_t)^2$.

One is interested in finding the min-max rate of regret against a non-parametric sequence class. We define the space of total variation (TV) bounded sequences as:

$$\mathcal{TV}(C_n) := \{\theta_{1:n} | \mathcal{TV}(\theta_{1:n}) \leq C_n\}.$$

Translated into the setup of Rakhlin and Sridharan (2014), one can aim to control the regret against $\mathcal{TV}(C_n)$ which is:

$$R_n := \sum_{t=1}^{n}(y_t - x_t)^2 - \inf_{\theta_{1:n} \in \mathcal{TV}(C_n)} \sum_{t=1}^{n}(y_t - \theta_t)^2. \tag{21}$$

The TV class is known to be sandwiched between two Besov spaces having the same minimax rate (see for eg. (DeVore and Lorentz, 1993)). So the results of Rakhlin and Sridharan (2014) based on characterizing the sequential Rademacher complexity of the Besov class leads to $O(n^{1/3})$ as the minimax rate of R_n wrt n. The rate wrt C_n was not provided in their work. However, we remark that they establish an $O(n^{1/3})$ upper bound also via non-constructive arguments.

In contrast, the lower bound we provided in the proof of Theorem 5 is for $\sum_{t=1}^{n} E[(y_t - u_t^{\text{alg}}[1])^2 - (y_t - a_t)^2|\mathcal{E}]$ (Eq.(20)) where $\mathcal{TV}(a_{1:n}) \leq C_n$ under the high probability event $\mathcal{E}$ trivially lower bounds R_n in Eq.(21) with high probability.