Dynamic Regret Minimization for Control of Non-stationary Linear Dynamical Systems

YUWEI LUO, Stanford University, USA
VARUN GUPTA, University of Chicago, USA
MLADEN KOLAR, University of Chicago, USA

We consider the problem of controlling a Linear Quadratic Regulator (LQR) system over a finite horizon T with fixed and known cost matrices Q, R, but unknown and non-stationary dynamics $\{A_t, B_t\}$. The sequence of dynamics matrices can be arbitrary, but with a total variation, V_T, assumed to be $o(T)$ and unknown to the controller. Under the assumption that a sequence of stabilizing, but potentially sub-optimal controllers is available for all t, we present an algorithm that achieves the optimal dynamic regret of $\widetilde{O}\left(V_T^{2/5}T^{3/5}\right)$. With piecewise constant dynamics, our algorithm achieves the optimal regret of $\widetilde{O}(\sqrt{ST})$ where S is the number of switches. The crux of our algorithm is an adaptive non-stationarity detection strategy, which builds on an approach recently developed for contextual Multi-armed Bandit problems. We also argue that non-adaptive forgetting (e.g., restarting or using sliding window learning with a static window size) may not be regret optimal for the LQR problem, even when the window size is optimally tuned with the knowledge of V_T. The main technical challenge in the analysis of our algorithm is to prove that the ordinary least squares (OLS) estimator has a small bias when the parameter to be estimated is non-stationary. Our analysis also highlights that the key motif driving the regret is that the LQR problem is in spirit a bandit problem with linear feedback and locally quadratic cost. This motif is more universal than the LQR problem itself, and therefore we believe our results should find wider application.

CCS Concepts: • **Theory of computation** → **Stochastic control and optimization**; *Online learning algorithms*; • **Mathematics of computing** → Stochastic processes.

Additional Key Words and Phrases: Linear Quadratic Regulator, dynamic regret, non-stationary learning, ordinary least squares estimator

ACM Reference Format:
Yuwei Luo, Varun Gupta, and Mladen Kolar. 2022. Dynamic Regret Minimization for Control of Non-stationary Linear Dynamical Systems. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 1, Article 9 (March 2022), 72 pages. https://doi.org/10.1145/3508029

1 INTRODUCTION

We look at the control of a Linear Quadratic Regulator (LQR) system with unknown and time-varying linear dynamics:

$$x_{t+1} = A_t x_t + B_t u_t + w_t,$$

with state $x_t \in \mathbb{R}^n$ and control $u_t \in \mathbb{R}^d$, stochastic *i.i.d.* sub-Gaussian noise process $\{w_t\}$, and a time-invariant known quadratic cost function $c(x, u) = x^\top Q x + u^\top R u$ over a horizon of T periods.

Authors' addresses: Yuwei Luo, Stanford University, 655 Knight Way, Stanford, USA, yuweiluo@stanford.edu; Varun Gupta, guptav@uchicago.edu, University of Chicago, 5807 S. Woodlawn Ave., Chicago, Illinois, USA, 60637; Mladen Kolar, mladen.kolar@chicagobooth.edu, University of Chicago, 5807 S. Woodlawn Ave., Chicago, Illinois, USA, 60637.

2476-1249/2022/3-ART9 $15.00
https://doi.org/10.1145/3508029

LQR systems are perhaps the simplest Markov Decision Processes (MDPs) and one of the most fundamental problems studied in control theory. To quote [37, Chapter 8], "one of the most powerful applications of time-varying LQR involves linearizing around a nominal trajectory of a nonlinear system and using LQR to provide a trajectory controller." More precisely, given a desired trajectory x_t^0, u_t^0 that one desires to track for a system with non-linear dynamics:

$$\mathbf{E}[x_{t+1} \mid x_t, u_t] = x_t + f(x_t, u_t),$$

we define the centered trajectories $\bar{x}_t = x_t - x_t^0, \bar{u}_t = u_t - u_t^0$, so that:

$$\begin{aligned}\mathbf{E}[\bar{x}_{t+1} \mid \bar{x}_t, \bar{u}_t] &= \bar{x}_t + f(x_t, u_t) - f(x_t^0, u_t^0) \\ &\approx \bar{x}_t + \frac{\partial f(x_t^0, u_t^0)}{\partial x_t^0}(x_t - x_t^0) + \frac{\partial f(x_t^0, u_t^0)}{\partial u_t^0}(u_t - u_t^0) =: A_t \bar{x}_t + B_t \bar{u}_t.\end{aligned}$$

See also [3] for a tutorial treatment of use of LQR in engineering design. LQR systems, and linear dynamical systems more broadly, have been used to model diverse applications, such as controlling robots [30], cooling data centers [12], control of brand dynamics in marketing [32], and macroeconomic policy [11] to name a few. As a result, LQR systems have also been the subject of a lot of research on reinforcement learning: from model-free vs. model-based approaches in episodic learning setting, to learning and control under unknown stationary dynamics, to robust control in the presence of an adversarial (non-stochastic) noise process. See related work in Section 2. The ability to adapt to changing dynamics lends another, arguably stronger, robustness to the control policy. However, to the best of our knowledge, the problem of learning non-stationary dynamics while controlling an LQR system has not been studied yet. We take the first steps towards this problem.

We quantify the non-stationarity of the sequence $\{\Theta_t = [A_t\ B_t]\}$ by the total variation $V_T = \sum_{t=1}^{T-1} \Delta_t$ with $\Delta_t := \|\Theta_{t+1} - \Theta_t\|_F$ denoting the Frobenius norm of change of dynamics matrix A and input matrix B from time t to $t+1$. In the case of piecewise constant dynamics, we measure the non-stationarity by the number of pieces $S_T \geq 1$.

We measure the performance of a control (and learning) policy π via dynamic regret metric:

$$\mathcal{R}^\pi(T) = \sum_{t=1}^{T} c(x_t, u_t) - J_t^*, \tag{1}$$

where u_t denotes the action taken by policy π, and J_t^* denotes the optimal average steady-state cost of the *stationary* LQR system with dynamics fixed as Θ_t. We also show that $\sum_t J_t^*$ is at most $O(V_T)$ larger than the expected cost of the dynamic optimal policy. A fundamental result in the theory of LQR systems states that the optimal policy for an LQR system is a linear feedback control policy $u_t = K_t x_t$ for some sequence of matrices K_t (see, e.g., [4]). If the LQR system is stationary, then the infinite horizon optimal policy satisfies $K_t = K^*$. Our central result states that, given access to a nominal sequence of controllers that are potentially sub-optimal but are guaranteed to stabilize the non-stationary LQR dynamics, the proposed algorithm Dyn-LQR guarantees:

$$\mathbf{E}\left[\mathcal{R}^{\text{Dyn-LQR}}(T)\right] = \widetilde{O}\left(V_T^{2/5} T^{3/5}\right),$$

without the knowledge of V_T upfront. We also demonstrate an instance showing that this regret rate is tight for any online learner/controller. The same algorithm guarantees $\mathbf{E}\left[\mathcal{R}^{\text{Dyn-LQR}}(T)\right] = \widetilde{O}\left(\sqrt{ST}\right)$ when the dynamics are piece-wise constant with at most S switches. The dependence of

the regret on the dimensions n, d for our algorithm and analysis is n^2d^2, but we believe this can be improved with a better choice of the tuning parameters in our algorithm.[1]

The design philosophy behind our algorithm Dyn-LQR is of using *certainty equivalent controllers*, that is, using the controller based on a point estimate of the model parameter (as opposed to confidence ellipsoids, for example). At a typical time t, Dyn-LQR employs a linear feedback control $\widehat{K}_t$ based on an estimate $\widehat{\Theta}_t$ of the current dynamics, with some extra exploration noise: $u_t = \widehat{K}_t x_t + \sigma_t \eta_t$. Here $\eta_t \sim \mathcal{N}(0, I_d)$, and σ_t denotes the "exploration energy." A fairly simple regret decomposition lemma shows that if the policies $\widehat{K}_t$ do not change very often, then the regret is dominated by *(i)* the total exploration energy $\sum_t \sigma_t^2$, and *(ii)* $\sum_t J_t(\widehat{K}_t) - J_t^*$, where $J_t(\widehat{K}_t)$ denotes the average steady-state cost of the *stationary* LQR system with time-invariant dynamics Θ_t and control $\widehat{K}_t$. A result of [35] shows that $J_t(\widehat{K}_t) - J_t^* \lessapprox C \cdot \left\| \widehat{\Theta}_t - \Theta_t \right\|_F^2$, *if* the estimation error is small enough. Thus, if we strip away the complexity introduced due to the dynamics itself, the essence of the non-stationary LQR problem is that of tracking Θ_t, which boils down to *a bandit problem with linear feedback and a locally quadratic loss function.* In Section 9 we give an example of a queueing system which also exhibits this motif, and for which we believe a similar algorithm as Dyn-LQR can give optimal dynamic regret.

Under non-stationary dynamics, it is important to forget the distant history when constructing an estimate of the current dynamics. Our approach for doing so is to adaptively restart the learning problem when "sufficient" change in the dynamics has accumulated, using a scheme motivated by the algorithm of Chen et al. [8] developed for contextual multi-armed bandits. The algorithm of Chen et al. [8] runs multiple tests in parallel, each tailored to detect changes of a different scale, by replaying (with carefully tailored probabilities) an older strategy and then comparing the new estimated reward distribution with the older reward distribution. As a result, Chen et al. [8] were the first to obtain the optimal dynamic regret for contextual bandit problems as a function of the total variation of the reward distribution *without the knowledge of the variation budget.* For the LQR problem, we modify this procedure in at least two directions. First, we keep using the current controller but inject a higher exploration noise. This change is critical for our regret analysis at two places: our current analysis includes a term involving the number of policy switches and minimizing the number of policy switches impacts the regret guarantee; and, we mention below, our analysis of the estimation error of dynamics crucially relies on the linear feedback control matrix being fixed throughout the interval of estimation. Second, the probabilities with which the exploration is carried out are different for the LQR problem owing to the quadratic cost. More recently, the authors in [39] outline that for many classes of episodic reinforcement learning problems, a similar strategy can be used to convert any Upper Confidence Bound (UCB) type stationary reinforcement learning algorithm to a dynamic regret minimizing algorithm. There are quite a few differences between [39] and our work: the LQR problem is not covered by the classes of MDPs they consider, we look at a non-episodic version of the LQR problem, and our algorithm is certainty equivalent controller-based and not a UCB-type.

Technical challenges and novelty: We next point out three areas where the analysis in the current paper contributes to the existing literature on online learning and control.

(1) *Ordinary Least Squares (OLS) under non-stationarity:* The biggest challenge we overcome is to prove a bound on the error of the estimated parameters $\widehat{\Theta}_t$. In particular, based on the

[1]For stationary LQR, [35] prove that the optimal dependence is $d\sqrt{n}$, we leave the task of achieving the same dependence in non-stationary LQR as a question for subsequent research.

observations in some interval $\mathcal{I}$, the OLS estimate $\widehat{\Theta}_{\mathcal{I}}$ of the dynamics is given by:

$$\widehat{\Theta}_{\mathcal{I}} = \operatorname*{argmin}_{\Theta} \sum_{t \in \mathcal{I}} \left\| x_{t+1} - \Theta (x_t^\top \; u_t^\top)^\top \right\|^2 = \operatorname*{argmin}_{\Theta} \sum_{t \in \mathcal{I}} \left\| (\Theta_t - \Theta) \cdot (x_t^\top \; u_t^\top)^\top + w_t \right\|^2 .$$

A linear feedback controller $u_t = K_{\mathcal{I}} x_t$, with $K_{\mathcal{I}}$ fixed during the interval $\mathcal{I}$, allows estimating the component of Θ_t parallel to the n-dimensional column space of $[x_t^\top \; u_t^\top] = [I_n \; K_{\mathcal{I}}^\top]^\top x_t$, but not in the orthogonal subspace. This problem shows up even in stationary LQR, and is the reason we use the exploration noise $\sigma_t \eta_t$ in u_t. However, for stationary LQR, this is only a mild problem – the estimate is unbiased by default and the condition number of the (ill-conditioned) Hessian is sufficient to bound the variance of the OLS estimator. Under non-stationary Θ_t, even proving that the OLS estimate $\widehat{\Theta}_{\mathcal{I}}$ is "unbiased," i.e., close to Θ_t for $t \in \mathcal{I}$ even when all the Θ_t in $\mathcal{I}$ are close to each other, is not trivial. Naively using the condition number of the Hessian would require a larger σ_t, and, thus, result in a suboptimal regret. A major chunk of the technical analysis is to show that a small exploration cost is sufficient to guarantee that $\widehat{\Theta}_{\mathcal{I}}$ has small bias. This requires quite a delicate analysis of the geometry of the Hessian, as well as an interplay with the algorithm itself where we need to keep the policy $K_{\mathcal{I}}$ fixed so that the column space of $[I_n \; K_{\mathcal{I}}^\top]^\top$ is fixed. This is where we crucially take advantage of the fact that instead of replaying an old policy as in [8] to detect non-stationarity, we continue playing the same linear feedback controller and only increase the exploration noise.

(2) *Continuous and unbounded state space:* The second challenge comes from the fact that the LQR system has unbounded state space. A particular complication this creates is that the certainty equivalent controller need not stabilize the dynamics under non-stationarity, and therefore the norm of the state can blow up. Algorithmically, we solve this problem by falling back on the nominal sequence of controllers when the norm of the state crosses a threshold, and until it falls below another threshold. Analytically, this requires some careful analysis to bound the total cost incurred during such phases.

(3) *An impossibility result for non-adaptive restart algorithms:* We prove a novel regret lower bound that outlines a shortcoming of a popular strategy for non-stationary bandits/reinforcement learning. As we mentioned earlier, to forget distant history for non-stationary bandits and episodic reinforcement learning, almost all existing algorithms restart learning at a fixed schedule, or use sliding window based estimators with a fixed window size. For all the flavors of non-stationary bandit or reinforcement learning problems studied in the literature, this strategy yields the optimal regret *if the window size is tuned optimally with the knowledge of the variation budget, or using a bandit-on-bandit technique.* In Theorem 8.3 we prove that for the non-stationary LQR problem, for a wide class of fixed window size based algorithms, this approach can not give the optimal regret rate even with the knowledge of V_T. This crucially uses the fact that the LQR problem behaves like a bandit problem with non-linear (in particular quadratic) loss function. We believe that the same lower bound should extend to non-linear bandit problems more generally.

Paper Outline: We survey some of the relevant literature in Section 2. In Section 3, we first present some classical results on control of stationary LQR and recent results on learning and control. Then in Section 4 we present the model assumptions for the non-stationary LQR problem that is the subject of our study. In Section 5, we present our proposed algorithm Dyn-LQR. We devote Section 6 to highlighting the technical challenge in studying the error of the OLS estimator for non-stationary LQR. In Section 7 we present the regret upper bound for Dyn-LQR, and in Section 8 we present two lower bound results.

Notation: All vectors are column vectors. For a matrix A, we use $\|A\| = \sup_{\|x\|=1} \|Ax\|$ to denote the operator norm and $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ to denote the Frobenius norm. For two square matrices A, B, we use $A \preccurlyeq B$ to denote that the matrix $B - A$ is positive semidefinite. The $O()$ notation will used to suppress problem dependent constants, including the dimensions d, n; the $\widetilde{O}()$ notation further suppresses polylog T factors.

2 RELATED WORK

Our work touches on many themes in online learning and control. For each, we mention only a few papers relevant to the present work and make no attempt to present an exhaustive survey.

Learning and control of stationary LQR:. The study of learning and control of LQRs was initiated in Abbasi-Yadkori and Szepesvári [2], who presented an $O(\sqrt{T})$ regret algorithm based on the Optimism in the Face of Uncertainty (OFU) principle, but with an exponential dependence on the dimensionality of the problem. Ibrahimi et al. [27] improved dependence on the dimensionality to polynomial. Cohen et al. [13] was the first paper that provided a computationally efficient algorithm with $O(\sqrt{T})$ regret for the stationary LQR problem by solving for the optimal steady-state covariance of $[x_t^\top u_t^\top]$ via a semi-definite program and extracting a controller from this covariance. Faradonbeh et al. [17] and Mania et al. [31] proved that the certainty equivalent controller is efficient and yields $O(\sqrt{T})$ regret. Simchowitz and Foster [35] proved a matching upper and lower bound on the regret of the stationary LQR problem of $\widetilde{\Theta}(\sqrt{nd^2T})$, settling the open question of whether logarithmic regret may be possible for LQR (due to the strongly convex loss function). Notably, the upper bound in Simchowitz and Foster [35] was achieved by a variant of the certainty equivalent controller. Cassel et al. [7] proved an $\Omega(\sqrt{T})$ lower bound and showed that naive exploration based algorithms can indeed attain logarithmic regret when the problem is sufficiently non-degenerate. [28] developed a certainty equivalent controller based strategy for stationary LQR, but allow the controller to change arbitrarily quickly, rather than according to a fixed doubling schedule as in prior work.

Dynamic regret minimization for experts and bandits: Due to the weakness of static regret as a metric for environments with non-stationary or adversarial losses/rewards, numerous stronger notions of regret have been proposed and studied. One of the first such results was in the seminal paper of Zinkevich [42], where a regret parameterized by the total variation of the comparator sequence of actions was proved. Herbster and Warmuth [26] proposed the FixedShare algorithm for prediction with expert advice problem, where the best expert may switch during the time horizon. Hazan and Seshadhri [25] looked at online convex optimization with changing loss functions, and proposed a metric for adaptive regret, defined to be the maximum over all windows of the regret of the algorithm on that window compared to the best fixed action for that window. Daniely et al. [14] introduced a metric of strongly adaptive regret and proved that no algorithm can be strongly adaptive in the bandit feedback setting. For the bandit setting, the most common approach towards dynamic regret is to assume that the non-stationary sequence has bounded total variation, and providing min-max regret guarantees as a function of the variation, e.g., Besbes et al. [5]. The common design technique is to use periodic restarts or discounting with the knowledge of the variation of rewards, e.g., [20, 34], or a bandit-on-bandit technique to learn the optimal window size as in Cheung et al. [9], but with a suboptimal regret guarantee. A recent breakthrough was achieved by the algorithm of Chen et al. [8], which performs a very delicate exploration and uses an adaptive restart argument to attain the optimal regret rate for contextual multi-armed bandits without any prior knowledge of the variation.

Reinforcement learning for non-stationary MDPs: While there is some literature on regret minimization for MDPs with fixed transition kernel, but a changing sequence of cost functions [33, 40], the work on unknown non-stationary dynamics is much more recent [10, 19]. The main idea is to use sliding window based estimators of the transition kernel and design a policy based on an optimistic model of the transition dynamics within the confidence set. As we mentioned earlier, sliding window based algorithms are provably regret-suboptimal for the LQR problem due to the quadratic cost function. In parallel with this work, [39] proposed an adaptive restart approach for non-stationary reinforcement learning that uses any UCB-type algorithm for stationary reinforcement learning as a black box. The authors show that for many tabular or linear MDP settings, their approach gives the state-of-the-art regret without knowledge of variation of the input instance. While the LQR problem is neither tabular nor linear, our approach is similar in its spirit to [39] – however, we use point estimates and explicit exploration instead of using a UCB-like approach.

Robust control of LQR under adversarial noise: While we consider the robust control of LQR systems from the perspective of changing transition dynamics, there have been some recent results on robust control of LQR when the noise w_t is adversarial. Hazan et al. [24] considered a "stationary" LQR system with known A, B, but with adversarial noise, and proposed an algorithm with $O(T^{2/3})$ regret against the best linear controller in hindsight. Simchowitz et al. [36] looked at the same problem when the A, B matrices may or may not be known, and proposed a Disturbance Feedback Control based online control policy with sublinear regret against all stabilizing policies. Finally, Goel and Hassibi [21], Gradu et al. [22] looked at non-stationary LQR problems with adversarial noise. Goel and Hassibi [21] assumed that the sequence A_t, B_t is known upfront and proposed a controller with optimal dependence of regret on the total noise. Gradu et al. [22] assumed that the dynamics matrices A_t, B_t are observed after the action u_t is taken and proposed a policy with strongly adaptive regret guarantee. Finally, we would like to point to [6] as a recent example of a work on learning and control of non-stationary non-linear dynamical systems, although in this work the non-stationary dynamics are linearly parameterized by a known non-stationary sequence of basis matrices and an unknown stationary parameter.

3 PRELIMINARIES – STATIONARY LQR

In this section, we give a brief summary of the classical theory of stationary LQR systems and some recent work on learning and control for stationary LQR systems that lays the groundwork for our work on non-stationary LQR. The stationary dynamics, parameterized by $\Theta = [A\ B]$, are given by:

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t \in [T],$$

and the cost function by:

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

where $x_t \in \mathbb{R}^n$ denotes the state, $u_t \in \mathbb{R}^d$ the control (or input), w_t are i.i.d. stochastic noise (disturbance) with covariance matrix W, and Q, R are positive-definite matrices.

A classical result in the theory of LQR problems is that the value function of the LQR problem is a quadratic function of the state. This is true even for non-stationary dynamics and can be most easily seen by solving for the optimal control for a finite horizon problem via backward Dynamic Programming. As a consequence, the optimal controller turns out to be a linear feedback controller $u_t = K_t x_t$, for some sequence of control matrices $\{K_t\}$. In the special case of infinite horizon average cost minimization, the control is stationary with $K_t = K^*$. For an arbitrary linear feedback controller K that is stabilizing, i.e., the spectral radius of $A + BK$ is upper bounded away from 1, we denote by $J(\Theta, K)$ the infinite horizon average cost and by the symmetric positive definite matrix $P(\Theta, K)$ we denote the quadratic *relative value function* (also called the *bias function*) for the infinite horizon

average cost problem, satisfying the following Bellman equation:

$$x^\top P(\Theta,K)x = c(x,Kx) - J(\Theta,K) + \mathbf{E}\left[x_1^\top P(\Theta,K)x_1 | x_0 = x\right]$$
$$= x^\top (Q + K^\top RK)x - J(\Theta,K) + x^\top (A+BK)^\top P(\Theta,K)(A+BK)x + \mathbf{E}\left[w^\top P(\Theta,K)w\right].$$

Matching the quadratic and the constant terms, we get that $P(\Theta,K)$ solves the following equation

$$P = Q + K^\top RK + (A+BK)^\top P(A+BK)$$

and $J(\Theta,K) = \mathrm{Tr}(P(\Theta,K)W)$. Let the optimal bias function be denoted by $P^*(\Theta)$ and the optimal linear feedback controller by $K^*(\Theta)$. Given $P^*(\Theta) = P^*$, the optimal linear feedback controller $K^* = K^*(\Theta)$ can be obtained by solving for the cost minimizing action in the Bellman equation:

$$K^* = -(R + B^\top P^* B)^{-1} B^\top P^* A. \tag{2}$$

Plugging the above in the equation for $P(\Theta,K)$ gives a fixed point equation (called the Discrete Algebraic Ricatti Equation) for $P^*(\Theta)$:

$$P^* = Q + A^\top P^* A - A^\top P^* B(R + B^\top P^* B)B^\top P^* A. \tag{3}$$

While the explicit forms of $K^*(\Theta), P^*(\Theta)$ are not essential for following the results in the paper, we would like to point out that neither of them depend on the covariance of the noise process, even though the optimal cost $J^*(\Theta)$ does.

Finally, consider the policy $u_t = Kx_t + \sigma\eta_t$, where η_t are *i.i.d.* with covariance I_d and $\sigma > 0$. Denote the average cost for this policy by $J(\Theta,K,\sigma)$ and the relative value function by $P(\Theta,K,\sigma)$. Then,

$$P(\Theta,K,\sigma) = P(\Theta,K),$$
$$J(\Theta,K,\sigma) = J(\Theta,K) + \sigma^2 \,\mathrm{Tr}\left(R + B^\top P(\Theta,K)B\right). \tag{4}$$

That is, the effect of additive noise in the controller completely decouples from the cost of the noiseless control Kx_t.

Cost of model estimation error: The following lemma from [35] will be central for the intuition and analysis behind learning and control of LQR.

Lemma 3.1 (Simchowitz and Foster [35, Theorem 5]). *Let $\Theta = [A\ B]$ be a stabilizable system and $\widehat{\Theta} = [\widehat{A}\ \widehat{B}]$ be an estimate of Θ. Then there exist constants C_1, C_2, depending on R, Q, W, such that if $\max\left\{\left\|A - \widehat{A}\right\|, \left\|B - \widehat{B}\right\|\right\} \le C_1 \|P^*(\Theta)\|^{-5}$, then*

$$J^*(\Theta) - J(\Theta, K^*(\widehat{\Theta})) \le C_2 \|P^*(\Theta)\|^8 \left\|\Theta - \widehat{\Theta}\right\|_F^2.$$

The lemma implies that the certainty equivalent controller $K^*(\widehat{\Theta})$ based on the estimate $\widehat{\Theta}$ with sufficiently small error ϵ leads to a suboptimality of at most a problem-dependent constant times ϵ^2. Note that the closer the spectral norm of the closed loop $A + BK^*(\Theta)$ is to 1, the larger is $\|P^*(\Theta)\|$, and the harder it is to satisfy the condition in Lemma 3.1.

A naive exploration algorithm: To get some intuition on the fundamental exploration-exploitation trade-off for the LQR problem, we describe a bare bones version of the algorithm from [35] for the stationary setting. The authors assume (as is common in the literature) access to a stabilizing, but suboptimal controller K_0. The algorithm begins by playing $u_t = K_0 x_t + \sigma_0 \eta_t$ with $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ and $\sigma_0^2 = 1$ for a sufficiently long warm-up period L. Based on this warm-up period, an initial estimate $\widehat{\Theta}_1$ is constructed using the ordinary least squares (OLS) estimator. The quantity σ_0^2 denotes the *exploration noise/energy*. Even though the LQR dynamics adds *i.i.d.* noise w_t to the

state, the exploration noise $\sigma_0^2\eta_t$ is necessary because the vector $[x_t^\top\ u_t^\top]^\top = [I_n\ K_0^\top]^\top x_t$ lives in an n-dimensional subspace instead of the full $(n+d)$-dimensional subspace. The algorithm then proceeds in blocks of doubling length, indexed by $i = 1, 2, \ldots$. Block i is of length $\tau_i = L \cdot 2^i$. In block 1, the control is chosen as $u_t = K_1 x_t + \sigma_1 \eta_t$ where $K_1 = K^*(\widehat{\Theta}_1)$ and $\sigma_1^2 = 1/\sqrt{\tau_1}$. The observations from block 1 are used to construct an estimate $\widehat{\Theta}_2$ and the control in block 2 is $u_t = K_2 x_t + \sigma_2 \eta_2$ with $K_2 = K^*(\widehat{\Theta}_2)$ and $\sigma_2^2 = 1/\sqrt{\tau_2}$. More generally, observations from block $(i-1)$ are used to create an estimate $\widehat{\Theta}_i$ and controller $K_i = K^*(\widehat{\Theta}_i)$. The control in block i is $u_t = K_i x_t + \sigma_i \eta_t$, with exploration noise $\sigma_i^2 = 1/\sqrt{\tau_i}$. The intuition behind the choice of exploration noise is the following. The total exploration energy invested in block i is $\tau_i \sigma_i^2$, which, by (4), increases the cost by an order $\tau_i \sigma_i^2$. Furthermore, the variance of the OLS estimator $\widehat{\Theta}_{i+1}$ is inversely proportional to the exploration noise, and is therefore $O(1/\tau_i \sigma_i^2)$. Lemma 3.1 then says that the per step exploitation cost from using controller K_{i+1} based on $\widehat{\Theta}_{i+1}$ is of the order $1/\tau_i \sigma_i^2$. Therefore, the total regret is of order $1/\sigma_i^2$ during block $(i+1)$. Balancing the exploration cost $\tau_i \sigma_i^2$ during block i and the total exploitation cost $1/\sigma_i^2$ during block $i+1$ gives the choice $\sigma_i^2 \approx \tau_i^{-1/2}$.

4 MODEL AND PRELIMINARIES – NON-STATIONARY LQR

The non-stationary LQR problem has dynamics:

$$x_{t+1} = A_t x_t + B_t u_t + w_t, \quad t \in [T],$$

and time-invariant cost function:

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

where $x_t \in \mathbb{R}^n$ denotes the state, $u_t \in \mathbb{R}^d$ the control (or input), $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, W)$ denotes the stochastic noise (disturbance) with covariance matrix $W = \psi^2 I_n$ (the assumption on w_t is for exposition purposes; our results readily extend to sub-Gaussian w_t with $\psi^2 I_n \preccurlyeq W \preccurlyeq \Psi^2 I_n$ for $0 < \psi < \Psi < \infty$). We use $\{\mathcal{F}_t\}_{t\in[T]}$ to denote the filtration generated by $\{w_1, \ldots, w_T\}$. We will use $\Theta_t = [A_t\ B_t]$ to succinctly denote the dynamics of the LQR at time period t. Cost matrices Q, R are assumed to be symmetric positive definite with $r_{\min} I_d \preccurlyeq R \preccurlyeq r_{\max} I_d$, $q_{\min} I_n \preccurlyeq Q \preccurlyeq q_{\max} I_n$.

The learner/controller knows the cost matrices Q, R, but not the dynamics $\{\Theta_t\}_{t\in[T]}$. For any interval $\mathcal{I} = [s, e]$, we define the total variation of the model parameter within the interval as

$$\Delta_{\mathcal{I}} = \Delta_{[s,e]} := \sum_{t=s}^{e-1} \Delta_s = \sum_{t=s}^{e-1} \|\Theta_s - \Theta_{s+1}\|_F,$$

so that the total variation $V_T = \Delta_{[1,T]}$. In the case of piecewise constant dynamics, we use $S_{\mathcal{I}} \geq 1$ to denote the number of such constant dynamics pieces in interval $\mathcal{I}$.

A common assumption in the literature on online learning and control of stationary LQR systems is the availability of a baseline controller K_0 that may be suboptimal, but stabilizes the system. Such a controller can be played in an initial warm-up phase until a good initial estimate of the dynamics can be learned. This assumption allows one to focus on the algorithmic challenge of minimizing regret and not worry about the stability of the system. From the point of view of applications, often there are default actions which guarantee this condition (e.g., shutting a data center will prevent over-heating of servers), or crude forecasts of the dynamics may be enough to derive such controls. Theoretically, a stabilizing controller can be found by following the strategy proposed in [16]. Similarly, we also assume that our algorithm is given a sequence of controllers $\{K_t^{\text{stab}}\}$ that stabilizes the dynamics given by $\{\Theta_t\}$. More formally, Assumption 4.2 states that the exogenous sequence of controllers satisfies a property called *sequentially strong stability*.

Definition 4.1 (Sequentially Strong Stability [12]). For the non-stationary LQR problem with parameters $\{\Theta_t\} = \{[A_t\ B_t]\}$, a sequence of controllers $\{K_1, \ldots, K_T\}$ is called (κ, γ) sequentially strongly-stabilizing (for $\kappa \geq 1$ and $0 < \gamma \leq 1$) if there exist matrices $H_1, H_2, \ldots, H_T \succ 0$ and $L_1, L_2, \ldots, L_T$ such that $A_t + B_t K_t = H_t L_t H_t^{-1}$ for all $t \in [T]$, and the following properties hold:

(i) $\|L_t\| \leq 1 - \gamma$ and $\|K_t\| \leq \kappa$ for $t \in [T]$;
(ii) $\|H_t\| \leq B_0$ and $\left\|H_t^{-1}\right\| \leq 1/b_0$ with $\kappa = B_0/b_0$ for $t \in [T]$;
(iii) $\left\|H_{t+1}^{-1} H_t\right\| \leq 1 + \gamma/2$ for $t \in [T-1]$.

Assumption 4.2. The online algorithm has access to a sequence of (κ, γ) sequentially strongly-stabilizing controllers $\{K_1^{\text{stab}}, K_2^{\text{stab}}, \ldots, K_T^{\text{stab}}\}$, for constants $\kappa \geq 1$ and $0 < \gamma \leq 1$.

A (κ', γ') sequentially strongly stabilizing sequence of controllers is also (κ, γ) strongly stabilizing for $\kappa \geq \kappa'$ and $\gamma \leq \gamma'$. Therefore, we take $\kappa \geq 1$ as a convenient convention. An intuitive explanation for this assumption is the following. Denote

$$\Phi_t := A_t + B_t K_t^{\text{stab}} \qquad \text{and} \qquad \Phi_{b:a} := \Phi_b \Phi_{b-1} \cdots \Phi_a, \quad \text{for } 1 \leq a \leq b \leq T.$$

Then

$$\begin{aligned} \|\Phi_{b:a}\| &= \left\|H_b L_b (H_b^{-1} H_{b-1}) L_{b-1} \cdots (H_{a+1}^{-1} H_a) L_a H_a^{-1}\right\| \\ &\leq \|H_b\| \cdot \|L_b\| \cdot \left\|H_b^{-1} H_{b-1}\right\| \cdot \|L_{b-1}\| \cdots \left\|H_{a+1}^{-1} H_a\right\| \cdot \|L_a\| \cdot \left\|H_a^{-1}\right\| \\ &\leq \kappa \left(1 + \frac{\gamma}{2}\right)^{b-a} (1-\gamma)^{b-a+1} \ \leq\ \kappa \left(1 - \frac{\gamma}{2}\right)^{b-a}. \end{aligned} \tag{5}$$

As a consequence of (5) and noting that

$$x_b = \Phi_{b-1:a} x_a + \Phi_{b-1:a+1} w_a + \Phi_{b-1:a+2} w_{a+1} + \cdots + \Phi_{b-1:b-1} w_{b-2} + w_{b-1},$$

we can bound the norm of the state under the stabilizing controllers as:

$$\|x_b\| \leq \kappa e^{-\gamma(b-a)/2} \|x_a\| + \frac{2\kappa}{\gamma} \max_{a \leq t \leq b-1} \|w_t\|,$$

and,

$$\begin{aligned} \mathbf{E}\left[\|x_b\|^2\right] &= \|\Phi_{b-1:a} x_a\|^2 + \mathbf{E}\left[\|\Phi_{b-1:a+1} w_a\|^2\right] + \cdots + \mathbf{E}\left[\|\Phi_{b-1:b-1} w_{b-2}\|^2\right] + \mathbf{E}\left[\|w_{b-1}\|^2\right] \\ &\leq\ \kappa^2 e^{-\gamma(b-a)} \|x_a\|^2 + \frac{2\kappa^2}{\gamma} \mathbf{E}\left[\|w\|^2\right]. \end{aligned} \tag{6}$$

While assuming $\|\Phi_t\| < 1 - \gamma/2$ also ensures (5), it is a much more restrictive condition. A weaker condition is that the spectral radius is bounded: $\rho(\Phi_t) \leq 1 - \gamma/2$, but the spectral radius is not submultiplicative and does not imply (5).

A second assumption we will make is on the stability of the controller derived from an *accurate* estimate of the true dynamics.

Assumption 4.3. For any $t \in [T]$, let Θ_t be the true dynamics, $\widehat{\Theta}_t$ be an estimate of the true dynamics, and $\widehat{K} = K^*(\widehat{\Theta}_t)$ be the optimal closed-loop controller for the estimated dynamics. Then, there exist constants C_3, C_4 such that $\left\|\widehat{\Theta}_t - \Theta_t\right\|_F^2 \leq C_3$ implies $J^*(\Theta_t) - J(\Theta_t, \widehat{K}) \leq C_4 \left\|\Theta_t - \widehat{\Theta}_t\right\|_F^2$. For convenience, we assume $C_3 \leq 1$, since the assumption continues to hold if we choose a smaller value of C_3 than sufficient.

Assumption 4.3 is without loss of generality due to Lemma 3.1. As mentioned earlier, the constants C_3, C_4 depend on the maximum operator norm of P_t^*, which we assume to be bounded independent of T and V_T. The constants C_3, C_4 are only used in the analysis, not as a part of the algorithm.

Just like Assumption 4.2, under non-stationary dynamics, we need a stronger sequential stability property for controllers $K^*(\widehat{\Theta}_t)$ than in Assumption 4.3. Towards that end, we introduce a strengthening of the (κ, γ) sequential strong stability criterion. The main difference is that condition *(iii)* involves the variation $\|\Theta_{t+1} - \Theta_t\|$ and hence allows us to prove exponential stability for non-stationary dynamics with small total variation.

Definition 4.4 ((κ, γ, ν)-Sequentially Strong Stability). For the non-stationary LQR problem and an interval $[a, b]$, a sequence of controllers $\{K_a, \ldots, K_b\}$ is called (κ, γ, ν)-sequentially strongly-stabilizing (for $\kappa \geq 1$ and $0 < \gamma \leq 1$) if there exist matrices $H_a, H_{a+1}, \ldots, H_b \succ 0$ and $L_a, L_{a+1}, \ldots, L_b$ such that $A_t + B_t K_t = H_t L_t H_t^{-1}$ for all $t \in [a, b]$, and the following properties hold:

(i) $\|L_t\| \leq 1 - \gamma$ and $\|K_t\| \leq \kappa$ for $t \in [a, b]$;
(ii) $\|H_t\| \leq B_0$ and $\left\|H_t^{-1}\right\| \leq 1/b_0$ with $\kappa = B_0/b_0$ for $t \in [a, b]$;
(iii) $\left\|H_{t+1}^{-1} H_t\right\| \leq 1 + \nu \cdot \|\Theta_{t+1} - \Theta_t\|$ for $t \in [a, b-1]$.

The next lemma states that if the provided estimate $\widehat{\Theta}$ satisfies $\|\widehat{\Theta} - \Theta_t\|_F^2 \leq C_3$ for all t in an interval $\mathcal{I}$, then the controller $\widehat{K} = K(\widehat{\Theta})$ is (κ, γ, ν)-sequentially strongly stable for the dynamics in $\mathcal{I}$.

Lemma 4.5. *For an interval $\mathcal{I}$, let $\widehat{\Theta}$ be an estimate of the dynamics such that $\|\widehat{\Theta} - \Theta_t\|_F^2 \leq C_3$ for all $t \in \mathcal{I}$. Let $\widehat{K} = K^*(\widehat{\Theta})$ be the optimal linear feedback controller with respect to the estimate $\widehat{\Theta}$. Define*

$$\nu = \frac{2(1-\gamma)^2}{1-(1-\gamma)^2}\left((1-\gamma) + (\kappa+1)\right).$$

Then $\widehat{K}$ is a (κ, γ, ν)-sequentially strongly stable control sequence for interval $\mathcal{I}$ with the following setting of parameters: $H_t = P_t^{1/2}$ and $L_t = P_t^{-1/2}\left(A_t + B_t\widehat{K}\right)P_t^{1/2}$, where $P_t := P(\Theta_t, \widehat{K})$, $\kappa = \sqrt{\frac{\widetilde{J}_{\mathcal{I}}^*}{\psi^2 r_{\min}}}$, $\gamma = \frac{q_{\min}\psi^2}{2J_{\mathcal{I}}^*}$, $J_{\mathcal{I}}^* = \max_{t\in\mathcal{I}} J^*(\Theta_t)$, and $\widetilde{J}_{\mathcal{I}}^* = J_{\mathcal{I}}^* + C_3C_4$.

As a corollary, similar to the calculations in (6), the following lemma bounds the norm of x_t.

Lemma 4.6. *Let the controller $\widehat{K}$ and interval $\mathcal{I} = [s_{\mathcal{I}}, e_{\mathcal{I}}]$ satisfy the conditions in Lemma 4.5. Then for an action sequence $u_t = \widehat{K}x_t + \sigma_t\eta_t$, $t \in \mathcal{I}$, there exists a constant C_{ss} such that*

$$\|x_t\| \leq \kappa e^{-\gamma(t-1)+C_{ss}V_{[1,t-1]}}\|x_1\| + \frac{\kappa e^{-\gamma(t-s)+C_{ss}V_{[s,t-1]}}}{\gamma}\max_{1<s<t}\|w_s + \sigma_s B_s\eta_s\|, \quad t \in \mathcal{I}.$$

Later we will see that the controllers used in our proposed Algorithm 1 satisfy the conditions of Lemmas 4.5 and 4.6, and hence stabilize the dynamics and the state has bounded norm with high probability.

Finally, we introduce some constants that we will use as a parameterization of the input instance. We assume that they are known to the learner/controller.

Additional Constants: Let the norm upper bounds for the parameters of the instance be given by: $A_u = \max_{t\in[T]}\|A_t\|$, $B_u = \max_{t\in[T]}\|B_t\|$, $\Theta_u = \max_{t\in[T]}\|\Theta_t\|$, and $P_u = \max_{t\in[T]}\left\|P_t^*\right\|$. Define

$$\beta := \max\left\{\psi, \max_{i,t}\beta_{i,t}\right\},$$

where $\beta_{i,t}$ are singular values of B_t. Define K_u as:

$$K_u = \max_t \left\{ \left\| K_t^{\text{stab}} \right\|, \max_{\widehat{\Theta}: \left\| \widehat{\Theta} - \Theta_t \right\}_F^2 \le C_3} \left\| K^*(\widehat{\Theta}) \right\| \right\}.$$

Finally, define $\rho_0 = 1 - \gamma_{\min}/2$ and $\kappa = \kappa_{\max}$, where $\gamma_{\min}$ is the smaller of γ values from Assumptions 4.2 and Lemma 4.5, and similarly $\kappa_{\max}$ is the larger of the κ values.

5 ALGORITHM DYN-LQR

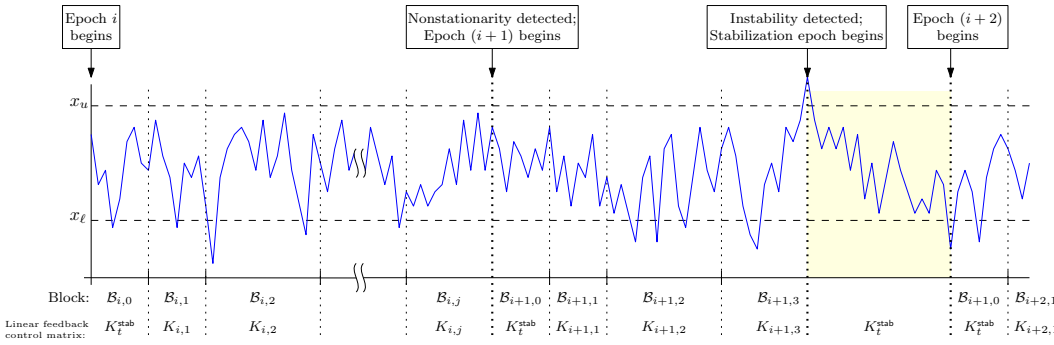


Fig. 1. Illustration of Regular epochs, blocks, and stabilization epochs for Algorithm Dyn-LQR. Epoch i ends in block $\mathcal{B}_{i,j}$ when an EndOfExplorationTest fails. Epoch $i+1$ ends because $\|x_t\|$ exceeds the threshold x_u, indicating that the current controller $K_{i+1,3}$ is potentially unstable. This triggers a stabilization epoch which ends the first time $\|x_t\|$ falls below x_ℓ, and starts epoch $i+2$.

Our algorithm Dyn-LQR is presented as Algorithm 1. At a high level, the algorithm divides the time horizon into epochs $\{\mathcal{E}_1, \mathcal{E}_2, \ldots\}$ where the squared total variation $\Delta^2_{\mathcal{E}_i}$ within epoch $\mathcal{E}_i$ is of the order $\sqrt{1/|\mathcal{E}_i|}$. This should be reminiscent of the trade-off described in the last paragraph of Section 3 where the variance of the OLS estimator for a block was proportional to the inverse square root of the length of the block. The end of an epoch signals that a sufficient change in Θ_t has accumulated and the algorithm starts a new epoch, whereby it forgets the past history and restarts the procedure to estimate the dynamics Θ_t. Since the length of an epoch is unknown to the online controller a priori, within each epoch we follow a doubling strategy (again similar to the naive algorithm in Section 3) by further splitting it into non-overlapping blocks (indexed by $j = 0, 1, \ldots$) of geometrically increasing duration. We denote the j-th block of epoch i as $\mathcal{B}_{i,j}$. During block 0, or the *warm-up block*, the algorithm plays an action $u_t = K_t^{\text{stab}} x_t + \nu_0 \eta_t$ where $\eta_t \sim \mathcal{N}(0, I_d)$ are *i.i.d.* Gaussian random vectors, and $\nu_0 = 1$ is the added exploration noise. We denote by $\{\mathcal{G}_t\}_{t \in [T]}$ the filtration generated by $\{\eta_1, \ldots, \eta_T\}$. The duration of the warm-up blocks is $L = O((n+d)\log^3 T)$. The $O(1)$ exploration noise reduces the estimation error of the OLS estimate computed at the end of the block. Observations from block j are used to create an estimate $\widehat{\Theta}_{i,j}$ of the dynamics, which in turn gives the linear feedback controller for block $j+1$ as $K_{i,j+1} := K^*(\widehat{\Theta}_{i,j})$, and action $u_t = K_{i,j+1} x_t + \nu_{j+1} \eta_t$. For a block $\mathcal{B}_{i,j}$ with $j \ge 1$, we choose $\nu_j^2 \approx \frac{1}{\sqrt{|\mathcal{B}_{i,j}|}}$ as the exploration noise similar to the stationary LQR case. If the estimate based on a block $\mathcal{B}_{i,j}$ "differs statistically" from the estimate from the previous block $\mathcal{B}_{i,j-1}$ (Algorithm 3), epoch $\mathcal{E}_i$ is ended and $\mathcal{E}_{i+1}$ started. Figure 1 gives an illustration of epochs and blocks.

Algorithm 1: DYN-LQR

Input: Horizon T, stabilizing controllers $\{K_t^{\text{stab}}\}$, input instance parameters $\rho_0, \psi, \kappa, \beta$

Definition: $v_0 = 1$; $v_j^2 = \sqrt{\frac{C_0}{2^j L}}$ for $j \geq 1$ where $C_0 = 4\log T$, $L = \frac{16(n+d)\log^3 T}{1-\rho_0}$;

  $\mathcal{B}_{i,j} = [\tau_i + 2^{j-1}L, \tau_i + 2^j L - 1]$, where τ_i is the start of exploration epoch $\mathcal{E}_i$;

  Bounds on $\|x_t\|$ for stabilization epochs: $x_u = 2\kappa e^{C_{ss}\left(\frac{\sqrt{8(n+d)}\beta}{\sqrt{1-\rho_0}}\sqrt{\log T} + \frac{(n+d)B}{1-\rho_0}\right)}$, $x_\ell = \frac{2\psi\kappa\sqrt{n}}{1-\rho_0}$;

  $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_n)$;

Initialize: $t = 1, i = 1$;

$\tau_i \leftarrow t$; /* Start of exploration epoch $\mathcal{E}_i$ */

for $t = \tau_i, \ldots, \tau_i + L - 1$ **do** /* Block 0 (warm-up) */

  Play $u_t = K_t^{\text{stab}} x_t + v_0 \eta_t$;

end

for $j = 1, 2, \ldots$ **do**

  Let $\widehat{\Theta}_{i,j-1}$ be the OLS estimator based on $\mathcal{B}_{i,j-1}$, and define $K_{i,j} = K^*(\widehat{\Theta}_{i,j-1})$;

  $\mathcal{M} \leftarrow \emptyset$; /* Initialize the set of exploration phases */

  while $t \leq \tau_i + 2^j L - 1$ **do**

    $E \sim \text{Ber}\left(\frac{1}{L} 2^{-j/2} \sum_{m=0}^{j-1} 2^{-m/2}\right)$; /* Sample exploration indicator */

    if $E = 1$ **then**

      Sample exploration scale index $m \in \{0, 1, 2, \ldots, j-1\}$ with probability $\Pr(m = b) \propto 2^{-b/2}$;

      $\mathcal{M} \leftarrow \mathcal{M} \cup \{(m, t)\}$;

    end

    Let $M_t = \{(m, s) \in \mathcal{M} \mid s \leq t \leq s + 2^m L - 1\}$; /* Active exploration phases */

    if $M_t \neq \emptyset$ **then**

      Set $m_t = \min\{m \mid \exists (m, s) \in M_t\}$;

      Play $u_t = K_{i,j} x_t + v_{m_t} \eta_t$;

    else

      Play $u_t = K_{i,j} x_t + v_j \eta_t$;

    end

    Observe x_{t+1};

    for $(m, s) \in \mathcal{M}$ *with* $t = s + 2^m L - 1$ **do**

      if *EndOfExplorationTest*$(i, j, m, s) = \textit{Fail}$ **then**

        $t \leftarrow t + 1$, $i \leftarrow i + 1$; Go to line 6 ; /* Start a new epoch */

      end

    end

    if $t = \tau_i + 2^j L - 1$ ***and*** *EndOfBlockTest*$(i, j) = \textit{Fail}$ **then**

      $t \leftarrow t + 1$, $i \leftarrow i + 1$; Go to line 6 ; /* Start a new epoch */

    end

    $t \leftarrow t + 1$;

    if $\|x_t\| \geq x_u$ **then** /* Instability detected */

      while $\|x_t\| \geq x_\ell$ **do**

        Play $u_t = K_t^{\text{stab}} x_t$, observe x_{t+1};

        $t \leftarrow t + 1$;

      end

      $i \leftarrow i + 1$; Go to line 6 ; /* Start a new epoch */

    end

  end

end

Algorithm 2: EndOfExplorationTest(i, j, m, s)

Construct OLS estimator $\widehat{\Theta}_{i,j,(m,s)}$;
$\widehat{\Theta}_{i,j,(m,s)} = \text{argmin}_{\Theta} \sum_{t=s}^{s+2^m L-1} \left\|x_{t+1} - \Theta[x_t^\top u_t^\top]^\top\right\|_F^2$;
if $\left\|\widehat{\Theta}_{i,j-1} - \widehat{\Theta}_{i,j,(m,s)}\right\|_F^2 \geq (1 + \bar{C}_{bias} + 2\bar{C}_{var})^2 (2^m L)^{-1/2}$ **then** /* See (7) */
  Return *Fail*;
end
Return *Pass*;

Algorithm 3: EndOfBlockTest(i, j)

Construct OLS estimator $\widehat{\Theta}_{i,j}$;
$\widehat{\Theta}_{i,j} = \text{argmin}_{\Theta} \sum_{t \in \mathcal{B}_{i,j}} \left\|x_{t+1} - \Theta[x_t^\top u_t^\top]^\top\right\|_F^2$;
if $\left\|\widehat{\Theta}_{i,j-1} - \widehat{\Theta}_{i,j}\right\|_F^2 \geq (1 + \bar{C}_{bias} + 2\bar{C}_{var})^2 (2^{j-1} L)^{-1/2}$ **then** /* See (7) */
  Return *Fail*;
else
  Return *Pass*;
end

The vanilla policy mentioned above suffers from the problem that we could potentially commit to a controller for a long block – and hence fail to detect a large change, which could in turn potentially lead to $O(T)$ regret. This is where the crucial novelty of the scheme of [8] (designed for contextual multi-armed bandits) comes into play: to detect non-stationarity, which may happen at different scales (few large or many small changes), at each time within the block $\mathcal{B}_{i,j}$, the authors' algorithm enters a *replay phase* where the policy from an earlier block in the same epoch (together with the larger exploration noise) is played. If at the end of some replay phase, the estimate of reward differs significantly from the history, the current epoch is ended. The algorithm could potentially be in multiple replay phases simultaneously, in which case the policy to replay is picked uniformly at random from active replays. Replay phases with different indexes are intended to detect changes of different magnitudes.

To adapt to the LQR setting, we simplify the above strategy. In particular, at any time t in a block $\mathcal{B}_{i,j}$, we enter an *exploration phase* with probability proportional to $1/\sqrt{|\mathcal{B}_{i,j}|}$ and given this event happens, the 'scale' of the exploration phase is chosen to be m with probability proportional to $1/\sqrt{2^m}$. A scale m exploration phase lasts for $2^m L$ time steps, during which we play the action $u_t = K_{i,j} x_t + \sigma_t \eta_t$. That is, we keep playing the same linear feedback controller, but with exploration noise increased to $\sigma_t^2 \approx \frac{1}{\sqrt{2^m}}$. Therefore, a scale m exploration phase allows us to detect variation in Θ_t of size $\sqrt{1/2^m}$. There can be multiple exploration phases active at any time t. We denote them by $M_t = \{(m_1, t_1), (m_2, t_2), \ldots\}$ where m_k denotes the scale and t_k denotes the starting time of the k-th active exploration phase. In this case, we play the most aggressive (i.e., the smallest m) exploration phase, with the feedback used by all active exploration phases to improve their estimates. At the end of the exploration phase (m, s), we first compute the OLS estimator $\Theta_{i,j,(m,s)}$, and declare non-stationarity and end the epoch if $\left\|\widehat{\Theta}_{i,j-1} - \widehat{\Theta}_{i,j,(m,s)}\right\|^2 \gtrapprox \frac{1}{\sqrt{2^m}}$ (Algorithm 2).

One crucial difference between LQR and the contextual bandit setting off [8] is that LQR has a quadratic cost, while contextual bandit is a special case of a linear bandit problem, which affects the choice of σ_t. Yet another crucial difference from the contextual bandit setting is that since

the LQR system has a state, the system could potentially become unstable through an inaccurate estimate before the non-stationarity is detected. We thus create a third criterion for ending an epoch: whenever $\|x_t\| \geq x_u = O\left(\frac{\sqrt{(n+d)\log T}}{1-\rho_0}\right)$, we end the current epoch and enter a *stabilization epoch*. In a stabilization epoch we keep playing the stabilizing controllers without any exploration noise until $\|x_t\|$ drops below $x_\ell = O\left(\frac{n}{1-\rho_0}\right)$. At this point, we begin a regular *exploration epoch*.

6 ESTIMATION ERROR FOR OLS WITH NON-STATIONARY Θ_t

A central ingredient of our algorithm is the ordinary least squares estimator used to learn the approximate dynamics. While the study of the variance of the OLS estimator is a well-understood topic, when the parameter sequence is non-stationary, the OLS estimator can be biased. Studying this bias is quite non-trivial, especially for the LQR problem.

We state our results on the estimation error of the OLS estimator for non-stationary LQR at the end of this section and devote Appendix C to the formal proofs of the results. However, we will highlight in brief the reason that these results are challenging and non-trivial. For intuition, the reader should keep the trade-off we pointed to at the end of Section 3 in mind: during an interval $\mathcal{I}$ of length $|\mathcal{I}|$, to balance the exploration-exploitation trade-off we would like to create an estimator that has error of order $|\mathcal{I}|^{-1/4}$. With a non-stationary parameter sequence, this error comes from both the variance of the estimator as well as the bias. Therefore, if the variation in Θ_t during this interval, $\Delta_{\mathcal{I}}$, is of smaller order than $|\mathcal{I}|^{-1/4}$, then we would like the bias of our estimator to be $O(\Delta_{\mathcal{I}})$.

Failure of a naive proof-strategy. We first show that an obvious first line of attack to bound the estimation error of OLS does not work. Define $z_t = [x_t^\top, u_t^\top]^\top$ and $\Upsilon_{\mathcal{I}} := \sum_{t\in\mathcal{I}} z_t z_t^\top$ for an interval $\mathcal{I} = [s, e]$. Then we can write the error in the OLS estimator compared to a 'representative' $\bar{\Theta}$ (e.g., $\bar{\Theta} = \Theta_e$) as:

$$\widehat{\Theta}_{\mathcal{I}} - \bar{\Theta} = \underbrace{\left(\sum_{t\in\mathcal{I}} \left(\Theta_t - \bar{\Theta}\right) z_t z_t^\top\right)\Upsilon_{\mathcal{I}}^{-1}}_{\text{"bias"}} + \underbrace{\left(\sum_{t\in\mathcal{I}} w_t z_t^\top\right)\Upsilon_{\mathcal{I}}^{-1}}_{\text{"variance"}}.$$

The above shows that if Θ_t is constant in $\mathcal{I}$, then the estimator is unbiased. Lacking that, we may try to bound the first term as follows (this proof strategy was followed in [9]). Let $\bar{\Theta} = \Theta_e$, then

$$\begin{aligned}\left\|\left(\sum_{t\in\mathcal{I}} (\Theta_t - \Theta_e) z_t z_t^\top\right)\Upsilon_{\mathcal{I}}^{-1}\right\|_F &= \left\|\left(\sum_{t\in\mathcal{I}}\sum_{p=t}^{e-1} \left(\Theta_p - \Theta_{p+1}\right) z_t z_t^\top\right)\Upsilon_{\mathcal{I}}^{-1}\right\|_F \\ &= \left\|\sum_{p=s}^{e-1} \left(\Theta_p - \Theta_{p+1}\right)\left(\sum_{t=s}^{p-1} z_s z_s^\top\right)\Upsilon_{\mathcal{I}}^{-1}\right\|_F \leq \sum_{p=s}^{e-1} \left\|\left(\Theta_p - \Theta_{p+1}\right)\right\|_F \lambda_{\max}\left(\left(\sum_{t=s}^{p-1} z_s z_s^\top\right)\Upsilon_{\mathcal{I}}^{-1}\right).\end{aligned}$$

If $\lambda_{\max}\left(\left(\sum_{t=s}^{p-1} z_s z_s^\top\right)\Upsilon_{\mathcal{I}}^{-1}\right) \leq 1$, then the analysis above would bound the bias by $\Delta_{\mathcal{I}}$. While this may seem intuitive (e.g., it is true if z_s are scalars), this was shown to be false for an arbitrary $\{z_s\}$ sequence even for the case of $z_s \in \mathbb{R}^2$ by [41].

An illustrative example. To further highlight why a technically challenging analysis is necessary for the study of OLS with a non-stationary parameter sequence, we consider a simple example of OLS estimation *without noise*. Consider a 2-dimensional example with two data points:

$$\theta_1 = [1 \;\; 1], \quad \theta_2 = [1-\epsilon \;\; 1]; \qquad z_1 = [\cos\alpha \;\; \sin\alpha], \quad z_2 = [1 \;\; 0].$$

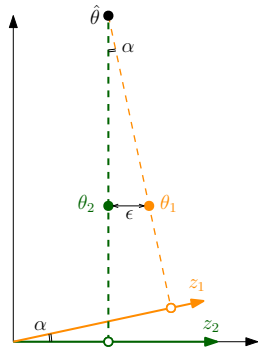


Fig. 2. Illustration for the large bias of OLS estimator with non-stationary parameters.

Figure 2 shows the geometric intuition behind the OLS estimator. In this specific example, the estimate is given by the intersection of two (for $t = 1, 2$) lines: perpendicular to z_t and passing through θ_t. The bias of the OLS estimate $\widehat{\theta}$ in this noiseless case is given as $|\widehat{\theta} - \theta_2| = \epsilon/\tan\alpha$. With $\alpha \approx \epsilon \ll 1$, the bias approaches 1 even though θ_1, θ_2 are ϵ-close to each other. The matrix Υ for this case is

$$\Upsilon = \begin{bmatrix} 1 + \cos^2\alpha & \cos\alpha \cdot \sin\alpha \\ \cos\alpha \cdot \sin\alpha & \sin^2\alpha \end{bmatrix},$$

which is ill-conditioned when $\alpha \ll 1$. In particular, $\lambda_{\max}(\Upsilon)/\lambda_{\min}(\Upsilon) \approx 1/2\alpha^2$. It might seem that such an ill-conditioned Υ is an extreme case that is unlikely to bother our study. However, with the exploration noise chosen in Algorithm 1, we give evidence in Lemma C.8 that the condition number of $\Upsilon_{\mathcal{I}}$ for intervals $\mathcal{I}$ of interest is concentrated around $O(\sqrt{|\mathcal{I}|})$, while we are trying to get unbiased estimates when the variation of Θ_t in interval $\mathcal{I}$ is $\Delta_{\mathcal{I}} = O(|\mathcal{I}|^{-1/4})$. This precisely corresponds to the problematic setting $\alpha \sim \epsilon \ll 1$ in our toy example above.

Our proof approach. We begin by decomposing the problem into bounding the estimation error for each row of the estimate $\widehat{\Theta}_{\mathcal{I}}$. For a given row, $\widehat{\theta}_{\mathcal{I}}$, the key obstacle in the analysis of the estimation error $\left\|\widehat{\theta}_{\mathcal{I}} - \bar{\theta}\right\|^2$ is that while z_t lives in $\mathbb{R}^{n+d}$, most of its variance is in the n-dimensional column space of $[I_n\ K_{\mathcal{I}}^\top]^\top$, where $K_{\mathcal{I}}$ is the fixed linear feedback controller used during interval $\mathcal{I}$. This is because the LQR dynamics naturally adds the noise w_{t-1} to arrive at the state x_t allowing efficient exploration/estimation of the component of $\bar{\theta}$ lying in the column space of $[I_n\ K_{\mathcal{I}}^\top]^\top$. In particular, the total energy in this column space is $O(|\mathcal{I}|)$ through w_t, while the energy in the orthogonal subspace through the exploration noise $\xi_t = \sigma_t \eta_t$ is $\sum_{t\in\mathcal{I}} \sigma_t^2 = O(\sqrt{|\mathcal{I}|})$. Therefore, as our toy example points out, a naive analysis based on a lower bound on the eigenvalues of the matrix $\sum_{t\in\mathcal{I}} z_t z_t^\top$ fails, because it does not exploit the statistical independence between ξ_t and x_t.

Our approach is to instead to look at one-dimensional OLS problems parameterized by directions $v \in \mathbb{S}^{n+d} := \{v \in \mathbb{R}^{n+d}, \|v\| = 1\}$:

$$\lambda_v = \operatorname*{argmin}_{\lambda} \mathcal{L}(\bar{\theta} + \lambda \cdot v),$$

where $\mathcal{L}$ is the quadratic loss function for OLS. We argue that $|\lambda_v|$ are small for 'enough' directions v. That is, in enough directions, the minimizer $\left(\bar{\theta} + \lambda_v \cdot v\right)$ of the 1-dimensional quadratic defined above is close to the candidate $\bar{\theta}$. Furthermore, since the loss function looks very different for v lying close to the column space of $[I_n\ K^\top]^\top$ versus v lying close to its orthogonal subspace, we consider two cases: v lying only in the column space or lying only in its orthogonal subspace, and

prove that the geometry of Hessian implies that it is sufficient to look at these two cases. The complete proof is presented in Appendix C.

Results. We state our lemmas for the estimation error for the OLS estimators used in Algorithm 1. Lemma 6.1 states it for intervals within exploration blocks $\mathcal{B}_{i,j}$, while Lemma 6.2 states it for warm-up blocks $\mathcal{B}_{i,0}$. The reason for the two separate results is that within a warm-up block, the controller K_t^{stab} is changing, which does not allow a subspace decomposition we mentioned earlier, but the $O(1)$ exploration noise still allows us to bound the estimation error. Within an exploration block $\mathcal{B}_{i,j}$, the exploration noise is of a much smaller magnitude (to control regret due to exploration), but the controller K_t is fixed, which allows the decomposition.

Lemma 6.1. *Consider an interval $\mathcal{I}$ in block $\mathcal{B}_{i,j}$ for some epoch $\mathcal{E}_i$ in Algorithm 1, such that $|\mathcal{I}| \geq L$ and $\max_{t\in\mathcal{I}} \|x_t\| \leq x_u$. Let $\widehat{\Theta}_{\mathcal{I}}$ be the corresponding OLS estimate from observations in $\mathcal{I}$ and $\bar{\Theta} = \Theta_t$ for some $t \in \mathcal{I}$. Then, there exists a T_0, such that for $T \geq T_0$, with probability at least $1-\varepsilon$:*

$$\left\|\widehat{\Theta}_{\mathcal{I}} - \bar{\Theta}\right\|_F \leq \check{C}_1 \Delta_{\mathcal{I}} + \check{C}_2 |\mathcal{I}|^{-\frac{1}{4}},$$

where

$$\check{C}_1 = C_{bias}\sqrt{\ln\frac{1}{\varepsilon} + \ln T} \quad \text{and} \qquad \check{C}_2 = C_{var}\left(\sqrt{\ln T} + \sqrt{\ln\frac{1}{\varepsilon}}\right),$$

for problem dependent constants C_{bias}, C_{var} (precise expressions are shown in (43)).

Lemma 6.2. *Consider a warm-up block $\mathcal{B}_{i,0}$ in Algorithm 1 and let $\bar{\Theta} = \Theta_t$ for some $t \in \mathcal{B}_{i,0}$. There exists a T_0, such that for $T \geq T_0$ and the choice of L in Algorithm 1, the OLS estimate $\widehat{\Theta}_{\mathcal{B}_{i,0}}$ of a warm-up block $\mathcal{B}_{i,0}$ satisfies*

$$\left\|\widehat{\Theta}_{\mathcal{B}_{i,0}} - \bar{\Theta}\right\|_F \leq \check{C}_{1,stab} \Delta_{\mathcal{B}_{i,0}} + \check{C}_{2,stab} |\mathcal{B}_{i,0}|^{-\frac{1}{4}},$$

with probability at least $1-\varepsilon$ where

$$\check{C}_{1,stab} = C_{bias,stab}\sqrt{\ln T} \quad \text{and} \qquad \check{C}_{2,stab} = C_{var,stab}\sqrt{\ln\frac{1}{\varepsilon} + \ln\ln T},$$

for problem dependent constants $C_{bias,stab}, C_{var,stab}$ (precise expressions are shown in (49)).

Applying Lemma 6.1 and Lemma 6.2 with $\varepsilon = 1/T^3$ to all the intervals (at most T^2) that may be considered during the execution of Algorithm 1 and a union bound immediately gives the following result.

Lemma 6.3. *Define Event 1 as the event that for each warm-up block $\mathcal{B}_{i,0}$ in Algorithm 1 it holds that*

$$\left\|\widehat{\Theta}_{\mathcal{B}_{i,0}} - \bar{\Theta}\right\|_F \leq C_{bias,stab}\sqrt{\ln T}\,\Delta_{\mathcal{B}_{i,0}} + 3C_{var,stab}\sqrt{\ln T}\,|\mathcal{B}_{i,0}|^{-\frac{1}{4}},$$

and for each phase and non-warmup block, denoted by $\mathcal{I} = [s, e]$, it holds that

$$\left\|\widehat{\Theta}_{\mathcal{I}} - \bar{\Theta}\right\|_F \leq 3C_{bias}\sqrt{\ln T}\,\Delta_{\mathcal{I}} + 3C_{var}\sqrt{\ln T}\,|\mathcal{I}|^{-\frac{1}{4}}.$$

Then we have that $\mathbf{Pr}[\text{Event 1}] \geq 1 - 1/T$.

For succinctness, define

$$\bar{C}_{\text{bias}} = \sqrt{\ln T}\max\{3C_{\text{bias}}, C_{\text{bias,stab}}\} \quad \text{and} \quad \bar{C}_{\text{var}} = \sqrt{\ln T}\max\{3C_{\text{var}}, 3C_{\text{var,stab}}\}. \tag{7}$$

7 REGRET UPPER BOUND FOR DYN-LQR

Our main regret upper bound for Dyn-LQR is shown below.

Theorem 7.1. *Under Assumption 4.2, the expected regret of Dyn-LQR is upper bounded as:*

$$\mathbf{E}\left[\mathcal{R}^{\text{Dyn-LQR}}(T)\right] = \widetilde{O}\left(V_T^{2/5}T^{3/5}\right).$$

If the dynamics $\{\Theta_t\}$ are piecewise constant with at most S switches, then the regret of Dyn-LQR is upper bounded as:

$$\mathbf{E}\left[\mathcal{R}^{\text{Dyn-LQR}}(T)\right] = \widetilde{O}\left(\sqrt{ST}\right).$$

Our definition of $\mathcal{R}^{\text{Dyn-LQR}}(T)$ in (1) measures the regret relative to the benchmark $\sum_{t=1}^{T} J_t^*$. In the next proposition, we prove that this benchmark is at most $\widetilde{O}(V_T)$ larger than the expected cost of the dynamic optimal policy. This additive error is dominated by the regret $\widetilde{O}(V_T^{2/5}T^{3/5})$ proved in Theorem 7.1. Proposition 7.2 is proved in Appendix D.

Proposition 7.2. *Let $\{u_t\}_{t=1}^T$ be an arbitrary non-anticipative policy for the non-stationary LQR control problem. Then,*

$$\mathbf{E}\left[\sum_{t=1}^{T} x_t^\top Q x_t + u_t^\top R u_t\right] \geq \sum_{t=1}^{T} J_t^* - O(V_T + \log T).$$

We will conduct our analysis under the assumption that Event 1 specified in Lemma 6.3 occurs. Since Dyn-LQR uses K_t^{stab} whenever $\|x_t\| \geq x_u$, outside this event, the total cost is bounded by $\widetilde{O}(T)$. Note that this happens with probability at most $1/T$.

7.1 Regret Decomposition

We begin with an informal regret decomposition lemma which highlights the key exploration-exploitation trade-off for non-stationary LQR.

Informal Lemma. The expected regret for a policy π with $u_t = K_t x_t + \sigma_t \eta_t$ where K_t, σ_t are adapted to the filtration $(\mathcal{F}, \mathcal{G})$ is given by:

$$\begin{aligned}
\mathbf{E}[\mathcal{R}^{\pi}(T)] &= \mathbf{E}\left[\sum_{t=1}^{T} x_t^\top Q x_t + u_t^\top R u_t - J_t^*\right] \\
&= \underbrace{\sum_{t=1}^{T} \mathbf{E}\left[J_t(K_t) - J_t^*\right]}_{\text{exploitation regret}} + \underbrace{\sum_{t=1}^{T} \mathbf{E}\left[\sigma_t^2 \operatorname{Tr}\left(R + B_t^\top P_t(K_t) B_t\right)\right]}_{\text{exploration regret}} \\
&\quad + \underbrace{\sum_{t=1}^{T-1} \mathbf{E}\left[x_{t+1}^\top \left(P_{t+1}(K_{t+1}) - P_t(K_t)\right) x_{t+1}\right]}_{\text{policy/parameter variation}} + \mathbf{E}\left[x_1^\top P_1(K_1) x_1 - x_{T+1}^\top P_T(K_T) x_{T+1}\right]. \qquad (8)
\end{aligned}$$

We term the lemma informal because it relies on $J_t(K_t)$ and $P_t(K_t)$ being defined for all t. This need not always be true for Dyn-LQR since K_t is the certainty equivalent controller based on an estimate of Θ_t, and therefore the *stationary* system corresponding to Θ_t and K_t need not even be stable, and $J_t(K_t)$ could be unbounded. We shortly address how we handle such time periods, but their contribution to regret will be asymptotically of a smaller order. The decomposition points out that the dominant terms in the analysis will be the exploitation regret and the exploration

regret. The policy/parameter variation depends on how much the pair (Θ_t, K_t) changes during non-warmup blocks of an exploration epoch. By design, the policies $\{K_t\}$ are piece-wise constant with at most $\log T$ changes per epoch, and we will prove that the number of epochs is $O(V_T^{4/5}T^{1/5})$. Finally, for a fixed K, $\|P(\Theta_t, K) - P(\Theta_{t+1}, K)\| = O(\|\Theta_t - \Theta_{t+1}\|)$, and hence this contributes at most $O(V_T)$ to the regret across the entire horizon.

To refine the regret decomposition, we recapitulate Algorithm 1, and in particular the classification of *exploration epochs*, *stabilization epochs*, *blocks* within exploration epochs, and another concept we define for the purpose of analysis alone – *bad intervals*.

(i) Stabilization epochs – such epochs begin whenever $\|x_t\|$ exceeds the upper bound x_u, indicating the potential instability of the current controller. We use τ_i^{stab} to denote the start of the i-th stabilization epoch. During a stabilization epoch, we use the controller $K_t = K_t^{\text{stab}}$. The i-th stabilization phase ends at θ_i^{stab} (inclusive) where

$$\theta_i^{\text{stab}} = \min\{t \geq \tau_i^{\text{stab}} + 1 \ : \ \|x_{t+1}\| \leq x_\ell\}.$$

We use $\mathcal{S}_i$ to denote the interval $[\tau_i^{\text{stab}}, \theta_i^{\text{stab}}]$ as well as the i-th stabilization epoch symbolically.

(ii) Exploration epochs – such epochs begin either at the end of a stabilization epoch, or at the end of another exploration epoch if sufficient non-stationarity is detected through failure of EndOfExplorationTest or EndOfBlockTest. We will denote the start and end of the i-th exploration epoch by τ_i and θ_i respectively, and use $\mathcal{E}_i$ to denote the interval $[\tau_i, \theta_i]$ as well as the epoch symbolically.

(iii) Blocks – The i-th exploration epoch $\mathcal{E}_i$ is partitioned into non-overlapping blocks of geometrically increasing duration. Block 0 (also called the *warm-up block*) is the interval $[\tau_i, \tau_i + L - 1]$, and the j-th block ($j = 1, \ldots$) is the interval $[\tau_i + L \cdot 2^{j-1}, \tau_i + L \cdot 2^j - 1] \cap \mathcal{E}_i$ of maximum length $L \cdot 2^{j-1}$. We denote by $\mathcal{B}_{i,j}$ both the interval as well as the block symbolically. The controller used at time $t \in \mathcal{B}_{i,j}$ ($j \geq 1$) is given by $K^*(\widehat{\Theta}_{i,j-1})$, where $\widehat{\Theta}_{i,j-1}$ is the OLS estimator based on the block $j-1$ of epoch $\mathcal{E}_i$. For succinctness, we use the notation

$$\widehat{\Theta}_t := \widehat{\Theta}_{i,j-1}, \quad \text{for } t \in \mathcal{B}_{i,j}.$$

We will use B_i to denote the number of blocks in epoch $\mathcal{E}_i$.

(iv) Bad/Good intervals – It can happen that for some time steps during an exploration epoch, the controller is unstable and therefore $J_t(K_t)$ is undefined, but the $\|x_t\|$ has not exceeded x_u. To study the regret due to such t, we define the notion of bad intervals within epochs. The k-th bad interval of an epoch i begins at $\tau_{i,k}^{\text{bad}}$ and ends at $\theta_{i,k}^{\text{bad}}$ where these are defined recursively as:

$$\tau_{i,1}^{\text{bad}} := \min\left\{t \in [\tau_i + L, \theta_i] : \left\|\widehat{\Theta}_t - \Theta_t\right\|_F^2 \geq C_3\right\},$$

$$\tau_{i,k}^{\text{bad}} := \min\left\{t \in [\theta_{i,k-1}^{\text{bad}}, \theta_i] : \left\|\widehat{\Theta}_t - \Theta_t\right\|_F^2 \geq C_3\right\},$$

$$\theta_{i,k}^{\text{bad}} := \min\left\{t \in [\tau_{i,k}^{\text{bad}} + 1, \theta_i] \ : \ \left\|\widehat{\Theta}_{t+1} - \Theta_{t+1}\right\|_F^2 \leq \frac{C_3}{2}\right\},$$

with the constant C_3 defined in Assumption 4.3. Note that we do not create bad intervals during the block $\mathcal{B}_{i,0}$, which is analyzed separately. We denote the k-th bad interval of an epoch i as $\mathcal{I}_{i,k}^{\text{bad}}$. By $\mathcal{I}_i^{\text{bad}}$, we denote the union of all bad intervals in $\mathcal{E}_i$, and by $\mathcal{I}^{\text{bad}}$, the union of all bad intervals. All time periods that not in bad intervals, i.e., they are in $\mathcal{E}_i \setminus \{\mathcal{B}_{i,0} \cup \mathcal{I}_i^{\text{bad}}\}$, will be called *good* and split into *good intervals*. For analysis purposes, we further split the good time periods based on the blocks. That is, a good interval can end at time t if *(i)* either a bad interval begins at time $t+1$, or *(ii)* a block ends at time t in which case another good interval can begin at time $t+1$. Using a

similar notation $\mathcal{I}^{\text{good}}_{i,j,k}$ denotes the k-th good interval of a block $\mathcal{B}_{i,j}$ (which must lie entirely inside $[\tau_i + L \cdot 2^{j-1}, \tau_i + L \cdot 2^j - 1]$. We will use N^{bad}_i to denote the total number of bad intervals in an epoch i and $N^{\text{good}}_{i,j}$ to denote the number of good intervals in a block $\mathcal{B}_{i,j}$. The advantage of defining the good intervals to lie within a block is that for the purposes of analysis, the good intervals within a block $\mathcal{B}_{i,j}$ are completely defined based on history before the start of the block $\mathcal{B}_{i,j}$.

We will use E to denote the total number of exploration epochs and E_S to denote the total number of stabilization epochs. Finally, we come to the regret decomposition that we use in the subsequent section:

$$\begin{aligned}\mathcal{R}^{\text{Dyn-LQR}}(T) &= \sum_{t=1}^{T} c_t - J^*_t \\ &\leq \underbrace{\sum_{i=1}^{E_S} \sum_{t \in \mathcal{S}_i} c_t}_{T_1\text{: Stabilization epochs}} + \underbrace{\sum_{i=1}^{E} \sum_{t \in \mathcal{B}_{i,0}} c_t}_{T_2\text{: Warm-up blocks}} + \underbrace{\sum_{i=1}^{E} \sum_{k=1}^{N^{\text{bad}}_i} \sum_{t \in \mathcal{I}^{\text{bad}}_{i,k}} c_t}_{T_3\text{: Bad intervals}} + \underbrace{\sum_{i=1}^{E} \sum_{j=1}^{B_i} \sum_{k=1}^{N^{\text{good}}_{i,j}} \sum_{t \in \mathcal{I}^{\text{good}}_{i,j,k}} \left(c_t - J^*_t\right)}_{T_4\text{: Good intervals}}. \end{aligned} \tag{9}$$

7.2 Regret analysis for Dyn-LQR

The main result of this section is the following lemma, which provides an intermediate characterization of $\mathbf{E}\left[\mathcal{R}^{\text{Dyn-LQR}}(T)\right]$ based on (9). In particular, the characterization highlights that to bound the regret, it is sufficient to bound *(i)* the number E of exploration epochs (Section 7.3) and *(ii)* the total squared norm of the estimation error of dynamics Θ_t for the good periods (Section 7.4).

Lemma 7.3. *The expected regret for Dyn-LQR is bounded as follows:*

$$\begin{aligned}\mathbf{E}\left[\mathcal{R}^{\text{Dyn-LQR}}(T)\right] &\leq \widetilde{O}\left(\mathbf{E}\left[\sum_{i=1}^{E}\sum_{j=1}^{B_i}\sum_{t\in\mathcal{B}_{i,j}} \min\left\{\left\|\widehat{\Theta}_{i,j-1} - \Theta_t\right\|_F^2, C_3\right\} + \sqrt{|\mathcal{B}_{i,j}|}\right]\right) + \widetilde{O}(E + V_T), \\ &\leq \widetilde{O}\left(\mathbf{E}\left[\sum_{i=1}^{E}\sum_{j=1}^{B_i}\sum_{t\in\mathcal{B}_{i,j}} \min\left\{\left\|\widehat{\Theta}_{i,j-1} - \Theta_t\right\|_F^2, C_3\right\}\right]\right) + \widetilde{O}(\sqrt{E \cdot T} + V_T). \end{aligned} \tag{10}$$

Proof. We proceed by bounding the terms in (9).

Upper bound for Term 1. Since the controllers $\{K^{\text{stab}}_t\}$ used in a stabilization epoch satisfy sequentially strong stability (Assumption 4.2), in Lemma 7.4 we prove that the expected total cost per stabilization epoch is $\widetilde{O}(1)$. Since the number of stabilization epochs is bounded by the number of exploration epochs E, the total contribution of Term 1 in (9) is $\widetilde{O}(E)$.

Lemma 7.4. *Let $[\tau^{stab}, \theta^{stab}]$ be a stabilization epoch. The expected total cost during the stabilization epoch is bounded by*

$$\mathbf{E}\left[\sum_{t=\tau^{stab}}^{\theta^{stab}} c_t \,\middle|\, \mathcal{F}_{\tau^{stab}-1}, \mathcal{G}_{\tau^{stab}-1}\right] = O\left(\frac{\kappa^2 x_u^2}{1-\rho_0}\right) = O\left(\frac{\kappa^2\beta^2(n+d+\log T)}{(1-\rho_0)^2}\right).$$

Upper bound for Term 2. Similar to Lemma 7.4, the use of K^{stab}_t during warm-up blocks gives a bound of $\widetilde{O}(1)$ per epoch in Lemma 7.5, which gives a $\widetilde{O}(E)$ contribution due to Term 2.

Lemma 7.5. *Let $[\tau_i, \tau_i + L - 1]$ denote the warm-up block $\mathcal{B}_{i,0}$ of an exploration epoch $\mathcal{E}_i$. The expected total cost during $\mathcal{B}_{i,0}$, for any i, is bounded by*

$$\mathbf{E}\left[\sum_{t=\tau_i}^{\tau_i+L-1} c_t \,\middle|\, \mathcal{F}_{\tau_i-1}, \mathcal{G}_{\tau_i-1}\right] = O\left(\frac{\kappa^2\beta^2(n+d)\log^3 T}{(1-\rho_0)^2}\right).$$

Upper bound for Term 3. Since $\|x_t\|$ is bounded by $x_u = \widetilde{O}(1)$ for any time period in a bad interval by definition, the cost is bounded by $\widetilde{O}(1)$ per time step. We can bound the number of bad time periods within an arbitrary interval $\mathcal{I}$ noting that for $t \in \mathcal{I}^{\text{bad}}$, $\left\|\widehat{\Theta}_t - \Theta_t\right\|_F^2 \geq C_3/2$ and thus:

$$\left|\mathcal{I}^{\text{bad}} \cap \mathcal{I}\right| = \sum_{t \in \mathcal{I}^{\text{bad}} \cap \mathcal{I}} 1 \leq \frac{2}{C_3} \sum_{t \in \mathcal{I}} \min\left\{\left\|\widehat{\Theta}_t - \Theta_t\right\|_F^2, C_3\right\}. \tag{11}$$

Then the total contribution of Term 3 is $O\left(\mathbf{E}\left[\sum_{i=1}^{E}\sum_{j=1}^{B_i}\sum_{t\in\mathcal{B}_{i,j}} \min\left\{\left\|\widehat{\Theta}_{i,j-1} - \Theta_t\right\|_F^2, C_3\right\}\right]\right)$.

Upper bound for Term 4.

Lemma 7.6. *For some epoch $\mathcal{E}_i$, a block $\mathcal{B}_{i,j}$ in epoch $\mathcal{E}_i$, and a good interval $\mathcal{I}_{i,j,k}^{good} = [\tau, \theta]$ in block $\mathcal{B}_{i,j}$, the expected regret is bounded as follows:*

$$\begin{aligned}
&\mathbf{E}\left[\mathcal{R}^{\pi}(\mathcal{I}_{i,j,k}^{good}) \mid \mathcal{F}_{\tau-1}, \mathcal{G}_{\tau-1}\right] \\
&\quad\leq \sum_{t=\tau}^{\theta}\left(J_t(K_t) - J_t^*\right) + \left|\mathcal{I}_{i,j,k}^{good}\right| \frac{C_0^{1/2}}{L^{3/2}} \cdot \frac{j}{\sqrt{2^j}} + O\left(\frac{n+d+\log T}{1-\rho_0}\left(1 + \Delta_{\mathcal{I}_{i,j,k}}\right)\right) \\
&\quad\leq \sum_{t=\tau}^{\theta} C_4 \left\|\widehat{\Theta}_{i,j-1} - \Theta_t\right\|_F^2 + \left|\mathcal{I}_{i,j,k}^{good}\right| \frac{C_0^{1/2} C_7}{L^{3/2}} \cdot \frac{j}{\sqrt{2^j}} + O\left(\frac{n+d+\log T}{1-\rho_0}\left(1 + \Delta_{\mathcal{I}_{i,j,k}}\right)\right),
\end{aligned}$$

where the constant $C_7 := \max_t \sup\left\{\mathrm{Tr}\left(R + B_t^\top P_t(K_t) B_t\right) \mid K_t = K^*(\widehat{\Theta}), \left\|\widehat{\Theta} - \Theta_t\right\|_F^2 \leq C_3\right\}$.

Combining the results above, we can bound the first term in (10) immediately from Term 3 and the first summand in Lemma 7.6 for Term 4. Summing the second term in Lemma 7.6 over all the good intervals within a block $\mathcal{B}_{i,j}$ (which is of length at most 2^j) contributes $\widetilde{O}(\sqrt{|\mathcal{B}_{i,j}|})$. Since the blocks within an epoch are doubling in length, $\sum_j \sqrt{|\mathcal{B}_{i,j}|} \leq 8\sqrt{|\mathcal{E}_i|}$ and $\sum_i \sqrt{|\mathcal{E}_i|} \leq E\sqrt{T/E} = \sqrt{E \cdot T}$. The contribution of the third term in Lemma 7.6 is proportional to the number of good intervals, which is bounded by $V_t/\sqrt{C_3/2} + E\log T$. To see this, note that without any bad intervals, there would be one good interval per block and there are at most $\log T$ blocks per epoch. For a good interval to begin due to a bad interval ending, the bad interval must 'eat up' $\sqrt{C_3/2}$ of the variation due to the criterion chosen for the end of a bad interval. Hence, there can be at most $V_T/\sqrt{C_3/2}$ good intervals created because of the bad intervals. The last term in Lemma 7.6 contributes $\widetilde{O}(V_T)$ to (10). □

7.3 Bounding the Number of Epochs

There are two ways of generating epochs in Algorithm 1: (1) epochs end due to the detection of non-stationarity (lines 29 and 33), and (2) epochs end due to the detection of instability (line 41). This section is devoted to bounding the number of epochs from these two sources separately.

Bounding the number of epochs generated by non-stationarity tests. In the subsequent analysis, we will bound the number of epochs terminated due to the detection of non-stationarity in Θ_t by $O(T^{1/5}V_T^{4/5})$, which dominates $O(V_T)$. Recall that an epoch ends if the non-stationarity tests in Algorithms 2 or 3 fail, which happens if the distance between the new OLS estimate and the estimate based on the previous block exceeds some threshold. The thresholds there are carefully designed according to the concentration results proved in Section 6, which allow us to prove the following lemma characterizing the variation budget needed for an epoch to fail the tests in Algorithms 2 and 3.

LEMMA 7.7. *Assume EVENT 1 holds. Let $\mathcal{E}_i$ be an epoch with total variation $\Delta_{[\tau_i,t]} \le (t-\tau_i+1)^{-1/4}$, then the epoch does not end because of nonstationarity detection.*

The following corollary bounds the number of restarts due to detection of non-stationarity.

COROLLARY 7.8. *Assume EVENT 1 holds. The number of epochs that end due to detection of non-stationarity is bounded by $O(C_0^{-2/5}T^{1/5}V_T^{4/5})$.*

Bounding the number of epochs generated by instability tests. Lemma 7.9 characterizes the variation budget needed to trigger the end of an epoch due to instability detection, which leads to Corollary 7.10 bounding the number of epochs ended due to instability.

LEMMA 7.9. *Let $\mathcal{E}_i$ be an epoch with total variation $\Delta_{\mathcal{E}_i} \le \left(\sqrt{\frac{C_3}{4}} - \bar{C}_{var}L^{-1/4}\right)/\bar{C}_{bias}$. Then under EVENT 1, with probability at least $1-O(1/T^3)$, the epoch does not end because of instability detection.*

COROLLARY 7.10. *The expected number of epochs that end due to the instability test is bounded by $O(V_T\sqrt{\ln T})$.*

Combining the two bounds, we get $E = O(T^{1/5}V_T^{4/5})$. Therefore, we can bound the $\widetilde{O}(\sqrt{E\cdot T})$ term in (10) by $\widetilde{O}(T^{3/5}V_T^{2/5})$.

7.4 Bounding the Total Square Norm of the Estimation Error

In this section, we analyse the regret due to the estimation error, i.e., the first term in (10). For succinctness, define the following *loss* function for an arbitrary interval $\mathcal{I}$:

$$\mathcal{L}(I) := \sum_{t\in\mathcal{I}} \min\left\{C_4\left\|\widehat{\Theta}_{i,j-1}-\Theta_t\right\|_F^2, C_3\right\}. \tag{12}$$

In the sequel, we first focus on an exploration epoch $\mathcal{E}_i$ and bound $\mathcal{L}(\mathcal{E}_i)$. We then combine the regret of epochs to get the requisite regret bound of Theorem 7.1.

Our proof decomposes into three parts. First, we focus on one block, say block j, of epoch i, and prove a lemma that provides an upper bound for $\mathcal{L}(\mathcal{I})$ for any interval $\mathcal{I}\subseteq\mathcal{B}_{i,j}$. Second, we partition a block into intervals with small total variation within each interval. We use the just mentioned bound to bound $\mathcal{L}(\mathcal{B}_{i,j})$ of each block j in an exploration epoch i in terms of the length of the block and the total variation within the block. Finally, we upper bound the total number of blocks within an epoch i and sum up the bound on $\mathcal{L}(\mathcal{B}_{i,j})$ for all the blocks in an epoch $\mathcal{E}_i$ to obtain a bound on $\mathcal{L}(\mathcal{E}_i)$.

LEMMA 7.11. *For an arbitrary interval $\mathcal{I}=[s,e]$ that lies in block $\mathcal{B}_{i,j}$, define $\varepsilon_{\mathcal{I}} := \left\|\widehat{\Theta}_{i,j-1}-\Theta_s\right\|_F^2$ and $\alpha_{\mathcal{I}} := \frac{\log|\mathcal{I}|}{\sqrt{\mathcal{I}}}$. Then, $\mathcal{L}(\mathcal{I})$ can be bounded as*

$$\mathcal{L}(\mathcal{I}) = O\left(|\mathcal{I}|\alpha_{\mathcal{I}} + |\mathcal{I}|\Delta_{\mathcal{I}}^2 + |\mathcal{I}|\varepsilon_{\mathcal{I}}\mathbb{1}\{\varepsilon_{\mathcal{I}}\ge\alpha_{\mathcal{I}}\}\right).$$

To get a bound for the regret for a block, we need to partition $\mathcal{B}_{ij}$ into intervals with small variation. Specifically, we have the following lemma adapted from [8].

Lemma 7.12. *There is a way to partition any block $\mathcal{B}$ into $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_\Gamma$ such that*

$$\Delta^2_{\mathcal{I}_k} \le \frac{\log_2 T}{\sqrt{|I_k|}} = \alpha_{I_k}, \qquad k \in [\Gamma],$$

and the number of blocks Γ satisfies $\Gamma = O\left(\min\left\{S_{\mathcal{B}}, (\log|\mathcal{B}|)^{-\frac{2}{5}} \Delta_{\mathcal{B}}^{\frac{4}{5}} |\mathcal{B}|^{\frac{1}{5}} + 1\right\}\right)$.

The partition in Lemma 7.12 is for the analysis only. The intuition for this partition is to create small enough intervals so that their regret can be shown to be small, while at the same time not creating too many intervals. Applying Lemma 7.11 to each interval of the partition of block $\mathcal{J}$:

$$\mathcal{L}(\mathcal{B}) \le \widetilde{O}\Big(\sum_{k=1}^{\Gamma-1} |\mathcal{I}_k| \alpha_{\mathcal{I}_k} + \sum_{k=1}^{\Gamma-1} |\mathcal{I}_k| \varepsilon_{\mathcal{I}_k} \mathbb{1}\{\varepsilon_{\mathcal{I}_k} \ge \alpha_{\mathcal{I}_k}\}\Big) + \mathcal{L}(\mathcal{I}_\Gamma). \tag{13}$$

Plugging in the definition of $\alpha_{\mathcal{I}_k}$, we get $|\mathcal{I}_k| \alpha_{\mathcal{I}_k} = \sqrt{\mathcal{I}_k} \log |\mathcal{I}_k|$. Then by the Cauchy-Schwartz inequality and the upper bound for Γ from Lemma 7.12, we have

$$\sum_{k=1}^{\Gamma-1} \sqrt{|\mathcal{I}_k|} \log |\mathcal{I}_k| \le \sqrt{(\Gamma-1) \sum_{k=1}^{\Gamma-1} |\mathcal{I}_k| \log^2 |\mathcal{I}_k|} = \widetilde{O}\left(\sqrt{(\Gamma-1) \sum_{k=1}^{\Gamma-1} |\mathcal{I}_k|}\right) = \widetilde{O}\left(|\mathcal{B}|^{\frac{3}{5}} \Delta_{\mathcal{B}}^{\frac{2}{5}}\right).$$

We defer the bound for the remaining terms of (13) to Appendix E.3. The following lemma presents the resulting upper bound for the loss function of a block $\mathcal{B}$.

Lemma 7.13. *Let $\mathcal{B} = \mathcal{B}_{i,j}$ be a block of some epoch i with $j > 0$. It holds with high probability that Dyn-LQR guarantees*

$$\mathcal{L}(\mathcal{B}) \le \widetilde{O}\left(|\mathcal{B}|^{\frac{3}{5}} \Delta_B^{\frac{2}{5}} + \sqrt{|\mathcal{B}|}\right).$$

From the geometrically increasing size of $\mathcal{B}_{ij}$, we get $\sum_j \sqrt{|\mathcal{B}_{ij}|} = O(|\mathcal{E}_i|)$. From the Hölder's inequality, we get

$$\sum_j |\mathcal{B}_{ij}|^{\frac{3}{5}} \Delta_{B_{i,j}}^{\frac{2}{5}} \le \left(\sum_j |\mathcal{B}_{ij}|\right)^{\frac{3}{5}} \left(\sum_j \Delta_{B_{i,j}}\right)^{\frac{2}{5}} = |\mathcal{E}_i|^{\frac{3}{5}} \Delta_{\mathcal{E}_i}^{\frac{2}{5}};$$

so that $\mathcal{L}(\mathcal{E}_i) = \widetilde{O}(|\mathcal{E}_i|^{3/5} \Delta_{\mathcal{E}_i}^{2/5} + \sqrt{\mathcal{E}_i})$. One more application of the Hölder's inequality gives the bound of $\widetilde{O}(T^{\frac{3}{5}} V_T^{\frac{2}{5}})$, proving Theorem 7.1.

8 REGRET LOWER BOUNDS

In this section, we prove two lower bounds for the regret of the non-stationary LQR problem. First, in Theorem 8.1 we prove that for any given $V_T = o(T)$, no learning algorithm can guarantee a regret $o(V_T^{3/5} T^{2/5})$, showing that the regret of Dyn-LQR is minimax optimal as a function of V_T. Next, in Theorem 8.3 we prove that a broad class of static-window based online learning algorithms are regret suboptimal for non-stationary LQR – even if the algorithm has the knowledge of the variation V_T. This rules out several popular approaches that have been used in the literature for learning under non-stationary such as UCB with static restart schedule or bandit-on-bandit approaches to optimize the window size.

Theorem 8.1. *There exists a T_0 such that for any $T \geq T_0$, and a total variation V_T of dynamics, for any randomized online algorithm Alg (which knows T, V_T), there exists a non-stationary LQR instance with regret lower bounded as*

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

Under switching dynamics with S switches, for any randomized algorithm Alg (which knows T, S), there exists an instance with regret lower bounded as

$$\mathbf{E}\left[\mathcal{R}^{Alg}(T)\right] = \Omega\left(\sqrt{ST}\right).$$

Proof. We build on the lower bound instance from [7]. Consider a randomly generated one dimensional LQR problem instance with dynamics and cost:

$$\begin{aligned} x_{t+1} &= ax_t + bu_t + w_t\,, \\ c_t &= x_t^2 + u_t^2\,, \end{aligned} \tag{14}$$

where $w_t \sim \mathcal{N}(0,1)$. The dynamics are given by $a = 1/\sqrt{5}$ and $b = \chi\sqrt{\epsilon}$, with χ being a Rademacher random variable that takes values ± 1 with equal probability. Standard results show that the optimal linear feedback controller for the above LQR system is:

$$k^* = -\frac{abp^*}{1 + b^2 p^*} \tag{15}$$

where p^* solves

$$p^* = 1 + \frac{a^2 p^*}{1 + b^2 p^*}. \tag{16}$$

In Cassel et al. [7], the authors prove the following lower bound on the regret of any algorithm.

Theorem 8.2 (Cassel et al. [7, Theorem 13]). *For $T \geq 12000$ and $\epsilon = \sqrt{T}/4$, the expected regret of any deterministic learning algorithm for system (14) satisfies*

$$\mathbf{E}[\mathcal{R}(T)] \geq \frac{\sqrt{T}}{3100} - 4.$$

By Yao's theorem, the above implies that for any randomized learning algorithm, there is an LQR instance with expected regret $\Omega(\sqrt{T})$.

We create a lower bound instance for a non-stationary LQR problem with the total variation V_T by pasting a sequence of these one-dimensional instances. In particular, we concatenate $\lfloor \frac{V_T}{2\sqrt{\epsilon}} \rfloor$ instances of (14) with horizon $\lfloor \frac{1}{4\epsilon^2} \rfloor$ each, where ϵ satisfies $\frac{V_T}{2\sqrt{\epsilon}} = T \cdot 4\epsilon^2$, or equivalently $\epsilon = \left(\frac{V_T}{8T}\right)^{2/5}$. That is, we re-randomize χ for every sub-instance. To demonstrate a lower bound, we further allow the learner the knowledge of the time instants at which a new sub-instance begins, and the duration of the sub-instance. Theorem 8.2 implies that the regret of the learner for each sub-instance is $\Omega\left(\frac{1}{2\epsilon}\right)$, for a total regret over the entire time horizon of $\Omega\left(\frac{V_T}{\epsilon^{3/2}}\right) = \Omega\left(V_T^{2/5} T^{3/5}\right)$.

If, instead of bounded total variation, the non-stationary LQR instance has a piecewise constant dynamics with S switches, we create a lower bound instance similarly with S sub-instances of horizon $\lfloor T/S \rfloor$ each, and $\epsilon = \frac{\sqrt{T/S}}{4}$. The regret per sub-instance for any learner is $\Omega(\sqrt{T/S})$ for a total regret lower bound of $\Omega(\sqrt{ST})$. □

Necessity of Adaptive Restarts. A common technique to handle non-stationary learning environments is to use random restarts or sliding window algorithms to forget the distant history. In learning problems where the rewards are linear in the unknown parameters (e.g., in multi-armed bandit problems), this gives the optimal regret rate in terms of the total variation of the instance *if* the window size is chosen optimally – in the lower bound instance, the adversary changes the instance by $O\left(V_T^{1/3}T^{-1/3}\right)$ at regularly spaced times. In the LQR problem, we instead have that the per-step regret $J^*(\Theta) - J(\Theta, K^*(\widehat{\Theta}))$ is *quadratic* in $\left\|\Theta - \widehat{\Theta}\right\|_F$. Intuitively, the adversary can maximally penalize a non-adaptive restart based algorithm by changing the instance by as much as $\Theta(1)$ at regularly spaced, but randomly chosen times. This strategy fails against an adaptive restart algorithm such as Dyn-LQR because big changes are easy to detect with less exploration effort. To give a little more formal intuition, we consider the one-dimensional LQR problem (14) from [7], but with non-stationary b_t, and a fairly general static window based algorithm for this non-stationary LQR instance. We prove that even with optimal tuning of the window size and an arbitrary exploration strategy, it can incur a regret as large as $\Omega(V_T^{1/3}T^{2/3})$.

We first describe the one-dimensional instance and the family of sliding window algorithms we consider. **Instance:** The cost function is $x_t^2 + u_t^2$ and the dynamics are given by:

$$x_{t+1} = ax_t + b_t u_t + w_t,$$

with $x_1 = 0$ and $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. The dynamics parameter is time-invariant $a = 1/\sqrt{5}$ and known to the algorithm (therefore, there is no learning needed for a). The sequence $\{b_t\}$ is random and generated as follows. Let $\epsilon = 0.05 \cdot (V_T/T)^{1/6}$. We choose $b_1 = \epsilon$. For each subsequent t, with probability $\frac{V_T}{2T}$, b_t is chosen to be ± 0.05 with equal probability, or, with probability $(V_T/4T)^{5/6}$, b_t is chosen to be $\pm\epsilon$ with equal probability, otherwise $b_t = b_{t-1}$. The key feature of the instance is that while most of the time b_t is small of size ϵ and most of the changes in b_t are of order ϵ as well, there are much rarer changes in b_t of $O(1)$ size. These two scales of changes make any fixed window size suboptimal for the regret.

Non-adaptive Restart with Exploration (RestartLQR(W)) Algorithm. We consider a family of algorithms parametrized by a window size W. Let $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. The algorithm splits the horizon T into non-overlapping phases of duration W each, and for time t in phase i, the algorithm plays $u_t = \widehat{k}_{(i)} x_t + \sigma_t \eta_t$, where $\widehat{k}_{(i)}$ is a linear feedback controller estimated by the algorithm based *only* on the trajectory observed in phase $(i-1)$, and σ_t is an arbitrary adapted sequence of exploration noise (energy) injected by the algorithm. To emphasize, the algorithm is restricted in two senses. First, it is restricted to playing a fixed linear feedback controller within each phase with Gaussian exploration noise. Second, at the beginning of each phase, the algorithm forgets the entire history and restarts the estimation of the dynamics.

Theorem 8.3. *The expected regret of RestartLQR under optimally tuned window size W and exploration strategy is at least $\Omega\left(V_T^{1/3}T^{2/3}\right)$.*

9 CONCLUDING THOUGHTS

In this paper, we have tried to fill an obvious gap in the literature – the absence of any low dynamic regret algorithm for the control of a non-stationary LQR system under stochastic noise. We discuss the possibility of wider applicability of our results and some open questions.

A Queueing Application. While in the paper we focused on the LQR problem, the key motif of the LQR problem that drove our results was that *(i)* given the state and action, the feedback we receive was a linear function (i.e., linear feedback); and *(ii)* given an ϵ error in the parameter estimates, the

optimal controller for the estimated parameters has an $O(\epsilon^2)$ additive suboptimality (i.e., quadratic cost). Similar motif shows up in numerous other applications where we believe a similar regret trade-off would show up. Here we mention a queueing example. Consider the following discrete time queueing system with a configurable server: the arrivals per period are i.i.d. Bernoulli with a known mean $\lambda < 1$. The server has two resources (say CPU and memory) and the operator can choose a configuration $(x, y) \in \{x^2+y^2 \leq 1; x, y \geq 0\}$ of the two resources. Given the configuration, the number of departures per period is also a Bernoulli random variable with mean,

$$\mu_t = \alpha_t x + \beta_t y,$$

where $\alpha_t, \beta_t \geq 0, \lambda < \alpha_t^2+\beta_t^2 \leq 1$ represent the resource requirements of the jobs, are non-stationary, and unknown to the operator. Assume a job that arrives in time step t can not be served before time step $t+1$. The cost at time step t is N_t, the number of jobs in the system. This system fits the motif of linear feedback and quadratic cost. The linear feedback can be seen by noting that the feedback at time step t is the Bernoulli random variable for the number of departures, which can be written as $\alpha_t x_t + \beta_t y_t + \eta_t$, where η_t is a mean 0 bounded random variable (independent across time periods). To see the quadratic cost part, consider the steady-state problem with stationary (α, β), and a stationary action (x, y) giving $\mu_t = \mu = \alpha x + \beta y$. The steady-state average cost would be $N(\mu, \lambda) = \frac{\lambda(1-\lambda)}{\mu-\lambda}$. In this case, the optimal action is to choose (x_t, y_t) in the direction (α, β) under which $\mu^* = \alpha^2 + \beta^2$ with optimal cost $N^* = N_{\mu^*,\lambda}$. Consider an estimate $(\widehat{\alpha}, \widehat{\beta})$ such that $|\alpha - \widehat{\alpha}| + |\beta - \widehat{\beta}| = \epsilon$. If $\lambda \leq \alpha^2 + \beta^2 - \frac{1}{100}$, then the controller based on the estimated $\widehat{\alpha}, \widehat{\beta}$ gives cost $N^* + \Theta(\epsilon^2)$, which is what we mean by a quadratic cost. We therefore expect that our results for the LQR problem would extend to the control of such queueing systems.

Open Questions. We believe both our algorithm and the regret analysis can be tightened, e.g., using sequential hypothesis testing to detect instability instead of our current threshold based approach, and made parameter free. An algorithm with a bound on regret of the following flavor would be desirable: There exist constant ϵ_0, T_0 such that for a non-stationary LQR problem with variation $V_T = \epsilon T$, where $\epsilon \leq \epsilon_0$ and $T \geq T_0$, the regret attained is at most $\epsilon^{2/5}T + o(T)$. It is also desirable to develop a notion of instance-optimal regret – instead of using the summary V_T and presenting minimax optimal guarantees.

Yet another challenging direction is that there seem to be two prevalent approaches to studying robustness for online control of LQR systems – one with non-stochastic/adversarial noise and another with unknown non-stationary dynamics. This leaves an open problem of finding a controller which achieves both types of robustness simultaneously or proving the impossibility of doing so. A second open problem is to consider more general convex cost functions. Many of the elegant results in LQR theory, and indeed the regret bounds in our paper, depend on the quadratic objective function. A starting point would be to study a bandit problem with linear feedback, but a general convex cost function. Finally, a notoriously hard problem is to study the robust control where the action set may depend on the state, which touches upon the theme of *safe exploration*. Doing so in the context of LQR could be fruitful.

REFERENCES

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*. 2312–2320.

[2] Yasin Abbasi-Yadkori and Csaba Szepesvári. 2011. Regret bounds for the adaptive control of linear quadratic systems. In *Conference on Learning Theory*. 1–26.

[3] Michael Athans. 1971. The role and use of the stochastic linear-quadratic-Gaussian problem in control system design. *IEEE transactions on automatic control* 16, 6 (1971), 529–552.

[4] Dimitri Bertsekas. 2012. *Dynamic programming and optimal control: Volume I*. Vol. 1. Athena scientific.

[5] Omar Besbes, Yonatan Gur, and Assaf Zeevi. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems* 27 (2014), 199–207.

[6] Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. 2021. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*. PMLR, 471–483.

[7] Asaf Cassel, Alon Cohen, and Tomer Koren. 2020. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning*. PMLR, 1328–1337.

[8] Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. 2019. A New Algorithm for Non-stationary Contextual Bandits: Efficient, Optimal and Parameter-free. In *COLT*. 696–726. http://proceedings.mlr.press/v99/chen19b.html

[9] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. 2019. Learning to Optimize under Non-Stationarity. In *Proceedings of Machine Learning Research (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 1079–1087. http://proceedings.mlr.press/v89/cheung19b.html

[10] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. 2019. Non-stationary reinforcement learning: The blessing of (more) optimism. *Available at SSRN 3397818* (2019).

[11] Gregory C Chow. 1976. Control methods for macroeconomic policy analysis. *The American Economic Review* 66, 2 (1976), 340–345.

[12] Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. 2018. Online linear quadratic control. *arXiv preprint arXiv:1806.07104* (2018).

[13] Alon Cohen, Tomer Koren, and Yishay Mansour. 2019. Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret. arXiv:1902.06223 [cs.LG]

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

[15] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. 2018. Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 4192–4201.

[16] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. 2018. Finite-time adaptive stabilization of linear systems. *IEEE Trans. Automat. Control* 64, 8 (2018), 3498–3505.

[17] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. 2020. Input perturbations for adaptive control and learning. *Automatica* 117 (2020), 108950.

[18] P.M. Gahinet, A.J. Laub, C.S. Kenney, and G.A. Hewer. 1990. Sensitivity of the stable discrete-time Lyapunov equation. *IEEE Trans. Automat. Control* 35, 11 (1990), 1209–1217. https://doi.org/10.1109/9.59806

[19] Pratik Gajane, Ronald Ortner, and Peter Auer. 2018. A sliding-window algorithm for Markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066* (2018).

[20] Aurélien Garivier and Eric Moulines. 2011. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*. Springer, 174–188.

[21] Gautam Goel and Babak Hassibi. 2021. Regret-optimal Estimation and Contro. *arXiv preprint arXiv:2106.12097* (2021).

[22] Paula Gradu, Elad Hazan, and Edgar Minasyan. 2020. Adaptive regret for control of time-varying dynamics. *arXiv preprint arXiv:2007.04393* (2020).

[23] Bruce Hajek. 1982. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability* (1982), 502–525.

[24] Elad Hazan, Sham Kakade, and Karan Singh. 2020. The nonstochastic control problem. In *Algorithmic Learning Theory*. PMLR, 408–421.

[25] Elad Hazan and Comandur Seshadhri. 2009. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*. 393–400.

[26] Mark Herbster and Manfred K Warmuth. 1998. Tracking the best expert. *Machine learning* 32, 2 (1998), 151–178.

[27] Morteza Ibrahimi, Adel Javanmard, and Benjamin V Roy. 2012. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*. 2636–2644.

[28] Yassir Jedra and Alexandre Proutiere. 2021. Minimal Expected Regret in Linear Quadratic Control. arXiv:2109.14429 [cs.LG]

[29] Beatrice Laurent and Pascal Massart. 2000. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* (2000), 1302–1338.

[30] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. 2016. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research* 17, 1 (2016), 1334–1373.

[31] Horia Mania, Stephen Tu, and Benjamin Recht. 2019. Certainty equivalence is efficient for linear quadratic control. *arXiv preprint arXiv:1902.07826* (2019).

[32] Prasad A Naik. 2014. Marketing dynamics: A primer on estimation and control. *Foundations and Trends in Marketing* 9, 3 (2014), 175–266.

[33] Ronald Ortner, Pratik Gajane, and Peter Auer. 2020. Variational regret bounds for reinforcement learning. In *Uncertainty in Artificial Intelligence*. PMLR, 81–90.

[34] Yoan Russac, Claire Vernade, and Olivier Cappé. 2019. Weighted Linear Bandits for Non-Stationary Environments. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/263fc48aae39f219b4c71d9d4bb4aed2-Paper.pdf

[35] Max Simchowitz and Dylan Foster. 2020. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning*. PMLR, 8937–8948.

[36] Max Simchowitz, Karan Singh, and Elad Hazan. 2020. Improper learning for non-stochastic control. In *Conference on Learning Theory*. PMLR, 3320–3436.

[37] Russ Tedrake. 2009. Underactuated robotics: Learning, planning, and control for efficient and agile machines course notes for MIT 6.832. *Working draft edition* 3 (2009).

[38] Roman Vershynin. 2010. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* (2010).

[39] Chen-Yu Wei and Haipeng Luo. 2021. Non-stationary Reinforcement Learning without Prior Knowledge: An Optimal Black-box Approach. *arXiv preprint arXiv:2102.05406* (2021).

[40] Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. 2009. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research* 34, 3 (2009), 737–757.

[41] Peng Zhao and Lijun Zhang. 2021. Non-stationary linear bandits revisited. *arXiv preprint arXiv:2103.05324* (2021).

[42] Martin Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*. 928–936.

A BASIC LEMMAS

Lemma A.1 (Laurent-Massart Bound [29]). *Let $a_1, \ldots, a_n$ be non-negative, and $X_1, \ldots, X_n$ be i.i.d. χ^2 random variables. Let*

$$|a|_\infty := \max_{i\in[n]} a_i \quad \text{and} \quad |a|_2^2 := \sum_{i\in[n]} a_i^2.$$

Then,

$$\mathbf{Pr}\left[\sum_{i\in[n]} a_i(X_i - 1) \geq 2|a|_2\sqrt{x} + 2|a|_\infty x\right] \leq e^{-x},$$

$$\mathbf{Pr}\left[\sum_{i\in[n]} a_i(X_i - 1) \leq -2|a|_2\sqrt{x}\right] \leq e^{-x}.$$

Lemma A.2. *Let $Y_1, \ldots, Y_T$ be i.i.d. χ_k^2 random variables. Then,*

$$\mathbf{E}\left[\max_{t\in T} Y_t\right] \leq k + \max\{12k, 3\ln T\} + 3,$$

$$\mathbf{E}\left[\max_{t\in T} \sqrt{Y_t}\right] \leq \sqrt{k} + \sqrt{8\ln T} + \sqrt{\frac{\pi}{2}}.$$

Proof. By Laurent-Massart bound (Lemma A.1),

$$\mathbf{Pr}\left[Y_t \geq k + 2\sqrt{kx} + 2x\right] \leq e^{-x}.$$

For $y \geq 12k$, we have the following sequence of implications

$$\begin{aligned}
&\mathbf{Pr}[Y_t \geq k + y] \leq e^{-\frac{y}{3}},\\
\implies &\mathbf{Pr}[Y_t \leq k + y] \geq 1 - e^{-\frac{y}{3}}\\
\implies &\mathbf{Pr}\left[\max_{t\in[T]} Y_t \leq k + y\right] \geq \left(1 - e^{-\frac{y}{3}}\right)^T\\
\implies &\mathbf{Pr}\left[\max_{t\in[T]} Y_t \geq k + y\right] \leq 1 - \left(1 - e^{-\frac{y}{3}}\right)^T.
\end{aligned}$$

Let $y = \max\{12k, 3\ln T\} + z$ for $z \geq 0$. Then,

$$\begin{aligned}
\mathbf{Pr}\left[\max_{t\in[T]} Y_t \geq k + y\right] &\leq 1 - \left(1 - e^{-\frac{\max\{12k, 3\ln T\} + z}{3}}\right)^T\\
&\leq 1 - \left(1 - \frac{1}{T}e^{-\frac{z}{3}}\right)^T\\
&\leq e^{-\frac{z}{3}}.
\end{aligned}$$

From the above,

$$\mathbf{E}\left[\max_{t\in T} Y_t\right] \leq k + \max\{12k, 3\ln T\} + \int_{z=0}^{\infty} e^{\frac{z}{3}} dz \leq k + \max\{12k, 3\ln T\} + 3.$$

For the second part, we again begin from the Laurent-Massart bound. For any $x \geq 0$,

$$\mathbf{Pr}\left[\sqrt{Y_t} \geq \sqrt{k} + \sqrt{2x}\right] \leq e^{-x},$$

which in turn implies for $y \geq 0$,

$$\mathbf{Pr}\left[\sqrt{Y_t} \geq \sqrt{k} + \sqrt{2\ln T} + y\right] \leq e^{-\frac{y^2 + 2y\sqrt{2\ln T}}{2}} \leq e^{-\frac{y^2}{2}}.$$

Further substituting $y = \sqrt{2\ln T} + z$ for $z \geq 0$,

$$\mathbf{Pr}\left[\max_t \sqrt{Y_t} \geq \sqrt{k} + 2\sqrt{2\ln T} + z\right] \leq 1 - \left(1 - \frac{1}{T}e^{-\frac{z^2 + 2z\sqrt{2\ln T}}{2}}\right)^T \leq e^{-\frac{z^2}{2}}.$$

Finally,

$$\begin{aligned}\mathbf{E}\left[\max_t \sqrt{Y_t}\right] &\leq \sqrt{k} + \sqrt{8\ln T} + \int_0^\infty e^{-\frac{z^2}{2}} dz \\ &= \sqrt{k} + \sqrt{8\ln T} + \sqrt{\frac{\pi}{2}}.\end{aligned}$$

□

The following lemma is adapted from [23], but we prove it here for completeness.

Lemma A.3. *Let $W_1, W_2, \ldots$ be a non-negative stochastic process, and $(\mathcal{W}_t)_{t\in\mathbb{N}}$ be the induced filtration. Let $Y_0, Y_1, \ldots$ be a non-negative stochastic process adapted to $\mathcal{W}_t$ such that for some $0 < \rho < 1$, for all $t \geq 0$,*

$$Y_{t+1} \leq \rho Y_t + W_{t+1}, \qquad \text{almost surely.}$$

Let $a \geq 0$ and $\rho \leq \widehat{\rho} < 1$ be such that for all $t \geq 1$,

$$\mathbf{E}\left[\rho + \frac{W_{t+1}}{a} \;\middle|\; \mathcal{W}_t\right] \leq \widehat{\rho}.$$

Define the a-hitting time of process $\{Y_t\}$ as:

$$\tau_a = \min_{k\geq 1}\{Y_k \leq a\}.$$

Then,

(1) $\mathbf{Pr}[\tau_a \geq k \mid \mathcal{W}_0] \leq \frac{Y_0}{a}\widehat{\rho}^t$,

(2) $\mathbf{E}\left[\sum_{k=0}^{\tau_a} Y_k^2 \mid \mathcal{W}_0\right] \leq \frac{Y_0^2}{1-\widehat{\rho}^2} \leq \frac{Y_0^2}{1-\widehat{\rho}}$.

Proof. Conditioning on the event $\{Y_t \geq a\}$ and using the definition of $\widehat{\rho}$ above,

$$\begin{aligned}\mathbf{E}[Y_{t+1} \mid \mathcal{W}_t, Y_t \geq a] &\leq \mathbf{E}[\rho Y_t + W_{t+1} \mid \mathcal{W}_t, Y_t \geq a] \\ &= \mathbf{E}\left[Y_t\left(\rho + \frac{W_{t+1}}{Y_t}\right) \;\middle|\; \mathcal{W}_t, Y_t \geq a\right] \\ &\leq \widehat{\rho}\cdot Y_t.\end{aligned}$$

Therefore, the stopped process $Y_{t\wedge\tau_a}/\widehat{\rho}^{t\wedge\tau_a}$ is a non-negative supermartingale, and hence

$$Y_0 \geq \mathbf{E}\left[\frac{Y_{k\wedge\tau_a}}{\widehat{\rho}^{k\wedge\tau_a}} \;\middle|\; \mathcal{W}_0\right] \geq \mathbf{E}\left[\frac{Y_{k\wedge\tau_a}}{\widehat{\rho}^{k\wedge\tau_a}}\mathbb{1}\{\tau_a \geq k\} \;\middle|\; \mathcal{W}_0\right] \geq \frac{a}{\widehat{\rho}^k}\mathbf{Pr}[\tau_a \geq k \mid \mathcal{W}_0].$$

That is, $\mathbf{Pr}[\tau_a \geq k \mid \mathcal{W}_0] \leq \frac{Y_0}{a}\widehat{\rho}^k$, proving the first part of the lemma. For the second part,

$$\sum_{k=0}^{\tau_a} Y_k^2 = \sum_{k=0}^{\infty} Y_k^2 \cdot \mathbb{1}\{\tau_a \geq k\}.$$

Taking the expectation and using the supermartingale result from above,

$$\mathbf{E}\left[\sum_{k=0}^{\tau_a} Y_k^2 \mid \mathcal{W}_0\right] = \sum_{k=0}^{\infty} \mathbf{E}\left[Y_k^2 \cdot \mathbb{1}\{\tau_a \geq k\} \mid \mathcal{W}_0\right] \;\leq\; \sum_{k=0}^{\infty} Y_0^2 \widehat{\rho}^{2k} \;=\; \frac{Y_0^2}{1-\widehat{\rho}^2}.$$

□

The following lemma on hitting times of exponentially ergodic random walks will be helpful for bounding the number of epochs that end because of instability detection through $\|x_t\|$ becoming large.

Lemma A.4. *Let $Y_0, Y_1, \ldots$ be a non-negative stochastic process satisfying*

$$Y_{t+1} \leq \rho Y_t + \sum_{i=1}^{m} \beta_{i,t+1} \mid W_{i,t+1}|$$

where $\rho < 1$, and $W_{i,t}$ are i.i.d $\mathcal{N}(0,1)$ random variables. Furthermore, let $\max_{i,t} \beta_{i,t} \leq B$, and $\bar{a} = \left(\frac{\sqrt{8m}B}{\sqrt{1-\rho}}\sqrt{\log T} + \frac{mB}{1-\rho}\right)$. Then,

$$\mathbf{Pr}\left[\max_{t\in[T]} Y_t \geq Y_0 + \bar{a}\right] \leq \frac{1}{T^3}.$$

Proof. Extending the sequence of random variables for $t \leq 0$, we get the following upper bound on Y_t:

$$Y_t \leq \rho^t Y_0 + \sum_{k=0}^{\infty} \rho^k B \sum_{i=1}^{m} |W_{i,t-k}|.$$

Let

$$S_t := \sum_{k=0}^{\infty} \rho^k B \sum_{i=1}^{m} |W_{i,t-k}|.$$

Therefore, for $a \geq 0$,

$$\mathbf{Pr}[Y_t \geq Y_0 + a] \leq \mathbf{Pr}\left[\rho^t Y_0 + S_t \leq Y_0 + a\right] \leq \mathbf{Pr}[S_t \leq a].$$

Furthermore,

$$\begin{aligned} S_t^2 &= B^2\left(\sum_{k=0}^{\infty}\sum_{i=1}^{m} \rho^{2k} W_{i,t-k}^2 + \sum_{(k_1,i_1)\neq(k_2,i_2)} \rho^{k_1+k_2} |W_{i_1,t-k_1}| \cdot |W_{i_2,t-k_2}|\right) \\ &\leq B^2\left(\sum_{k=0}^{\infty}\sum_{i=1}^{m} \rho^{2k} W_{i,t-k}^2 + \frac{1}{2}\sum_{(k_1,i_1)\neq(k_2,i_2)} \rho^{k_1+k_2}\left(W_{i_1,t-k_1}^2 + W_{i_2,t-k_2}^2\right)\right) \\ &= mB^2 \sum_{k=0}^{\infty}\sum_{i=1}^{m} \frac{\rho^k}{1-\rho} W_{i,k}^2. \end{aligned}$$

Applying Laurent-Massart bound from Lemma A.1,

$$\mathbf{Pr}\left[S_t^2 \geq \frac{m^2B^2}{(1-\rho)^2} + \frac{2m^{3/2}B^2}{(1-\rho)\sqrt{1-\rho^2}}\sqrt{x} + \frac{2mB^2}{1-\rho}x\right] \leq e^{-x}.$$

A simple upper bound on the right hand side within $\mathbf{Pr}[\cdot]$ gives,

$$\mathbf{Pr}\left[S_t \geq \frac{mB}{1-\rho} + \frac{\sqrt{2m}B}{\sqrt{1-\rho}}\sqrt{x}\right] \leq e^{-x}. \tag{17}$$

Substituting $x = 4 \ln T$,

$$\mathbf{Pr}\left[Y_t \geq Y_0 + \frac{mB}{1-\rho} + \frac{\sqrt{8m}B}{\sqrt{1-\rho}} \sqrt{\ln T} \right] \leq \frac{1}{T^4}.$$

A union bound completes the final argument. □

For reference, we note some basic matrix norm inequalities:

(1) $\frac{1}{2} \left\| \Theta - \widehat{\Theta} \right\| \leq \max \left\{ \left\| A - \widehat{A} \right\|, \left\| B - \widehat{B} \right\| \right\} \leq \left\| \Theta - \widehat{\Theta} \right\|$;

(2) $\frac{1}{\sqrt{n}} \left\| \Theta - \widehat{\Theta} \right\|_F \leq \left\| \Theta - \widehat{\Theta} \right\| \leq \left\| \Theta - \widehat{\Theta} \right\|_F$.

Lemma 3.1 states that $\left\| \Theta - \widehat{\Theta} \right\|_F \leq C_3$ implies $J^*(\Theta) - J(\Theta, K^*(\widehat{\Theta})) \leq C_4 \left\| \Theta - \widehat{\Theta} \right\|_F^2$.

B PROOF OF SEQUENTIAL STRONG STABILITY

Lemma B.1 ([18]). *Let X be the solution to the Lyapunov equation*

$$X - F^\top X F = M.$$

Let $X + \Delta X$ be the solution to the perturbed problem

$$Z - (F + \Delta F)^\top Z (F + \Delta F) = M.$$

Then the following inequality holds for the spectral norm:

$$\frac{\|\Delta X\|}{\|X + \Delta X\|} \leq 2 \left\| \sum_{k=0}^{+\infty} \left(F^\top\right)^k F^k \right\| \cdot (2\|F\| + \|\Delta F\|) \cdot \|\Delta F\|.$$

Proof of Lemma 4.5. Let $P_t := P(\Theta_t, \widehat{K})$ and $P_{t+1} := P(\Theta_{t+1}, \widehat{K})$ be the solutions to the following Lyapunov equations, respectively:

$$P_t = Q + \widehat{K}^\top R \widehat{K} + (A_t + B_t \widehat{K})^\top P_t (A_t + B_t \widehat{K}),$$
$$P_{t+1} = Q + \widehat{K}^\top R \widehat{K} + (A_{t+1} + B_{t+1} \widehat{K})^\top P_{t+1} (A_{t+1} + B_{t+1} \widehat{K}).$$

Taking $X = P_t$, $X + \Delta X = P_{t+1}$, $F = A_t + B_t \widehat{K}$, $F + \Delta F = A_{t+1} + B_{t+1} \widehat{K}$, and applying Lemma B.1, we get the following Lemma as a corollary.

Lemma B.2. *It holds that*

$$P_t \leq P_{t+1} \cdot \left(1 + \frac{2(1-\gamma)^2}{1 - (1-\gamma)^2} (2(1-\gamma) + (\kappa + 1)\|\Theta_{t+1} - \Theta_t\|) \right) \|\Theta_{t+1} - \Theta_t\|.$$

Proof. Applying Lemma B.1 with $X = P_t$, $X + \Delta X = P_{t+1}$, $F = A_t + B_t \widehat{K}$, and $\Delta F = A_{t+1} + B_{t+1} \widehat{K} - (A_t + B_t \widehat{K}) = (A_{t+1} - A_t) + (B_{t+1} - B_t) \widehat{K}$, we have

$$\begin{aligned} \frac{\|P_{t+1} - P_t\|}{\|P_{t+1}\|} &\leq 2 \left\| \sum_{k=0}^{+\infty} \left((A_t + B_t \widehat{K})^\top \right)^k (A_t + B_t \widehat{K})^k \right\| \\ &\quad \cdot (2\|(A_t + B_t \widehat{K})\| + \|(A_{t+1} - A_t) + (B_{t+1} - B_t)\widehat{K}\|) \cdot \|(A_{t+1} - A_t) + (B_{t+1} - B_t)\widehat{K}\| \\ &\leq \frac{2(1-\gamma)^2}{1 - (1-\gamma)^2} \cdot (2(1-\gamma) + (1+\kappa)\|\Theta_{t+1} - \Theta_t\|) \|\Theta_{t+1} - \Theta_t\|, \end{aligned}$$

where in the last inequality we use $\|A_t + B_t \widehat{K}\| \leq 1 - \gamma$ and $\|\widehat{K}\| \leq \kappa$. Then by direct computation, we have

$$\|P_t\| \leq \|P_{t+1}\| + \|P_{t+1} - P_t\|$$

$$\leq \|P_{t+1}\| \cdot \left(1 + \frac{2(1-\gamma)^2}{1-(1-\gamma)^2}(2(1-\gamma) + (\kappa+1)\|\Theta_{t+1} - \Theta_t\|)\|\Theta_{t+1} - \Theta_t\|\right).$$

□

In the sequel, we first prove that K is (κ, γ)-strongly stable for $A_t + B_t K = H_t L_t H_t^{-1}$. Note that by our assumption,

$$J^*(\Theta_t, \widehat{K}) \leq J^*(\Theta_t) + C_4 \left\|\Theta_t - \widehat{\Theta}\right\|_F^2 \leq J^*(\Theta_t) + C_4 C_3 \leq J_{\mathcal{I}}^* + C_4 C_3 \coloneqq \widetilde{J}_{\mathcal{I}}^*.$$

We have $\lambda_{\max}(P_t) \leq \widetilde{J}_{\mathcal{I}}^*/\psi^2$ and $\|H_t\| \leq \sqrt{\widetilde{J}_{\mathcal{I}}^*}/\psi =: B_0$. By definition, we have

$$\begin{aligned} P_t &= Q + K^\top R K + (A_t + B_t K)^\top P_t (A_t + B_t K) \\ &\geq q_{\min} I + r_{\min} K^\top K + (A_t + B_t K)^\top P_t (A_t + B_t K) \\ &\geq q_{\min} I + (A_t + B_t K)^\top P_t (A_t + B_t K). \end{aligned}$$

Specifically, we have $P_t \geq q_{\min} I$. Hence $\|H_t^{-1}\| \leq q_{\min}^{-1/2} =: 1/b_0$. Then setting $\kappa = B_0/b_0 = \sqrt{\frac{\widetilde{J}_{\mathcal{I}}^*}{\psi^2 q_{\min}}}$ will suffice. By $P_t \geq r_{\min} K^\top K$, we have

$$\|K\| \leq \sqrt{\frac{\|P_t\|}{r_{\min}}} \leq \sqrt{\frac{\widetilde{J}_{\mathcal{I}}^*}{\psi^2 r_{\min}}} =: \kappa.$$

Moreover,

$$\begin{aligned} L_t^\top L_t &= P_t^{-1/2} (A_t + B_t K)^\top P_t (A_t + B_t K) P_t^{-1/2} \\ &\leq P_t^{-1/2} (P_t - q_{\min} I) P_t^{-1/2} \\ &\leq I - q_{\min} P_t^{-1}. \end{aligned}$$

Then

$$\|L_t\|^2 \leq 1 - \frac{q_{\min}\psi^2}{\widetilde{J}_{\mathcal{I}}^*}$$

and

$$\|L_t\| \leq \sqrt{1 - \frac{q_{\min}\psi^2}{J^*}} \leq 1 - \frac{q_{\min}\psi^2}{2\widetilde{J}_{\mathcal{I}}^*}.$$

In the sequel, we prove the (κ, γ)-sequentially strongly stability. By direct computation, we have

$$\begin{aligned} \|H_{t+1}^{-1} H_t\|^2 &= \|P_{t+1}^{-1/2} P_t^{1/2}\|^2 \\ &= \|P_{t+1}^{-1/2} P_t P_{t+1}^{-1/2}\| \\ &\leq \left(1 + \frac{2(1-\gamma)^2}{1-(1-\gamma)^2}(2(1-\gamma) + (\kappa+1)\|\Theta_{t+1} - \Theta_t\|)\|\Theta_{t+1} - \Theta_t\|\right), \end{aligned}$$

where in the inequality we apply Lemma B.2. By the fact that $\sqrt{1+x} \leq 1 + \frac{1}{2}x$ for $x \geq 0$, we have

$$\begin{aligned} \|H_{t+1}^{-1} H_t\| &\leq 1 + \frac{(1-\gamma)^2}{1-(1-\gamma)^2} (2(1-\gamma) + (\kappa+1)\|\Theta_{t+1} - \Theta_t\|) \|\Theta_{t+1} - \Theta_t\| \\ &\leq 1 + \frac{2(1-\gamma)^2}{1-(1-\gamma)^2} ((1-\gamma) + (\kappa+1)) \|\Theta_{t+1} - \Theta_t\|, \end{aligned}$$

where in the last step we use that $\|\Theta_{t+1} - \Theta_t\| \le 2\sqrt{C_3} \le 2$ by our assumption that $C_3 \le 1$.

Proof of Lemma 4.6. Without loss of generality, we let $s_{\mathcal{I}} = 1$. Since $x_{t+1} = (A_t + B_t K)\, x_t + \sigma_t B_t \eta_t + w_t$, we have

$$x_t = M_1 x_1 + \sum_{s=1}^{t-1} M_{s+1}(\sigma_s B_s \eta_s + w_s),$$

where we define $M_t = I$ and $M_s = \prod_{j=s}^{t-1} \left(A_j + B_j K_j\right)$. Moreover,

$$\begin{aligned}
\|M_s\| &= \left\| \prod_{j=s}^{t-1} H_j L_j^\top H_j^{-1} \right\| \\
&\le \|H_{t-1}\| \left(\prod_{j=s}^{t-1} \left\|L_j\right\| \right) \left(\prod_{j=s}^{t-2} \left\|H_{j+1}^{-1} H_j\right\| \right) \left\|H_s^{-1}\right\| \\
&\le B_0 (1-\gamma)^{t-s} \left(\prod_{j=s}^{t-2} \left\|H_{j+1}^{-1} H_j\right\| \right) (1/b_0) \\
&\le \kappa (1-\gamma)^{t-s} \left(\prod_{j=s}^{t-2} \left\|H_{j+1}^{-1} H_j\right\| \right).
\end{aligned}$$

Using the fact that $1 + x \le e^x$, we have

$$\begin{aligned}
\prod_{j=s}^{t-2} \left\|H_{j+1}^{-1} H_j\right\| &\le e^{\sum_{j=s}^{t-2} \frac{2(1-\gamma)^2}{1-(1-\gamma)^2} ((1-\gamma)+(\kappa+1)) \|\Theta_{t+1} - \Theta_t\|} \\
&\le e^{C_{ss} V_{[s,t-1]}}
\end{aligned}$$

for some constant C_{ss}. Then it holds that

$$\begin{aligned}
\|M_s\| &\le \kappa (1-\gamma)^{t-s} e^{C_{ss} V_{[s,t-1]}} \\
&\le \kappa e^{-\gamma(t-s)} e^{C_{ss} V_{[s,t-1]}}.
\end{aligned}$$

Then we can bound the norm of x_t as

$$\begin{aligned}
\|x_t\| &\le \|M_1\|\, \|x_1\| + \sum_{s=1}^{t-1} \|M_{s+1}\|\, \|\sigma_s B_s \eta_s + w_s\| \\
&\le \kappa e^{-\gamma(t-1)} e^{C_{ss} V_{[1,t-1]}} \|x_1\| + \kappa e^{-\gamma(t-s)} e^{C_{ss} V_{[s,t-1]}} \sum_{s=1}^{t-1} (1-\gamma)^{t-s-1} \|\sigma_s B_s \eta_s + w_s\| \\
&\le \kappa e^{-\gamma(t-1)} e^{C_{ss} V_{[1,t-1]}} \|x_1\| + \kappa e^{-\gamma(t-s)} e^{C_{ss} V_{[s,t-1]}} \max_{1<s<t} \|\sigma_s B_s \eta_s + w_s\| \sum_{t=1}^{\infty} (1-\gamma)^t \\
&= \kappa e^{-\gamma(t-1) + C_{ss} V_{[1,t-1]}} \|x_1\| + \frac{\kappa e^{-\gamma(t-s) + C_{ss} V_{[s,t-1]}}}{\gamma} \max_{1<s<t} \|\sigma_s B_s \eta_s + w_s\|.
\end{aligned}$$

C ESTIMATION ERROR BOUNDS FOR OLS WITH NON-STATIONARY DYNAMICS

C.1 Proof of error bound for $\mathcal{I} \subseteq \mathcal{B}_{ij}$ (Lemma 6.1)

Given the OLS estimator for interval $\mathcal{I} = [s, e]$,

$$\widehat{\Theta}_{\mathcal{I}} = \operatorname*{argmin}_{\Theta} \sum_{t \in \mathcal{I}} \left\|x_{t+1} - \Theta [x_t^\top \; u_t^\top]^\top\right\|_F^2,$$

our goal is to bound the estimation error $\left\|\widehat{\Theta}_{\mathcal{I}} - \bar{\Theta}\right\|_F$ where $\bar{\Theta}$ is a 'representative' Θ for $\{\Theta_t\}_{t\in\mathcal{I}}$, for example Θ_e. We will assume that during the entire interval $\mathcal{I}$, the controller $K_t = K$ is stationary. Let $M := \begin{bmatrix} I_n \\ K \end{bmatrix}$. We will use the notation

$$y_t = Mx_t \quad \text{and} \quad \xi_t = \sigma_t \widetilde{\eta}_t = \sigma_t \begin{bmatrix} 0 \\ I_d \end{bmatrix} \eta_t, \quad \text{so that:} \quad z_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix} = y_t + \xi_t \quad \text{and} \quad x_{t+1} = \Theta_t z_t + w_t.$$

By our choice of σ_t, we have $\sigma_L^2 := \nu_1 = \sqrt{\frac{C_0}{L}} \geq \sigma_t^2 \geq \sigma_{\mathcal{I}}^2 := \sqrt{\frac{C_0}{|\mathcal{I}|}}$ for all $t \in \mathcal{I}$. With these notations, we can write the OLS loss function and estimator as:

$$\widehat{\Theta}_{\mathcal{I}} = \operatorname*{argmin}_{\Theta} \mathcal{L}(\Theta), \qquad \text{where} \quad \mathcal{L}(\Theta) = \sum_{t\in\mathcal{I}} \|x_{t+1} - \Theta z_t\|_F^2 = \sum_{t\in\mathcal{I}} \|\Theta_t z_t + w_t - \Theta z_t\|_F^2 .$$

Due to the OLS objective function, we can decompose this problem and estimate each row of $\widehat{\Theta}_{\mathcal{I}}$ separately. Towards that end, let us fix a row i. With abuse of notation, denote the ith rows of $\Theta_t, \Theta, \widehat{\Theta}_I, \bar{\Theta}$ by $\theta_t, \theta, \widehat{\theta}, \bar{\theta}$, respectively. Let us also use ω_t to denote the ith entry of w_t. The OLS estimation problem for the row i becomes:

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \mathcal{L}(\theta), \qquad \text{where,} \quad \mathcal{L}(\theta) = \sum_{t\in\mathcal{I}} \left(\langle \theta_t, z_t\rangle + \omega_t - \langle \theta, z_t\rangle\right)^2 .$$

The solution for this OLS estimation problem is given by the solution of the following linear system:

$$\widehat{\theta}\left(\sum_{t\in\mathcal{I}} z_t z_t^\top\right) = \left(\sum_{t\in\mathcal{I}} \theta_t z_t z_t^\top\right) + \sum_{t\in\mathcal{I}} \omega_t z_t^\top,$$

or

$$\widehat{\theta} = \left(\sum_{t\in\mathcal{I}} \theta_t z_t z_t^\top\right)\left(\sum_{t\in\mathcal{I}} z_t z_t^\top\right)^{-1} + \left(\sum_{t\in\mathcal{I}} \omega_t z_t^\top\right)\left(\sum_{t\in\mathcal{I}} z_t z_t^\top\right)^{-1}.$$

The second term above is a martingale sum, since ω_t is zero mean and independent of z_t, and contributes to the variance of the estimator. However, the first term which contributes to the 'bias' is non-trivial. In the stationary case, $\theta_t = \theta$ and the first term becomes θ, which implies that the OLS estimator is unbiased. However, in the non-stationary case, the first term can be far from $\bar{\theta}$ even when all the θ_t are close to each other. This necessitates a fresh analysis of the OLS estimator in the non-stationary setting.

The key obstacle in the analysis of the estimation error $\left\|\widehat{\theta} - \bar{\theta}\right\|^2$, is that while z_t lives in $\mathbb{R}^{n+d}$, most of its variance is in the n-dimensional column space of $[I_n\ K^\top]^\top$. This is because the LQR dynamics naturally adds the noise w_{t-1} to arrive at the state x_t. In fact, this is precisely the reason we add the exploration noise ξ_t: to be able to distinguish changes in $\Theta_t = [A_t\ B_t]$ that are orthogonal to the column space of $[I_n\ K^\top]^\top$. However, this also means that we can not use a naive analysis based on a lower bound on the eigenvalues of the matrix $\sum_{t\in\mathcal{I}} z_t z_t^\top$.

Our approach to bounding the estimation error of the OLS estimator is to begin by looking at the one dimensional OLS problems parametrized by $v \in \mathbb{S}^{n+d} := \{v \in \mathbb{R}^{n+d}, \|v\| = 1\}$:

$$\lambda_v = \operatorname*{argmin}_{\lambda} \mathcal{L}(\bar{\theta} + \lambda \cdot v),$$

and argue that $|\lambda_v|$ are small for 'enough' directions v. That is, in enough directions, the minimizer of the 1-dimensional quadratic defined above is close to the candidate $\bar{\theta}$. Finally, we will show via an ϵ-net argument that this implies that the true OLS estimator $\widehat{\theta}$ is also close to $\bar{\theta}$.

Step 1: Decomposing the problem into orthogonal subspaces. Fixing a direction v, the first order conditions for the minimizer λ_v of $\mathcal{L}(\bar{\theta} + \lambda \cdot v)$ gives:

$$\lambda_v \sum_t \langle v, z_t \rangle^2 = \sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \langle v, z_t \rangle + \sum_t \omega_t \langle v, z_t \rangle . \tag{18}$$

For a vector $u \in \mathbb{R}^{n+d}$, let $u^{\parallel}$ and $u^{\perp}$ denote the projections onto the column space of $[I_n \; K^\top]^\top$ and its orthogonal space, respectively. That is,

$$u^{\parallel} = \begin{bmatrix} I_n \\ K \end{bmatrix} (I + K^\top K)^{-1} \begin{bmatrix} I_n \\ K \end{bmatrix}^\top u \quad \text{and} \quad u^{\perp} = u - u^{\parallel}.$$

Similarly, let $\widehat{u}^{\parallel}$ and $\widehat{u}^{\perp}$ denote the unit vectors in the direction $u^{\parallel}$ and $u^{\perp}$, respectively.

For analysis, it will be convenient to generalize the one dimensional problem of finding the minimizer on the line $\bar{\theta} + \lambda \cdot v$ to instead finding the minimizer in the plane $\bar{\theta} + \lambda^{\parallel}\widehat{v}^{\parallel} + \lambda^{\perp}\widehat{v}^{\perp}$, where we seek the optimal values of $\lambda^{\parallel}$ and $\lambda^{\perp}$. From (18), denoting $V := \sum_t z_t z_t^\top$, the Hessian of the corresponding quadratic loss function is given by

$$H_{\widehat{v}^{\parallel},\widehat{v}^{\perp}} = \begin{bmatrix} (\widehat{v}^{\parallel})^\top V \widehat{v}^{\parallel} & (\widehat{v}^{\parallel})^\top V \widehat{v}^{\perp} \\ (\widehat{v}^{\perp})^\top V \widehat{v}^{\parallel} & (\widehat{v}^{\perp})^\top V \widehat{v}^{\perp} \end{bmatrix} = \begin{bmatrix} \sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel} \right\rangle^2 & \sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel} \right\rangle \left\langle \widehat{v}^{\perp}, \xi_t^{\perp} \right\rangle \\ \sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel} \right\rangle \left\langle \widehat{v}^{\perp}, \xi_t^{\perp} \right\rangle & \sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp} \right\rangle^2 \end{bmatrix}. \tag{19}$$

A careful analysis on $H_{\widehat{v}^{\parallel},\widehat{v}^{\perp}}$ later yields the following lemma that indicates it suffices to consider the following two simpler cases: $v = \widehat{v}^{\parallel}$ and $v = \widehat{v}^{\perp}$.

Lemma C.1. *Let* $\lambda_{\widehat{v}^{\parallel}} = \operatorname{argmin}_\lambda \mathcal{L}(\bar{\theta} + \lambda \cdot \widehat{v}^{\parallel})$ *and* $\lambda_{\widehat{v}^{\perp}} = \operatorname{argmin}_\lambda \mathcal{L}(\bar{\theta} + \lambda \cdot \widehat{v}^{\perp})$. *It holds with probability at least* $1 - 11\delta$ *that*

$$\lambda_v^2 \le 2\lambda_{\widehat{v}^{\parallel}}^2 + 2\lambda_{\widehat{v}^{\perp}}^2.$$

Proof. Combining Lemma C.2 and Lemma C.3, it suffices to prove that

$$\max\left\{ \frac{\sum_t \left| \left\langle \widehat{v}^{\parallel}, z_t^{\parallel} \right\rangle \left\langle \widehat{v}^{\perp}, \xi_t^{\perp} \right\rangle \right|}{\sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel} \right\rangle^2}, \frac{\sum_t \left| \left\langle \widehat{v}^{\parallel}, z_t^{\parallel} \right\rangle \left\langle \widehat{v}^{\perp}, \xi_t^{\perp} \right\rangle \right|}{\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp} \right\rangle^2} \right\} \le \frac{1}{33}$$

holds with probability at least $1 - 11\delta$.

Note that $z_t = y_t + \xi_t = Mx_{t-1} + \xi_t = M\Theta_{t-1}z_{t-1} + Mw_{t-1} + \xi_t$. We have

$$z_t^{\parallel} = M\Theta_{t-1}z_{t-1} + Mw_{t-1} + \xi_t^{\parallel}, \quad z_t^{\perp} = \xi_t^{\perp}.$$

Step 0: Bound on $\sum_{t\in\mathcal{I}} \|z_t\|^2$. We begin with the following corollary of Lemma C.6: For $|\mathcal{I}| \ge 16 \ln \frac{1}{\delta}$, conditioned on $\max_{t\in\mathcal{I}} \|x_t\| \le x_u$, it holds with probability at least $1 - \delta$ that

$$\sum_{t\in\mathcal{I}} \|z_t\|^2 \le 2|\mathcal{I}| \left((1 + K_u^2)x_u^2 + 2\sigma_L^2\right) =: |\mathcal{I}|z_u^2.$$

Step 1: Upper bound on the numerator. Recall $\xi_t = \sigma_t \widetilde{\eta}_t$ and $\sigma_t \le \nu_1 =: \sigma_L$. Let

$$\sigma_{\widehat{v}^{\perp}}^2 = \mathbf{E}\left[(\widehat{v}^{\perp})^\top \begin{pmatrix} 0_n \\ \eta_t \end{pmatrix} \begin{pmatrix} 0_n \\ \eta_t \end{pmatrix}^\top \widehat{v}^{\perp} \right]$$

denote the variance of $\left\langle \widehat{v}^{\perp}, \widetilde{\eta}_t^{\perp} \right\rangle$. Write $\widehat{v}^{\perp}$ as $\widehat{v}^{\perp} = [(\widehat{v}_1^{\perp})^\top \; (\widehat{v}_2^{\perp})^\top]^\top$, where $\widehat{v}_1^{\perp} \in \mathbb{R}^n, \widehat{v}_2^{\perp} \in \mathbb{R}^d$, and $\left\|\widehat{v}_1^{\perp}\right\| + \left\|\widehat{v}_2^{\perp}\right\| = 1$. Since $\widehat{v}^{\perp}$ is a unit vector in the orthogonal space of the columns space of $[I_n \; K^\top]^\top$, we must have $\widehat{v}_1^{\perp} + K^\top \widehat{v}_2^{\perp} = 0$. Then $\left\|\widehat{v}_1^{\perp}\right\| = \left\|-K^\top \widehat{v}_2^{\perp}\right\| \le \|K\| \left\|\widehat{v}_2^{\perp}\right\|$ and hence $\left\|\widehat{v}_2^{\perp}\right\| \ge \frac{1}{1+\|K\|}$. We have $\sigma_{\widehat{v}^{\perp}}^2 = \mathbf{E}\left[(\widehat{v}_2^{\perp})^\top \eta_t \eta_t^\top \widehat{v}_2^{\perp}\right] = (\widehat{v}_2^{\perp})^\top I_d \widehat{v}_2^{\perp} = \left\|\widehat{v}_2^{\perp}\right\|^2 \ge \frac{1}{(1+\|K\|)^2}$. Also, $\sigma_{\widehat{v}^{\perp}}^2 \le 1$.

Applying a supermartingale argument, we get

$$\mathbf{Pr}\left[\left|\sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel}\right\rangle \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle\right| \geq \sigma_{\widehat{v}^{\perp}}\sigma_L\sqrt{2\ln\frac{1}{\delta}\sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel}\right\rangle^2}\right] \leq 2\delta. \tag{20}$$

Next, we lower bound the denominator.

Step 2: Lower bound for $\sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel}\right\rangle^2$. By direct computation,

$$\begin{aligned}\sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel}\right\rangle^2 &= \left\langle \widehat{v}^{\parallel}, M\Theta_{t-1}z_{t-1} + Mw_{t-1} + \xi_t^{\parallel}\right\rangle^2 \\ &\geq \sum_t \left\langle \widehat{v}^{\parallel}, Mw_{t-1}\right\rangle^2 + 2\left\langle \widehat{v}^{\parallel}, Mw_{t-1}\right\rangle\left\langle \widehat{v}^{\parallel}, M\Theta_{t-1}z_{t-1}\right\rangle \\ &\quad + 2\left\langle \widehat{v}^{\parallel}, Mw_{t-1}\right\rangle\left\langle \widehat{v}^{\parallel}, \xi_t^{\parallel}\right\rangle + 2\left\langle \widehat{v}^{\parallel}, \xi_t^{\parallel}\right\rangle\left\langle \widehat{v}^{\parallel}, M\Theta_{t-1}z_{t-1}\right\rangle.\end{aligned}$$

Let $\sigma_1^2 = (\widehat{v}^{\parallel})^\top MWM^\top\widehat{v}^{\parallel}$ denote the variance of $\left\langle \widehat{v}^{\parallel}, Mw_{t-1}\right\rangle$. Write $\widehat{v}^{\parallel} = Mx_v$, where $1 = \left\|\widehat{v}^{\parallel}\right\|^2 = \left\|[x_v^\top\ x_v^\top K^\top]\right\|^2 = \|x_v\|^2 + \|Kx_v\|^2$. Recalling that $W \succcurlyeq \psi^2 I_n$, we have

$$\begin{aligned}\sigma_1^2 = (\widehat{v}^{\parallel})^\top MWM^\top\widehat{v}^{\parallel} &\geq \psi^2\cdot x_v^\top M^\top MM^\top Mx_v \\ &= \psi^2\cdot x_v^\top(I+K^\top K)(I+K^\top K)x_v \\ &= \psi^2\left(x_v^\top x_v + 2x_v^\top K^\top Kx_v + x_v^\top K^\top KK^\top Kx_v\right) \\ &= \psi^2\left(\|x_v\|^2 + 2\|Kx_v\|^2 + \left\|K^\top Kx_v\right\|^2\right) \\ &\geq \psi^2(\|x_v\|^2 + \|Kx_v\|^2) \\ &= \psi^2.\end{aligned}$$

We also have $W \preccurlyeq \Psi^2 I_n$. Then

$$\begin{aligned}\sigma_1^2 = (\widehat{v}^{\parallel})^\top MWM^\top\widehat{v}^{\parallel} &\leq \Psi^2\cdot x_v^\top M^\top MM^\top Mx_v \\ &= \Psi^2\cdot x_v^\top(I+K^\top K)(I+K^\top K)x_v \\ &= \Psi^2\left(x_v^\top x_v + 2x_v^\top K^\top Kx_v + x_v^\top K^\top KK^\top Kx_v\right) \\ &= \Psi^2\left(\|x_v\|^2 + 2\|Kx_v\|^2 + \left\|K^\top Kx_v\right\|^2\right) \\ &\leq \Psi^2\left(\|x_v\|^2 + \|Kx_v\|^2 + \|K\|^2\|x_v\|^2 + \left\|K^\top\right\|^2\|Kx_v\|^2\right) \\ &\leq \Psi^2(1+\|K\|^2).\end{aligned}$$

By standard Laurent-Massart bounds, we get

$$\mathbf{Pr}\left[\sum_t \left\langle \widehat{v}^{\parallel}, Mw_{t-1}\right\rangle^2 \geq \Psi^2(1+\|K\|^2)\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right)\right] \leq \delta, \tag{21}$$

$$\mathbf{Pr}\left[\sum_t \left\langle \widehat{v}^{\parallel}, Mw_{t-1}\right\rangle^2 \leq \psi^2\left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})}\right)\right] \leq \delta. \tag{22}$$

Note that

$$\sigma_L^2\sum_t \left\langle \widehat{v}^{\parallel}, \widetilde{\eta}_t^{\parallel}\right\rangle^2 \geq \sum_t \left\langle \widehat{v}^{\parallel}, \xi_t^{\parallel}\right\rangle^2 = \sum_t \sigma_t^2\left\langle \widehat{v}^{\parallel}, \widetilde{\eta}_t^{\parallel}\right\rangle^2 \geq \sigma_{\mathcal{I}}^2\sum_t \left\langle \widehat{v}^{\parallel}, \widetilde{\eta}_t^{\parallel}\right\rangle^2.$$

Let $\sigma^2_{\widehat{v}^{\|}} = \mathbf{E}\left[(\widehat{v}^{\|})^\top \begin{pmatrix} 0_n \\ \eta_t \end{pmatrix} \begin{pmatrix} 0_n \\ \eta_t \end{pmatrix}^\top \widehat{v}^{\|}\right]$ denote the variance of $\left\langle \widehat{v}^{\|}, \widetilde{\eta}_t^{\perp} \right\rangle$. Write $\widehat{v}^{\|} = [\widehat{v}_1^{\|} \ \widehat{v}_2^{\|}]$, where $\widehat{v}_1^{\|} \in \mathbb{R}^n, \widehat{v}_2^{\|} \in \mathbb{R}^d$ and $\left\|\widehat{v}_1^{\|}\right\|, \left\|\widehat{v}_2^{\|}\right\| \le 1$. We have $\sigma^2_{\widehat{v}^{\|}} = \mathbf{E}\left[(\widehat{v}_2^{\|})^\top \eta_t \eta_t^\top \widehat{v}_2^{\|}\right] = (\widehat{v}_2^{\|})^\top I_d (\widehat{v}_2^{\|}) = \left\|\widehat{v}_2^{\|}\right\|^2 \le 1$.

Again, by standard Laurent-Massart bounds, we have

$$\mathbf{Pr}\left[\sum_t \left\langle \widehat{v}^{\|}, \xi_t^{\|} \right\rangle^2 \ge \sigma_L^2\right] \left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right) \le \delta,$$

$$\mathbf{Pr}\left[\sum_t \left\langle \widehat{v}^{\|}, \xi_t^{\|} \right\rangle^2 \le \sigma_{\mathcal{I}}^2 \sigma^2_{\widehat{v}^{\|}} \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})}\right)\right] \le \delta.$$

Applying a supermartingale argument, we get

$$\mathbf{Pr}\left[\left|\sum_t \left\langle \widehat{v}^{\|}, M w_{t-1} \right\rangle \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle\right| \ge \Psi(1 + \|K\|^2)^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta} \sum_t \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle^2}\right] \le 2\delta, \quad (23)$$

$$\mathbf{Pr}\left[\left|\sum_t \left\langle \widehat{v}^{\|}, \xi_t^{\|} \right\rangle \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle\right| \ge \sigma_L \sqrt{2\ln\frac{1}{\delta} \sum_t \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle^2}\right] \le 2\delta, \quad (24)$$

$$\mathbf{Pr}\left[\left|\sum_t \left\langle \widehat{v}^{\|}, M w_{t-1} \right\rangle \left\langle \widehat{v}^{\|}, \xi_t^{\|} \right\rangle\right| \ge \sigma_L \sqrt{2\ln\frac{1}{\delta} \sum_t \left\langle \widehat{v}^{\|}, M w_{t-1} \right\rangle^2}\right] \le 2\delta. \quad (25)$$

Combining (21) and (25), we have

$$\mathbf{Pr}\left[\left|\sum_t \left\langle \widehat{v}^{\|}, M w_{t-1} \right\rangle \left\langle \widehat{v}^{\|}, \xi_t^{\|} \right\rangle\right| \ge \sigma_L \Psi(1 + \|K\|^2)^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right)}\right] \le 3\delta.$$

Combining the inequalities above, it holds with probability at least $1 - 8\delta$ that

$$\begin{aligned}
\sum_t \left\langle \widehat{v}^{\|}, z_t^{\|} \right\rangle^2 &\ge \sum_t \left\langle \widehat{v}^{\|}, M w_{t-1} \right\rangle^2 + 2\left\langle \widehat{v}^{\|}, M w_{t-1} \right\rangle \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle \\
&\quad + 2\left\langle \widehat{v}^{\|}, M w_{t-1} \right\rangle \left\langle \widehat{v}^{\|}, \xi_t^{\|} \right\rangle + 2\left\langle \widehat{v}^{\|}, \xi_t^{\|} \right\rangle \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle \\
&\ge \psi^2 \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})}\right) - 2\Psi(1 + \|K\|^2)^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta} \sum_t \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle^2} \\
&\quad - 2\sigma_L \sqrt{2\ln\frac{1}{\delta} \sum_t \left\langle \widehat{v}^{\|}, M\Theta_{t-1} z_{t-1} \right\rangle^2} \\
&\quad - 2\sigma_L \Psi(1 + \|K\|^2)^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right)} \\
&\ge \psi^2 \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})}\right) - 2\Psi(1 + \|K\|^2)^{\frac{1}{2}} \|M\| \, \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} \\
&\quad - 2\sigma_L \|M\| \, \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}}
\end{aligned}$$

$$
\begin{aligned}
&- 2\sigma_L \Psi(1+\|K\|^2)^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right)} \\
&\geq \psi^2 \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})}\right) - 2\Psi(1+\|K\|^2)^{\frac{1}{2}} \|M\| \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} \\
&\quad - 2\sigma_L \|M\| \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} - 2\sigma_L \Psi(1+\|K\|^2)^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta} 2I} \\
&= \psi^2 |\mathcal{I}| - \sqrt{2|\mathcal{I}|\ln\frac{1}{\delta}} \left(2\psi^2 + 2\Psi(1+\|K\|^2)^{\frac{1}{2}} \|M\| \Theta_u z_u\right) \\
&\qquad +2\sigma_L \|M\| \Theta_u z_u + 2\sqrt{2}\sigma_L \Psi(1+\|K\|^2)^{\frac{1}{2}}\Big) =: \Lambda_1. \qquad (26)
\end{aligned}
$$

In arriving at (26) we have used the assumption $|\mathcal{I}| \geq 16 \ln \frac{1}{\delta}$, which implies

$$
2\ln\frac{1}{\delta}\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right) \leq 2|\mathcal{I}|.
$$

To get a further cleaner expression, we further assume

$$
|\mathcal{I}| \geq 32\psi^{-4} \ln\frac{1}{\delta}\left(\psi^2 + \Psi(1+\|K\|^2)^{\frac{1}{2}} \|M\| \Theta_u z_u + \sigma_L \|M\| \Theta_u z_u + \sqrt{2}\sigma_L \Psi(1+\|K\|^2)^{\frac{1}{2}}\right)^2, \qquad (27)
$$

which in turn implies $|\mathcal{I}| \geq 16 \ln \frac{1}{\delta}$, under which the bound simplifies to

$$
\Lambda_1 \geq \frac{\psi^2}{2}|\mathcal{I}|. \qquad (28)
$$

Combining (20) and (26), it holds with probability at least $1 - 10\delta$ that

$$
\frac{\sum_t \left|\left\langle \widehat{v}^{\|}, z_t^{\|}\right\rangle \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle\right|}{\sum_t \left\langle \widehat{v}^{\|}, z_t^{\|}\right\rangle^2} \leq \frac{\sigma_{\widehat{v}^{\perp}} \sigma_L \sqrt{2\ln\frac{1}{\delta}}}{\sqrt{\sum_t \left\langle \widehat{v}^{\|}, z_t^{\|}\right\rangle^2}} \leq \sigma_L 2\sqrt{\ln\frac{1}{\delta}} \psi^{-1} |\mathcal{I}|^{-\frac{1}{2}} \leq \frac{1}{33}, \qquad (29)
$$

provided that

$$
|\mathcal{I}| \geq 4 \cdot 33^2 \sigma_L^2 \ln\frac{1}{\delta} \psi^{-2}. \qquad (30)
$$

Step 3: Lower bound on $\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle^2$. Recall that $\xi_t = \sigma_t \widetilde{\eta}_t$ and $\sigma_t \leq \sigma_L$. Let

$$
\sigma_{\widehat{v}^{\perp}}^2 = \mathbf{E}\left[(\widehat{v}^{\perp})^{\top} \begin{pmatrix} 0_n \\ \eta_t \end{pmatrix} \begin{pmatrix} 0_n \\ \eta_t \end{pmatrix}^{\top} \widehat{v}^{\perp}\right]
$$

denote the variance of $\left\langle \widehat{v}^{\perp}, \widetilde{\eta}_t^{\perp}\right\rangle$. Write $\widehat{v}^{\perp}$ as $\widehat{v}^{\perp} = [(\widehat{v}_1^{\perp})^{\top} \ (\widehat{v}_2^{\perp})^{\top}]^{\top}$, where $\widehat{v}_1^{\perp} \in \mathbb{R}^n, \widehat{v}_2^{\perp} \in \mathbb{R}^d$ and $\left\|\widehat{v}_1^{\perp}\right\| + \left\|\widehat{v}_2^{\perp}\right\| = 1$. Since $\widehat{v}^{\perp}$ is a unit vector in the orthogonal space of the columns space of $[I_n \ K^{\top}]^{\top}$, we must have $\widehat{v}_1^{\perp} + K^{\top}\widehat{v}_2^{\perp} = 0$. Then $\left\|\widehat{v}_1^{\perp}\right\| = \left\|-K^{\top}\widehat{v}_2^{\perp}\right\| \leq \|K\| \left\|\widehat{v}_2^{\perp}\right\|$ and hence $\left\|\widehat{v}_2^{\perp}\right\| \geq \frac{1}{1+\|K\|}$. We have $\sigma_{\widehat{v}^{\perp}}^2 = \mathbf{E}\left[(\widehat{v}_2^{\perp})^{\top} \eta_t \eta_t^{\top} \widehat{v}_2^{\perp}\right] = (\widehat{v}_2^{\perp})^{\top} I_d \widehat{v}_2^{\perp} = \left\|\widehat{v}_2^{\perp}\right\|^2 \geq \frac{1}{(1+\|K\|)^2}$. Also, $\sigma_{\widehat{v}^{\perp}}^2 \leq 1$.

By standard Laurent-Massart bounds, the denominator can be bounded from below by

$$
\mathbf{Pr}\left[\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle^2 \leq \frac{\sigma_{\mathcal{I}}^2}{(1+\|K\|)^2}\left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})}\right)\right] \leq \delta.
$$

Plugging in our choice of $\sigma_{\mathcal{I}}^2$, with probability at least $1-2\delta$ we have

$$\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp} \right\rangle^2 \geq \frac{1}{(1+\|K\|)^2}\sqrt{\frac{C_0}{|\mathcal{I}|}}\left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})}\right) = \frac{1}{(1+\|K\|)^2}C_0^{\frac{1}{2}}(|\mathcal{I}|^{\frac{1}{2}} - 2\sqrt{\ln\frac{1}{\delta}}) =: \Lambda_2.$$

Under the assumption that $|\mathcal{I}| \geq 16\ln\frac{1}{\delta}$, we get:

$$\Lambda_2 \geq \frac{1}{2}|I|^{\frac{1}{2}}\frac{1}{(1+\|K\|)^2}C_0^{\frac{1}{2}}. \tag{31}$$

Combining with (20), it holds with probability at least $1-4\delta$ that

$$\begin{aligned}
\frac{\sum_t \left|\left\langle \widehat{v}^{\|}, z_t^{\|}\right\rangle \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle\right|}{\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle^2} &\leq \frac{\sigma_{\widehat{v}^{\perp}}\sigma_L\sqrt{2\ln\frac{1}{\delta}\sum_t \left\langle \widehat{v}^{\|}, z_t^{\|}\right\rangle^2}}{\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle^2} \\
&\leq \frac{\sigma_{\widehat{v}^{\perp}}\sigma_L\sqrt{2\ln\frac{1}{\delta}\sum_t \left\|z_t^{\|}\right\|^2}}{\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle^2} \\
&\leq \frac{\sigma_L\sqrt{2\ln\frac{1}{\delta}}|\mathcal{I}|^{\frac{1}{2}}z_u}{\sum_t \left\langle \widehat{v}^{\perp}, \xi_t^{\perp}\right\rangle^2} \\
&\leq \frac{\sigma_L\sqrt{2\ln\frac{1}{\delta}}|\mathcal{I}|^{\frac{1}{2}}z_u}{\frac{1}{2}|\mathcal{I}|^{\frac{1}{2}}\frac{1}{(1+\|K\|)^2}C_0^{\frac{1}{2}}} \\
&= 2\sqrt{2}\sqrt{\ln\frac{1}{\delta}}\sigma_L z_u(1+\|K\|)^2 C_0^{-\frac{1}{2}} \\
&= 2\sqrt{2}\sqrt{\ln\frac{1}{\delta}}z_u(1+\|K\|)^2 L^{-\frac{1}{2}} \\
&\leq \frac{1}{33},
\end{aligned}$$

provided L satisfies:

$$L \geq 66^2 \cdot 2 \cdot \ln\frac{1}{\delta}\cdot z_u^2(1+\|K\|)^4. \tag{32}$$

□

Step 2: Bounding λ_v when $v = \widehat{v}^{\|}$. Noting that $z_t^{\|} = y_t + \xi_t^{\|}$, we can rewrite the left hand side of (18) in this case as $\lambda_{\widehat{v}^{\|}}\sum_t \left\langle v, z_t^{\|}\right\rangle^2$, and the right hand side of (18) as

$$\sum_t \omega_t \left\langle v, z_t^{\|}\right\rangle + \sum_t \left\langle \theta_t - \bar{\theta}, z_t\right\rangle \cdot \left\langle v, z_t^{\|}\right\rangle.$$

Therefore,

$$|\lambda_{\widehat{v}^{\|}}| \leq \frac{\left|\sum_t \left\langle \theta_t - \bar{\theta}, z_t\right\rangle \cdot \left\langle v, z_t^{\|}\right\rangle\right|}{\sum_t \left\langle v, z_t^{\|}\right\rangle^2} + \frac{\left|\sum_t \omega_t \left\langle v, z_t^{\|}\right\rangle\right|}{\sum_t \left\langle v, z_t^{\|}\right\rangle^2}. \tag{33}$$

By (26) and (28) it holds with probability at least $1 - 8\delta$ that

$$\sum_t \left\langle \widehat{v}^{\parallel}, z_t^{\parallel} \right\rangle^2 \geq \Lambda_1 \geq \frac{\psi^2}{2}|\mathcal{I}|,$$

if we have

$$|\mathcal{I}| \geq 32\psi^{-4} \ln \frac{1}{\delta} \left(\psi^2 + \Psi(1 + \|K\|^2)^{\frac{1}{2}} \|M\| \Theta_u z_u + \sigma_L \|M\| \Theta_u z_u + \sqrt{2}\sigma_L \Psi (1 + \|K\|^2)^{\frac{1}{2}} \right)^2.$$

It remains to upper bound the numerators of the two terms in (33). For the first term, Cauchy-Schwartz inequality gives:

$$\begin{aligned}
\left| \sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, z_t^{\parallel} \right\rangle \right| &\leq \sqrt{\sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle^2} \sqrt{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2} \\
&\leq \max_t \left| \theta_t - \bar{\theta} \right| \cdot \sqrt{\sum_t \|z_t\|^2} \sqrt{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2} \\
&\leq \Delta_{\mathcal{I}} \sqrt{z_u^2 |\mathcal{I}|} \sqrt{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2} \\
&\leq \Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2}. \qquad (34)
\end{aligned}$$

Plugging this into (33) gives

$$\frac{\left| \sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, z_t^{\parallel} \right\rangle \right|}{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2} \leq \frac{\Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}}}{\sqrt{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2}} \leq \Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}} \Lambda_1^{-\frac{1}{2}}. \qquad (35)$$

For the second term, let $\omega_t \overset{d}{=} \mathcal{N}(0, \psi_i^2)$. Note that assuming $w_t(i)$ is ψ_i^2 sub-Gaussian suffices. By the assumption on w_t, a standard supermartingale argument implies that

$$\mathbf{Pr}\left[\left| \sum_t \omega_t \left\langle v, z_t^{\parallel} \right\rangle \right| \geq \psi_i \sqrt{2 \ln \frac{1}{\delta} \sum_t \left\langle v, z_t^{\parallel} \right\rangle^2} \right] \leq 2\delta.$$

Plugging this into (33) gives

$$\frac{\left| \sum_t \omega_t \left\langle v, z_t^{\parallel} \right\rangle \right|}{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2} \leq \frac{\psi_i \sqrt{2 \ln \frac{1}{\delta}}}{\sqrt{\sum_t \left\langle v, z_t^{\parallel} \right\rangle^2}} \leq \psi_i \sqrt{2 \ln \frac{1}{\delta}} \Lambda_1^{-\frac{1}{2}}. \qquad (36)$$

Finally, we conclude it holds with probability at least $1 - 10\delta$ that

$$\begin{aligned}
|\lambda_{\widehat{v}^{\parallel}}| &\leq \Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}} \Lambda_1^{-\frac{1}{2}} + \psi_i \sqrt{2 \ln \frac{1}{\delta}} \Lambda_1^{-\frac{1}{2}} \\
&\leq \sqrt{2}\psi^{-1} \Delta_{\mathcal{I}} z_u + 2\psi^{-1} \psi_i \sqrt{\ln \frac{1}{\delta}} |\mathcal{I}|^{-\frac{1}{2}}. \qquad (37)
\end{aligned}$$

Step 3: Bounding λ_v when $v = \widehat{v}^{\perp}$. Noting that $z_t^{\perp} = \xi_t^{\perp}$, we can rewrite the left hand side of (18) as $\lambda_{\widehat{v}^{\perp}} \sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2$, and the right hand side of (18) as

$$\sum_t \omega_t \left\langle v, \xi_t^{\perp} \right\rangle + \sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, \xi_t^{\perp} \right\rangle.$$

Therefore,

$$|\lambda_{\widehat{v}^{\perp}}| \le \frac{\left|\sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, \xi_t^{\perp} \right\rangle\right|}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2} + \frac{\left|\sum_t \omega_t \left\langle v, \xi_t^{\perp} \right\rangle\right|}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2}. \tag{38}$$

For the first term, we observe that $\xi_t^{\perp}$ is normally distributed and is independent of z_t. Applying a supermartingale argument, we get

$$\mathbf{Pr}\left[\left|\sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, \xi_t^{\perp} \right\rangle\right| \ge \sigma_L \sqrt{2 \ln \frac{1}{\delta} \sum_t \langle \theta_t - \theta, z_t \rangle^2}\right] \le 2\delta,$$

and hence

$$\mathbf{Pr}\left[\left|\sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, \xi_t^{\perp} \right\rangle\right| \ge \sigma_L \Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2 \ln \frac{1}{\delta}}\right] \le 3\delta.$$

Then the first term is upper bounded as

$$\frac{\left|\sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, \xi_t^{\perp} \right\rangle\right|}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2} \le \frac{\sqrt{\sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle^2} \sqrt{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2}}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2} \le \sigma_L \Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2 \ln \frac{1}{\delta}} \Lambda_2^{-1}.$$

For the second term, a supermartingale argument implies that

$$\mathbf{Pr}\left[\left|\sum_t \omega_t \left\langle v, \xi_t^{\perp} \right\rangle\right| \ge \psi_i \sqrt{2 \ln \frac{1}{\delta} \sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2}\right] \le 2\delta.$$

Then

$$\frac{\left|\sum_t \omega_t \left\langle v, \xi_t^{\perp} \right\rangle\right|}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2} \le \frac{\psi_i \sqrt{2 \ln \frac{1}{\delta} \sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2}}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2} \le \psi_i \sqrt{2 \ln \frac{1}{\delta}} \Lambda_2^{-\frac{1}{2}}.$$

Finally, we conclude it holds with probability at least $1 - 6\delta$ that

$$\begin{aligned} |\lambda_{\widehat{v}^{\perp}}| &\le \frac{\left|\sum_t \left\langle \theta_t - \bar{\theta}, z_t \right\rangle \cdot \left\langle v, \xi_t^{\perp} \right\rangle\right|}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2} + \frac{\left|\sum_t \omega_t \left\langle v, \xi_t^{\perp} \right\rangle\right|}{\sum_t \left\langle v, \xi_t^{\perp} \right\rangle^2} \\ &\le \sigma_L \Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2 \ln \frac{1}{\delta}} \Lambda_2^{-1} + \psi_i \sqrt{2 \ln \frac{1}{\delta}} \Lambda_2^{-\frac{1}{2}}. \end{aligned}$$

Assuming $|\mathcal{I}| \ge 16 \ln \frac{1}{\delta}$ and using the bound on Λ_2 from (31), we have

$$|\lambda_{\widehat{v}^{\perp}}| \le 2\sigma_L \Delta_{\mathcal{I}} z_u \sqrt{2 \ln \frac{1}{\delta}} (1 + \|K\|)^2 C_0^{-\frac{1}{2}} + 2\psi_i \sqrt{\ln \frac{1}{\delta}} (1 + \|K\|) C_0^{-\frac{1}{4}} |\mathcal{I}|^{-\frac{1}{4}}. \tag{39}$$

Combining Lemma C.1, (37) and (39), we have that

$$\begin{aligned} \lambda_v^2 &\le 2\lambda_{\widehat{v}^{\parallel}}^2 + 2\lambda_{\widehat{v}^{\perp}}^2 \\ &\le 4\Delta_{\mathcal{I}}^2 z_u^2 |\mathcal{I}| \Lambda_1^{-1} + 8\psi_i^2 \ln \frac{1}{\delta} \Lambda_1^{-1} + 8\sigma_L^2 \Delta_{\mathcal{I}}^2 (1 + \|K\|)^2 x_u^2 \ln \frac{1}{\delta} |\mathcal{I}| \Lambda_2^{-2} + 8\psi_i^2 \ln \frac{1}{\delta} \Lambda_2^{-1} \end{aligned}$$

$$= 4\Delta_{\mathcal{I}}^2 z_u^2 \left(|\mathcal{I}|\Lambda_1^{-1} + 2C_0^{\frac{1}{2}} \ln\frac{1}{\delta}|\mathcal{I}|^{\frac{1}{2}}\Lambda_2^{-2} \right) + 8\psi_i^2 \ln\frac{1}{\delta}\left(\Lambda_1^{-1} + \Lambda_2^{-1}\right)$$

$$= 4\Delta_{\mathcal{I}}^2 z_u^2 |\mathcal{I}| \left(\Lambda_1^{-1} + 2\sigma_L^2 \ln\frac{1}{\delta}\Lambda_2^{-2} \right) + 8\psi_i^2 \ln\frac{1}{\delta}\left(\Lambda_1^{-1} + \Lambda_2^{-1}\right)$$

$$\leq 8\psi^{-2}\Delta_{\mathcal{I}}^2 z_u^2 + 16\psi^{-2}\psi_i^2 \ln\frac{1}{\delta}|\mathcal{I}|^{-1} + 32\sigma_L^2\Delta_{\mathcal{I}}^2 z_u^2 \ln\frac{1}{\delta}(1+\|K\|)^4 C_0^{-1} + 16\psi_i^2 \ln\frac{1}{\delta}(1+\|K\|)^2 C_0^{-\frac{1}{2}}|\mathcal{I}|^{-\frac{1}{2}}$$

$$= 8\Delta_{\mathcal{I}}^2 z_u^2 \left(\psi^{-2} + 4\sigma_L^2 \ln\frac{1}{\delta}(1+\|K\|)^4 C_0^{-1} \right) + 16\psi_i^2 \ln\frac{1}{\delta}\left(\psi^{-2}|\mathcal{I}|^{-1} + (1+\|K\|)^2 C_0^{-\frac{1}{2}}|\mathcal{I}|^{-\frac{1}{2}} \right)$$

holds with probability at least $1 - 27\delta$.

Hence, we conclude that

$$\begin{aligned} |\lambda_v| &\leq 2\sqrt{2}\Delta_{\mathcal{I}} z_u \sqrt{\psi^{-2} + 4\sigma_L^2 \ln\frac{1}{\delta}(1+\|K\|)^4 C_0^{-1}} + 4\psi_i \sqrt{\ln\frac{1}{\delta}} \sqrt{\psi^{-2}|\mathcal{I}|^{-1} + (1+\|K\|)^2 C_0^{-\frac{1}{2}}|\mathcal{I}|^{-\frac{1}{2}}} \\ &\leq 2\sqrt{2}\Delta_{\mathcal{I}} z_u \sqrt{\psi^{-2} + 4\sigma_L^2 \ln\frac{1}{\delta}(1+\|K\|)^4 C_0^{-1}} + 4\psi_i \sqrt{\ln\frac{1}{\delta}} \sqrt{\psi^{-2} + (1+\|K\|)^2 C_0^{-\frac{1}{2}}}|\mathcal{I}|^{-\frac{1}{4}} \\ &=: \widetilde{C}_1\Delta_{\mathcal{I}} + \widetilde{C}_2|\mathcal{I}|^{-\frac{1}{4}}, \end{aligned} \tag{40}$$

holds with probability at least $1 - 27\delta$, where we define

$$\begin{aligned} \widetilde{C}_1 &= 2\sqrt{2} z_u \sqrt{\psi^{-2} + 4\sigma_L^2 \ln\frac{1}{\delta}(1+\|K\|)^4 C_0^{-1}}, \\ \widetilde{C}_2 &= 4\psi_i \sqrt{\ln\frac{1}{\delta}} \sqrt{\psi^{-2} + (1+\|K\|)^2 C_0^{-\frac{1}{2}}}. \end{aligned} \tag{41}$$

Step 4: An ϵ-net argument. To summarize, thus far we have shown that for any fixed direction $v \in \mathbb{R}^{n+d}$, and for any row θ_i of the parameter matrix Θ, the minimizer of the one-dimensional quadratic loss function satisfies (40). We next invoke Lemma C.5, which implies that if this statement holds for all v in an ϵ-net of the $(n+d)$-dimensional unit sphere (where ϵ depends on the condition number κ_u of the Hessian $\sum_t z_t z_t^\top$ as $\epsilon \leq \frac{1}{5(1+\kappa_u)}$), then the Frobenius norm of the OLS estimator $\widehat{\theta}_i$ and $\bar{\theta}_i$ is upper bounded by $\frac{5}{3}\bar{\lambda}$. The bound on the condition number κ_u of the Hessian is proved in Lemma C.8. We make this more formal next.

We fix ε as the confidence level. First, substituting $\delta = \varepsilon/6$ in Lemma C.8 gives that with probability at least $1-\varepsilon/6$, the condition number of the Hessian is bounded from above as $\kappa_{\mathcal{I}} \leq C_9\sqrt{|\mathcal{I}|}$ provided $\varepsilon \leq 18/100$ and

$$|\mathcal{I}| \geq \frac{2000}{9}\left(2(n+d)\log\frac{6}{\varepsilon} + (n+d)\log\frac{\bar{x}^2(1+\|K\|^2) + \sigma_L^2}{\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2 + 2\|K\|^2}\right\}} \right). \tag{42}$$

We will thus choose $\epsilon = (5(1 + C_9\sqrt{I}))^{-1}$ in the ϵ-net result of Lemma C.5 and Lemma C.4. This gives an upper bound on the cardinality of the ϵ-net of $\left(1 + \frac{4}{\epsilon}\right)^{n+d} \leq (10(1 + C_9\sqrt{|\mathcal{I}|}))^{n+d}$.

Applying (40) by substituting $\delta = \frac{\varepsilon}{54n(10(1+C_9\sqrt{|\mathcal{I}|}))^{n+d}}$, it holds with probability at least $1-\varepsilon$ that for every row of Θ we have

$$\left\|\widehat{\theta}_{\mathcal{I}} - \bar{\theta}\right\|_F = \left\|\widehat{\theta}_{\mathcal{I}} - \bar{\theta}\right\| \leq \frac{5}{3}\left(\widetilde{C}_1\Delta_{\mathcal{I}} + \widetilde{C}_2|\mathcal{I}|^{-\frac{1}{4}}\right).$$

Combining n rows, we have

$$\begin{aligned}\left\|\widehat{\Theta}_{\mathcal{I}} - \bar{\Theta}\right\|_F &\le \frac{5\sqrt{n}}{3}\left(\widetilde{C}_1\Delta_{\mathcal{I}} + \widetilde{C}_2|\mathcal{I}|^{-\frac{1}{4}}\right)\\ &\le \check{C}_1\Delta_{\mathcal{I}} + \check{C}_2|\mathcal{I}|^{-\frac{1}{4}},\end{aligned}$$

where

$$\begin{aligned}\check{C}_1 &= 5z_u\sqrt{n}\left(\psi^{-1} + \sqrt{4\sigma_L^2(1+\|K\|)^4C_0^{-1}}\sqrt{\ln\frac{1}{\varepsilon} + \ln 54n + (n+d)\ln\left(10(1+C_9\sqrt{|\mathcal{I}|})\right)}\right),\\ \check{C}_2 &= 7\psi_i\sqrt{n}\sqrt{\psi^{-2} + (1+\|K\|)^2C_0^{-\frac{1}{2}}}\sqrt{\ln\frac{1}{\varepsilon} + \ln 54n + (n+d)\ln\left(10(1+C_9\sqrt{|\mathcal{I}|})\right)}.\end{aligned} \tag{43}$$

Under our choice of $C_0 = O(\log T)$, $L = O(\log^3 T)$, $\sigma_L^2 = \sqrt{C_0/L}$, $\varepsilon = O(T^{-3})$, and assuming that T is large enough so that L satisfies conditions (27), (30), (32), and (42), both $\check{C}_1, \check{C}_2$ are $O(\sqrt{\log T})$.

C.2 Proof of OLS Concentration for $\mathcal{B}_{i,0}$ (Lemma 6.2)

To bound the estimation error of the OLS estimator of the warm-up block $\mathcal{I} = \mathcal{B}_{i,0}$, we look at the one dimensional problem (18) again:

$$\lambda_v\sum_t \langle v, z_t\rangle^2 = \sum_t \omega_t\langle v, z_t\rangle + \sum_t \left\langle \theta_t - \bar{\theta}, z_t\right\rangle\cdot\langle v, z_t\rangle.$$

Since we are using a sequence of sequentially strongly stable policies $\{K_t^{\text{stab}}\}$ instead of a fixed policy, we do not have a fixed column space anymore. Nevertheless, the $O(1)$ exploration noise $\xi_t = \nu_0\widetilde{\eta}_t$ enables us to bound the error similar to the case $v = v^{\|}$. Recall that we set $\nu_0 = 1$ in Algorithm 1.

Let $M_t \coloneqq \begin{bmatrix} I_n \\ K_t^{\text{stab}}\end{bmatrix}$. Note that

$$\begin{aligned}\sum_t \langle v, z_t\rangle^2 &= \langle v, M_t\Theta_{t-1}z_{t-1} + M_t w_{t-1} + \xi_t\rangle^2\\ &\ge \sum_t \langle v, M_t w_{t-1} + \xi_t\rangle^2 + 2\langle v, M_t w_{t-1}\rangle\langle v, M_t\Theta_{t-1}z_{t-1}\rangle\\ &\quad + 2\langle v, M_t w_{t-1}\rangle\langle v, \xi_t\rangle + 2\langle v, \xi_t\rangle\langle v, M_t\Theta_{t-1}z_{t-1}\rangle.\end{aligned}$$

Let $v = v_t^{\|} + v_t^{\perp}$, where $v_t^{\|}$ is the projection of v onto the column space generated by M_t. We have

$$\langle v, M_t w_{t-1} + \xi_t\rangle^2 \ge \left\langle v_t^{\|}, M_t w_{t-1}\right\rangle^2 + \left\langle v_t^{\perp}, \widetilde{\eta}_t\right\rangle^2.$$

Let $\sigma_{1,t}^2 = (v_t^{\|})^\top M_t W M_t^\top v_t^{\|}$ denote the variance of $\left\langle v_t^{\|}, M_t w_{t-1}\right\rangle$. Write $v_t^{\|} = M_t x_{v,t}$, where $\left\|v_t^{\|}\right\|^2 = \left\|[x_{v,t}^\top\ x_{v,t}^\top K_t^\top]\right\|^2 = \left\|x_{v,t}\right\|^2 + \left\|(K_t^{\text{stab}})x_{v,t}\right\|^2$. Recall that $W \succcurlyeq \psi^2 I_n$. We have

$$\begin{aligned}\sigma_{1,t}^2 &= (v_t^{\|})^\top M_t W M_t^\top v_t^{\|}\\ &\ge \psi^2\cdot x_{v,t}^\top M_t^\top M_t M_t^\top M_t x_{v,t}\\ &= \psi^2\cdot x_{v,t}^\top(I + (K_t^{\text{stab}})^\top(K_t^{\text{stab}}))(I + (K_t^{\text{stab}})^\top(K_t^{\text{stab}}))x_{v,t}\\ &= \psi^2\left(x_{v,t}^\top x_{v,t} + 2x_{v,t}^\top(K_t^{\text{stab}})_t^\top(K_t^{\text{stab}})x_v + x_{v,t}^\top(K_t^{\text{stab}})^\top(K_t^{\text{stab}})(K_t^{\text{stab}})^\top(K_t^{\text{stab}})x_{v,t}\right)\\ &= \psi^2\left(\left\|x_{v,t}\right\|^2 + 2\left\|(K_t^{\text{stab}})x_{v,t}\right\|^2 + \left\|(K_t^{\text{stab}})^\top(K_t^{\text{stab}})x_{v,t}\right\|^2\right)\end{aligned}$$

$$\geq \psi^2(\left\|x_{v,t}\right\|^2 + \left\|(K_t^{\text{stab}})x_{v,t}\right\|^2)$$
$$= \psi^2 \left\|v_t^{\parallel}\right\|^2. \tag{44}$$

Let $\sigma_{2,t}^2$ denote the variance of $\left\langle v_t^{\perp}, \widetilde{\eta}_t \right\rangle$. Write $v_t^{\perp}$ as $v_t^{\perp} = [(v_{t,1}^{\perp})^\top \ (v_{t,2}^{\perp})^\top]^\top$, where $v_{t,1}^{\perp} \in \mathbb{R}^n$, $v_{t,2}^{\perp} \in \mathbb{R}^d$, and $\left\|v_{t,1}^{\perp}\right\| + \left\|v_{t,2}^{\perp}\right\| = \left\|v_t^{\perp}\right\|$. Since $\widehat{v}^{\perp}$ is in the orthogonal space of the columns space of $[I_n \ K_t^\top]^\top$, we must have $v_{t,1}^{\perp} + K_t^\top v_{t,2}^{\perp} = 0$. Then $\left\|v_{t,1}^{\perp}\right\| = \left\|-K^\top v_{t,2}^{\perp}\right\| \leq \left\|K_t^{\text{stab}}\right\| \left\|v_{t,2}^{\perp}\right\|$ and hence $\left\|v_{t,2}^{\perp}\right\| \geq \frac{1}{1+\left\|K_t^{\text{stab}}\right\|}$. We have

$$\sigma_{2,t}^2 = \mathbf{E}\left[(v_{t,2}^{\perp})^\top \eta_t \eta_t^\top v_{t,2}^{\perp}\right] = (v_{t,2}^{\perp})^\top I_d v_{t,2}^{\perp} = \left\|v_{t,2}^{\perp}\right\|^2 \geq \frac{1}{(1+\left\|K_t^{\text{stab}}\right\|)^2} \left\|v_t^{\perp}\right\|^2 \geq \frac{1}{(1+K_u)^2} \left\|v_t^{\perp}\right\|^2. \tag{45}$$

Recall that $\left\|v_t^{\parallel}\right\|^2 + \left\|v_t^{\perp}\right\|^2 = 1$. Combing (44) and (45), we have

$$\begin{aligned}
\mathbf{E}\left[\langle v, \xi_t + M_t w_{t-1}\rangle^2\right] &\geq \mathbf{E}\left[\left\langle v_t^{\parallel}, M_t w_{t-1}\right\rangle^2 + \left\langle v_t^{\perp}, \xi_t\right\rangle^2\right] \\
&\geq \psi^2 \left\|v_t^{\parallel}\right\|^2 + \frac{v_0^2}{(1+K_u)^2} \left\|v_t^{\perp}\right\|^2 \\
&\geq \min\left\{\psi^2, \frac{v_0^2}{(1+K_u)^2}\right\} \\
&=: \sigma_v^2.
\end{aligned}$$

Using the standard Laurent-Massart bound implies

$$\mathbf{Pr}\left[\sum_t \langle v, \xi_t + M_t w_{t-1}\rangle^2 \leq \sigma_v^2 \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})}\right)\right] \leq \delta.$$

Let $v = [v_1^\top \ v_2^\top]^\top$, where $v_1 \in \mathbb{R}^n$ and $v_2 \in \mathbb{R}^d$. Then

$$\begin{aligned}
\mathbf{E}\left[\langle v, M_t w_{t-1}\rangle^2\right] &= \mathbf{E}\left[\left\langle v_1 + (K_t^{\text{stab}})^\top v_2, w_t\right\rangle^2\right] \\
&\leq \mathbf{E}\left[2\langle v_1, w_t\rangle^2 + 2\left\langle (K_t^{\text{stab}})^\top v_2, w_t\right\rangle^2\right] \\
&\leq 2(\|v_1\|^2 + K_u^2 \|v_2\|^2)\mathbf{E}\left[\|w_t\|^2\right] \\
&\leq 2(1+K_u^2)\mathbf{E}\left[\|w_t\|^2\right] \\
&=: \sigma_{\text{stab}}^2.
\end{aligned}$$

By the standard Laurent-Massart bound, we have

$$\mathbf{Pr}\left[\sum_t \langle v, M_t w_{t-1}\rangle^2 \geq \sigma_{\text{stab}}^2 \left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}| \ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right)\right] \leq \delta.$$

Applying a supermartingale argument, we get

$$\mathbf{Pr}\left[\left|\sum_t \langle v, M_t w_{t-1}\rangle \langle v, M\Theta_{t-1} z_{t-1}\rangle\right| \geq \sigma_{\text{stab}} \sqrt{2\ln\frac{1}{\delta} \sum_t \langle v, M_t \Theta_{t-1} z_{t-1}\rangle^2}\right] \leq 2\delta,$$

$$\mathbf{Pr}\left[\left|\sum_t \langle v, \xi_t\rangle \langle v, M_t \Theta_{t-1} z_{t-1}\rangle\right| \geq \sigma_v v_0 \sqrt{2\ln\frac{1}{\delta} \sum_t \langle v, M_t \Theta_{t-1} z_{t-1}\rangle^2}\right] \leq 2\delta,$$

$$\mathbf{Pr}\left[\left|\sum_t \langle v, M_t w_{t-1}\rangle \langle v, \xi_t\rangle\right| \geq \sigma_v \nu_0 \sqrt{2\ln\frac{1}{\delta}\sum_t \langle v, M_t w_{t-1}\rangle^2}\right] \leq 2\delta.$$

By direct computation, we have that

$$\begin{aligned}
\sum_t \langle v, z_t\rangle^2 &\geq \sum_t \left\langle v_t^{\parallel}, M_t w_{t-1}\right\rangle^2 + \sum_t \left\langle v_t^{\perp}, \xi_t\right\rangle^2 + 2\langle v, M_t w_{t-1}\rangle \langle v, M_t \Theta_{t-1} z_{t-1}\rangle \\
&\quad + 2\langle v, M_t w_{t-1}\rangle \langle v, \xi_t\rangle + 2\langle v, \xi_t\rangle \langle v, M_t \Theta_{t-1} z_{t-1}\rangle \\
&\geq \sigma_v^2 \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})}\right) - 2\sigma_{\text{stab}} \sqrt{2\ln\frac{1}{\delta}\sum_t \langle v, M_t \Theta_{t-1} z_{t-1}\rangle^2} \\
&\quad - 2\sigma_v \nu_0 \sqrt{2\ln\frac{1}{\delta}\sum_t \langle v, M_t \Theta_{t-1} z_{t-1}\rangle^2} \\
&\quad - 2\sigma_v \nu_0 \sqrt{2\ln\frac{1}{\delta}\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right)} \\
&\geq \sigma_v^2 \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})}\right) - 2\sigma_{\text{stab}} M_u \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} \\
&\quad - 2\sigma_v \nu_0 M_u \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} - 2\sigma_v \nu_0 \sqrt{2\ln\frac{1}{\delta}\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right)} \\
&\geq \sigma_v^2 \left(|\mathcal{I}| - 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})}\right) - 2\sigma_{\text{stab}} M_u \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} \\
&\quad - 2\sigma_v \nu_0 M_u \Theta_u z_u |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} - 2\sigma_v \nu_0 \sqrt{2\ln\frac{1}{\delta} 2|\mathcal{I}|} \\
&= \sigma_v^2 |\mathcal{I}| - |\mathcal{I}|^{\frac{1}{2}} \sqrt{2\ln\frac{1}{\delta}} \left(\sqrt{2}\sigma_v^2 + 2\sigma_{\text{stab}} M_u \Theta_u z_u + 2\sigma_v \nu_0 M_u \Theta_u z_u + 2\sqrt{2}\sigma_v \nu_0\right) \\
&=: \Lambda_{\text{stab}},
\end{aligned}$$

holds with probability at least $1 - 8\delta$, where $M_u := \max_t \|M_t\|$. We can simplify the bound to

$$\Lambda_{\text{stab}} \geq \frac{1}{2}\sigma_v^2 L$$

given the warm-up block $\mathcal{I}$ is long enough:

$$|\mathcal{I}| = L \geq 8\sigma_v^{-4} \ln\frac{1}{\delta}\left(\sqrt{2}\sigma_v^2 + 2\sigma_{\text{stab}} M_u \Theta_u z_u + 2\sigma_v \nu_0 M_u \Theta_u z_u + 2\sqrt{2}\sigma_v \nu_0\right)^2. \tag{46}$$

Note that this condition is stricter than $L \geq 16\ln\frac{1}{\delta}$.

By a supermartingale argument, we have

$$\mathbf{Pr}\left[\left|\sum_t \omega_t \langle v, z_t\rangle\right| \geq \psi_i \sqrt{2\ln\frac{1}{\delta}\sum_t \langle v, z_t\rangle^2}\right] \leq 2\delta. \tag{47}$$

Hence, it holds with probability at least $1 - 9\delta$ that

$$|\lambda_v| \leq \frac{\left|\sum_t \left\langle \theta_t - \bar{\theta}, z_t\right\rangle \cdot \langle v, z_t\rangle\right|}{\sum_t \langle v, z_t\rangle^2} + \frac{\left|\sum_t \omega_t \langle v, z_t\rangle\right|}{\sum_t \langle v, z_t\rangle^2}$$

$$\leq \Delta_{\mathcal{I}} z_u |\mathcal{I}|^{\frac{1}{2}} \Lambda_{\text{stab}}^{-\frac{1}{2}} + \psi_i \sqrt{2 \ln \frac{1}{\delta}} \Lambda_{\text{stab}}^{-\frac{1}{2}}$$
$$\leq \sqrt{2} \sigma_v^{-1} z_u \Delta_{\mathcal{I}} + 2\psi_i \sigma_v^{-1} \sqrt{\ln \frac{1}{\delta}} |\mathcal{I}|^{-\frac{1}{2}}$$
$$=: \widetilde{C}_{1,\text{stab}} \Delta_{\mathcal{I}} + \widetilde{C}_{2,\text{stab}} |\mathcal{I}|^{-\frac{1}{2}}. \tag{48}$$

Here we have used Cauchy-Schwartz inequality, (47), and the upper bounds $|\theta_t - \bar{\theta}| \leq \Delta_{\mathcal{I}}$, $\left\|K_t^{\text{stab}}\right\| \leq K_u$ and the definition of z_u.

Finally, we combine (48) with an ϵ-net argument as in the proof of Lemma 6.1. We let ε be the confidence parameter. Choosing $\delta = \frac{\varepsilon}{6}$ in Lemma C.8 we get an upper bound on the condition number of the Hessian as $\kappa_0 = O(1)$. Setting $\epsilon = 5(1+\kappa_0)$, and applying (48) with $\delta = \frac{\varepsilon}{18n(5(1+\kappa_0))^{n+d}}$, it holds with probability at least $1-\varepsilon$ that for every row we have,

$$\left\|\widehat{\theta} - \bar{\theta}\right\|_F = \left\|\widehat{\theta}_{\mathcal{B}_{i,0}} - \bar{\theta}\right\| \leq \frac{5}{3}\left(\widetilde{C}_{1,\text{stab}} \Delta_{\mathcal{B}_{i,0}} + \widetilde{C}_{2,\text{stab}} |\mathcal{B}_{i,0}|^{-\frac{1}{4}}\right).$$

Combining the n rows, we have

$$\left\|\widehat{\Theta}_{\mathcal{B}_{i,0}} - \bar{\Theta}\right\|_F \leq \frac{5\sqrt{n}}{3}\left(\widetilde{C}_{1,\text{stab}} \Delta_{\mathcal{B}_{i,0}} + \widetilde{C}_{2,\text{stab}} |\mathcal{B}_{i,0}|^{-\frac{1}{4}}\right)$$
$$\leq \check{C}_{1,\text{stab}} \Delta_{\mathcal{B}_{i,0}} + \check{C}_{2,\text{stab}} |\mathcal{B}_{i,0}|^{-\frac{1}{4}},$$

where

$$\check{C}_{1,\text{stab}} = 3\sqrt{n}\sigma_v^{-1} z_u,$$
$$\check{C}_{2,\text{stab}} = 4\psi_i \sigma_v^{-1} \sqrt{n} \sqrt{\ln \frac{1}{\varepsilon} + \ln 18n + (n+d)\ln 5(1+\kappa_0)}. \tag{49}$$

For our choice of x_u, v_0, L, assuming that T is large enough so that L satisfies (46), and setting $\varepsilon = O(T^{-3})$, both $\check{C}_{1,\text{stab}}, \check{C}_{2,\text{stab}}$ are $O(\sqrt{\ln T})$.

C.3 Lemmas on the geometry of the Hessian

The following lemma gives a sufficient condition under which to upper bound the distance of a point $p' = (x', y')$ from the minimizer of a quadratic form $f(x, y)$, it suffices to upper bound the distance of p' from the minimizers of the one-dimensional functions $h(x) = f(x, y')$ and $g(y) = f(x', y)$. In a nut shell, the lemma states that if the level sets of f are "almost axis-parallel" (the precise requirement being given by the condition number of the Hessian), then it suffices to obtain upper bounds on the one-dimensional minimizers.

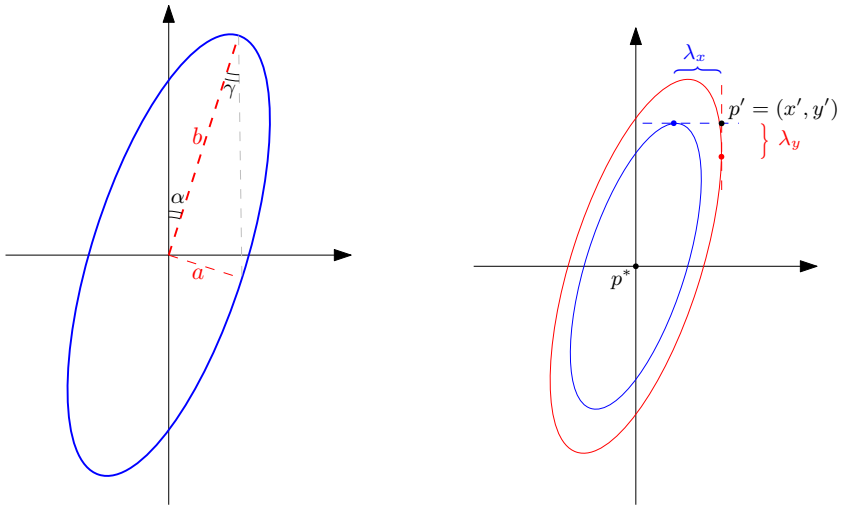
Lemma C.2. *Let $f(x, y)$ be a quadratic form with Hessian $H = \begin{bmatrix} A^2 & C \\ C & B^2 \end{bmatrix} \succ 0$. Let the level sets of $f(x, y)$ be given by ellipses that are clockwise rotation by an angle $\alpha \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right)$ of axis parallel ellipses: $\frac{x^2}{a^2} + \frac{y^2}{b^2} = r^2$. Define γ such that $\tan\gamma = \frac{\min\{a,b\}}{\max\{a,b\}} \leq 1$.*

For a given point $p' = (x', y')$, define

$$x'' = \operatorname*{argmin}_x f(x, y'), \qquad y'' = \operatorname*{argmin}_x f(x', y),$$

and $\lambda_x = x' - x''$, $\lambda_y = y' - y''$. Let $p^* = (x^*, y^*) = \operatorname{argmin}_{(x,y)} f(x, y)$ be the true minimizer of $f(\cdot)$. If (i) $\tan\gamma \geq \frac{5}{8}$, or (ii) $|\tan\alpha| \leq \frac{\tan^2\gamma}{4}$, then

$$\lambda_x^2 + \lambda_y^2 \geq \frac{1}{2}\left\|p' - p^*\right\|^2.$$



(a) A level set of $f(x, y)$

(b) Illustration of 1-d minimizers in Lemma C.2

Fig. 3. Illustration of the setup for Lemma C.2. The figure on the left shows one level set of the quadratic form $f(x, y)$; a α rotation of an axis-parallel ellipse with principal axes of lengths a and b. The figure on the right is a visual illustration of λ_x, λ_y in the Lemma statement. For example, the blue ellipse denotes the level set on which the minimizer (x'', y') (blue dot) lies, giving $\lambda_x = x' - x''$.

Proof. We will assume that the true minimizer $p^* = 0$ without loss of generality, so that we need to prove that $\lambda_x^2 + \lambda_y^2 \geq \frac{1}{2} \cdot (x'^2 + y'^2)$. We will also assume without loss of generality that $a \leq b$. Furthermore, in the proof we will focus on the case $\alpha \in [0, \pi/4]$ and $x', y' \geq 0$, as illustrated in Figure 3(b). This is indeed the hardest case: if $x' \leq 0, y' \geq 0, 0 \leq \alpha \leq \pi/4$, then $|\lambda_y| \geq |y'|, |\lambda_x| \geq x'$ and the Lemma follows. The other cases are symmetric to one of the above.

We begin by finding an expression for y''. Observe that at the point (x', y''), the tangent to the level set is parallel to the y-axis. If we now imagine rotating the level set, along with the point (x', y'') counter-clockwise by an angle α, so that the level set becomes axis-parallel and the point (x', y'') moves to $(\widehat{x}, \widehat{y})$, then the tangent at $(\widehat{x}, \widehat{y})$ to this axis-parallel ellipse has a slope of $m = -1/\tan\alpha$. We can now obtain one relationship between $\widehat{x}, \widehat{y}$ by differentiating the equation for the level set with respect to x:

$$\frac{d}{dx}\left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right)\Bigg|_{(\widehat{x},\widehat{y})} = 0 \implies \widehat{y} = \widehat{x}\frac{b^2}{a^2}\tan\alpha.$$

Let $\tan\beta := \frac{b^2}{a^2}\tan\alpha$, so that $\widehat{y} = \widehat{x}\tan\beta$. Since (x', y'') is obtained by clockwise rotation of $(\widehat{x}, \widehat{y})$ by α, we have

$$y'' = x'\tan(\beta - \alpha) = x'\frac{\tan\beta - \tan\alpha}{1 + \tan\beta\tan\alpha}.$$

Substituting $\tan\beta = \frac{b^2}{a^2}\tan\alpha = \tan\alpha/\tan^2\gamma$:

$$y'' = x'\tan\alpha\frac{1 - \tan^2\gamma}{\tan^2\gamma + \tan^2\alpha}. \tag{50}$$

A similar analysis gives,

$$x'' = y' \tan\alpha \frac{1-\tan^2\gamma}{1+\tan^2\gamma\tan^2\alpha}. \tag{51}$$

Writing (x', y') in polar coordinates (r, θ), we get

$$\lambda_x = x' - x'' = r\left(\cos\theta - \sin\theta\cdot\tan\alpha\frac{1-\tan^2\gamma}{1+\tan^2\gamma\tan^2\alpha}\right), \tag{52}$$

$$\lambda_y = y' - y'' = r\left(\sin\theta - \cos\theta\cdot\tan\alpha\frac{1-\tan^2\gamma}{\tan^2\gamma+\tan^2\alpha}\right). \tag{53}$$

By our assumptions, $\tan\alpha \geq 0$ and $0 \leq \tan\gamma \leq 1$. One can further verify that under the condition $\tan\alpha \leq \frac{\tan^2\gamma}{4}$, we have

$$0 \leq \tan\alpha\frac{1-\tan^2\gamma}{1+\tan^2\gamma\tan^2\alpha} \leq 1/4,$$

and

$$0 \leq \tan\alpha\frac{1-\tan^2\gamma}{\tan^2\gamma+\tan^2\alpha} \leq 1/4.$$

To see why, the above inequalities can be rearranged into the following quadratic inequalities in $\tan\alpha$ for any fixed value $\tan\gamma$:

$$\left(\frac{\tan^2\gamma}{1-\tan^2\gamma}\right)\tan^2\alpha - 4\tan\alpha + \frac{1}{1-\tan^2\gamma} \geq 0,$$
$$\left(\frac{1}{1-\tan^2\gamma}\right)\tan^2\alpha - 4\tan\alpha + \frac{\tan^2\gamma}{1-\tan^2\gamma} \geq 0.$$

It can then be shown that $\frac{\tan^2\gamma}{4}$ is a lower bound on the smaller roots of the quadratic expressions on the left hand side above. In fact, the second condition above is stricter than the first (when $\tan\gamma \leq 1$), and $\frac{\tan^2\gamma}{4}$ is a linear approximation to the smaller root $(= 2(1-\tan^2\gamma) - \sqrt{4\tan^2\gamma - 9\tan\gamma + 4})$ in the vicinity of $\tan^2\gamma = 0$ for the second inequality above. In fact, if $\tan\gamma \geq \frac{5}{8}$, then the two inequalities are always true.

Finally,

$$\begin{aligned}\lambda_x^2 + \lambda_y^2 &= (x' - x'')^2 + (y' - y'')^2 \\ &\geq r^2\left[\sin^2\theta + \cos^2\theta - 2\sin\theta\cos\theta\tan\alpha\left(\frac{1-\tan^2\gamma}{1+\tan^2\gamma\tan^2\alpha} + \frac{1-\tan^2\gamma}{\tan^2\gamma+\tan^2\alpha}\right)\right] \\ &\geq r^2\left(1 - \frac{1}{2}\sin 2\theta\right) \;\geq\; \frac{r^2}{2}.\end{aligned}$$

□

The Hessian for the quadratic described in Lemma C.2 is given by:

$$H = \begin{bmatrix} A^2 & C \\ C & B^2 \end{bmatrix} = \begin{bmatrix} \frac{\cos^2\alpha}{a^2} + \frac{\sin^2\alpha}{b^2} & \sin\alpha\cos\alpha\left(\frac{1}{b^2} - \frac{1}{a^2}\right) \\ \sin\alpha\cos\alpha\left(\frac{1}{b^2} - \frac{1}{a^2}\right) & \frac{\sin^2\alpha}{a^2} + \frac{\cos^2\alpha}{b^2} \end{bmatrix}. \tag{54}$$

Lemma C.3. *The Hessian in (54) satisfies the conditions of Lemma C.2 if* $\frac{|C|}{\min\{A^2, B^2\}} \leq \frac{1}{33}$.

Proof. Without loss of generality, assume $a \le b$, so that with $\alpha \in [-\pi/4, \pi/4]$, we have $B^2 \le A^2$. To neaten the exposition, we will further focus on $\alpha \in [0, \pi/4]$, since only $|\sin\alpha|$ and $|\cos\alpha|$ are involved in verifying the condition.

It suffices to prove that

$$\frac{C}{B^2} = \frac{1}{2} \cdot \frac{\sin 2\alpha}{a^2/b^2} \cdot \frac{1 - \frac{a^2}{b^2}}{1 + \left(\frac{b^2}{a^2} - 1\right)\sin^2\alpha} \le \frac{1}{33} \tag{55}$$

implies $\tan\alpha \le \frac{a^2}{4b^2}$, *under the assumption* that $\frac{a^2}{b^2} \le 25/64$, since otherwise $\tan\gamma \ge 5/8$ and the first condition in Lemma C.2 is satisfied. Since under this assumption $1 - \frac{a^2}{b^2} \ge \frac{39}{64}$, (55) implies

$$\frac{\sin 2\alpha}{a^2/b^2} \cdot \frac{1}{1 + \left(\frac{b^2}{a^2} - 1\right)\sin^2\alpha} \le \frac{128}{39 \cdot 33} \le \frac{1}{10}.$$

Rearranging,

$$\frac{b^2}{a^2} \le \frac{1 - \sin^2\alpha}{10\sin 2\alpha - \sin^2\alpha} \le \frac{1 - \sin^2\alpha}{10\sqrt{2}\sin\alpha - \sin^2\alpha}, \tag{56}$$

since in the interval $\alpha \in [0, \pi/4]$ we have $\sin 2\alpha \ge \sqrt{2}\sin\alpha$. Since $\frac{b^2}{a^2} \ge \frac{64}{25} \ge 2$, we get the quadratic inequality

$$\sin^2\alpha - 20\sqrt{2}\sin\alpha + 1 \ge 0,$$

which implies $\sin\alpha \le 0.04$ and therefore $\tan\alpha = \frac{\sin\alpha}{\cos\alpha} \le 1.01\sin\alpha$.

Starting from (56) again,

$$\left(\frac{b^2}{a^2} - 1\right)\sin^2\alpha - 10\sqrt{2}\frac{b^2}{a^2}\sin\alpha + 1 \ge 0,$$

which has roots $\frac{10\sqrt{2}b^2/a^2 \pm \sqrt{200b^4/a^4 - 4b^2/a^4 + 4}}{2(b^2/a^2 - 1)}$. Since $b^2/a^2 \ge 64/25$, we observe that the larger root is greater than $5\sqrt{2}$, and hence the smaller root is bounded above by $\frac{1}{5\sqrt{2}} \cdot \frac{1}{\frac{b^2}{a^2} - 1} \le \frac{64/39}{5\sqrt{2}} \cdot \frac{a^2}{b^2}$, which is also an upper bound on $\sin\alpha$. In the last inequality we have again used $b^2/a^2 \ge 64/25$. Finally, $\tan\alpha \le 1.01\sin\alpha \le 1.01 \cdot \frac{64/39}{5\sqrt{2}} \cdot \frac{a^2}{b^2} \le \frac{a^2}{4b^2}$ as needed. □

The following lemma adapted from the volume argument of ϵ-net w.r.t. Euclidean norm [38] gives an upper bound for the covering numbers of the sphere using ϵ-net w.r.t. tan.

Lemma C.4. *Let $N(\epsilon, \mathbb{S}^{n-1})$ be the minimal cardinality of an ϵ-net of $\mathbb{S}^{n-1}$ such that for every unit vector $v \in \mathbb{S}^{n-1}$, there exists a $v_\epsilon \in \mathcal{S}_\epsilon$ such that the* tan *of the angle between v and v_ϵ is in $[-\epsilon, \epsilon]$. If $\epsilon \le \frac{1}{5}$, we have that*

$$N(\epsilon, \mathbb{S}^{n-1}) \le \left(1 + \frac{4}{\epsilon}\right)^n.$$

Proof. Choose $\mathcal{N}_\epsilon$ to be the maximal subset of $\mathbb{S}^{n-1}$ such that the tan of the angle between two arbitrary vectors v_1, v_2 is larger than ϵ. By the maximality property, $\mathcal{N}_\epsilon$ is an ϵ-net. Moreover, using the fact that $x \ge \frac{1}{2}\tan(x)$ if $x \le \frac{1}{5}$, the balls of radii $\frac{\epsilon}{4}$ centered at the points in $\mathcal{N}_\epsilon$ are disjoint. Let $\mathcal{B}_{n,2}$ denote the unit Euclidean ball in $\mathbb{R}^n$ centered at the origin. By comparing the volumes, it holds that

$$N(\epsilon, \mathbb{S}^{n-1}) \cdot \left(\frac{\epsilon}{4}\right)^n vol(\mathcal{B}_n) = N(\epsilon, \mathbb{S}^{n-1}) \cdot vol\left(\frac{\epsilon}{4}\mathcal{B}_n\right) \le vol\left(\left(1 + \frac{\epsilon}{4}\right)\mathcal{B}_n\right) = \left(1 + \frac{\epsilon}{4}\right)^n \cdot vol(\mathcal{B}_n).$$

Hence we conclude

$$N(\epsilon, \mathbb{S}^{n-1}) \leq \frac{(1+\frac{\epsilon}{4})^n}{\frac{\epsilon}{4}^n} = \left(1+\frac{4}{\epsilon}\right)^n.$$

□

Lemma C.5. *Let $\mathcal{L}(\theta) : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form loss function with positive definite Hessian H and minimizer θ^*. Let $\kappa \geq 1$, the condition number, denote the ratio of the largest to the smallest eigenvalue of H. Let $\mathcal{S}_\epsilon \subset \mathbb{S}^{n-1}$ be an 'ϵ-net' of the n-dimensional unit sphere so that for every unit vector $v \in \mathbb{S}^{n-1}$, there exists a $v_\epsilon \in \mathcal{S}_\epsilon$ such that the* tan *of the angle between v and v_ϵ is in $[-\epsilon, \epsilon]$ and $\epsilon \leq \frac{1}{5+2\kappa}$. Let $\bar{\theta}$ be an approximate minimizer of $\mathcal{L}$, and let λ_v denote the minimizer of the scalar quadratic function $\mathcal{L}_v(\lambda) = \mathcal{L}(\bar{\theta} + \lambda v)$. If for all $v_\epsilon \in \mathcal{S}_\epsilon$, $|\lambda_{v_\epsilon}| \leq \bar{\lambda}$, then $|\bar{\theta} - \theta^*| \leq \frac{\bar{\lambda}}{1-2(1+\kappa)\epsilon}$.*

Proof. Let $v = (v_x, v_y)$ denote the unit vector in the direction $\bar{\theta} - \theta^*$, and let $u = (u_x, u_y) \in \mathcal{S}_\epsilon$ satisfy the condition in the Lemma statement with respect to v. That is, if $m_v = \frac{v_y}{v_x} = \tan\alpha$ and $m_u = \frac{u_y}{u_x} = \tan\beta$, then $\tan(\beta - \alpha) = \epsilon'$ with $\epsilon' \in [-\epsilon, \epsilon]$. Let P denote the point $\bar{\theta} + \lambda_u u$. The points $\bar{\theta}, \theta^*, P$ define a plane and the subsequent analysis will be restricted to this plane. Without loss of generality, let us translate and rotate our co-ordinate system so that θ^* is at the origin, the level sets of the loss function $\mathcal{L}$ restricted to the plane of interest have the form $\frac{x^2}{a^2} + \frac{y^2}{b^2} = r^2$ (with $\frac{1}{\kappa} \leq \frac{a^2}{b^2} \leq \kappa$), and $\bar{\theta} = (\bar{\theta}_x, \bar{\theta}_y) = r_{\bar{\theta}} \cdot v$ lie in the positive quadrant.

Using the fact that the point $P = \bar{\theta} + \lambda_u u$ is tangent to the level set, we get

$$-\lambda_u = \frac{\bar{\theta}_x b^2 + \bar{\theta}_y m_u a^2}{a^2 m_u^2 + b^2} \cdot \sqrt{1+m_u^2} = r_{\bar{\theta}} . \frac{v_x b^2 + v_y m_u a^2}{a^2 m_u^2 + b^2} \cdot \sqrt{1+m_u^2}.$$

Since $m_u = \tan\beta = \frac{\tan\alpha + \epsilon'}{1-\epsilon'\tan\alpha} = \frac{m_v + \epsilon'}{1-\epsilon' m_v}$, some calculations give,

$$\begin{aligned} |\lambda_u| &= r_{\bar{\theta}} . \frac{v_x b^2 + v_y m_u a^2}{a^2 m_u^2 + b^2} \cdot \sqrt{1+m_u^2} \\ &= r_{\bar{\theta}} \sqrt{1+\epsilon'^2} \left[1 - \epsilon' \underbrace{\frac{(m_v+\epsilon')a^2 - m_v(1-m_v\epsilon')b^2}{(m_v+\epsilon')^2 a^2 + (1-\mu\epsilon')^2 b^2}}_{D} \right]. \end{aligned}$$

We can bound D as:

$$D \leq \underbrace{\frac{(m_v+\epsilon')a^2}{(m_v+\epsilon')^2 a^2 + (1-\mu\epsilon')^2 b^2}}_{D_1} + \underbrace{\frac{m_v|1-m_v\epsilon'|b^2}{(m_v+\epsilon')^2 a^2 + (1-\mu\epsilon')^2 b^2}}_{D_2}.$$

Assuming $|\epsilon'| \leq \epsilon \leq 1/5$, if $m_v \geq 1$, then $D_1 \leq 1$ and $D_2 \leq \frac{2b^2}{a^2} \leq 2\kappa$. If $m_v \leq 1$, then $D_1 \leq \frac{2a^2}{b^2} \leq 2\kappa$ and $D_2 \leq 2$. Therefore, $D \leq 2(1+\kappa)$, finally giving:

$$|\bar{\theta} - \theta^*| = r_{\bar{\theta}} \leq \frac{|\lambda_u|}{\sqrt{1+\epsilon'^2}} \cdot \frac{1}{1-2(1+\kappa)\epsilon'} \leq \frac{\bar{\lambda}}{1-2(1+\kappa)\epsilon}.$$

□

C.4 Bound on $\sum_{t\in\mathcal{I}} \|z_t\|^2$

Lemma C.6. *For a $\delta \in (0,1)$ and an interval $\mathcal{I}$ lying within some block $\mathcal{B}_{ij}$ with $j \geq 1$ and $|\mathcal{I}| \geq 16 \ln \frac{1}{\delta}$, it holds with probability at least $1-\delta$ that*

$$\sum_{t\in\mathcal{I}} \|z_t\|^2 \leq |\mathcal{I}| \left(2\left((1+K_u^2)\max_{t\in\mathcal{I}} \|x_t\|^2 + 2\sigma_L^2\right)\right).$$

In particular, for any $\bar{x} \geq 0$, conditioned on $\max_{t\in\mathcal{I}} \|x_t\| \leq \bar{x}$, we have $\sum_{t\in\mathcal{I}} \|z_t\|^2 \leq |\mathcal{I}|\bar{z}$, where $\bar{z} := \sqrt{2\left((1+K_u^2)\bar{x}^2 + 2\sigma_L^2\right)}$. Here $\sigma_L^2 := v_1^2 = \sqrt{C_0/2L}$.

Proof. Recall that $z_t = y_t + \xi_t$ and $\|K_t\|, \left\|K_t^{\text{stab}}\right\| \leq K_u$. We have

$$\begin{aligned}\sum_{t\in\mathcal{I}} \|z_t\|^2 &= \sum_{t\in\mathcal{I}} \|y_t + \xi_t\|^2 \\ &\leq 2\sum_{t\in\mathcal{I}} \|y_t\|^2 + 2\sum_{t\in\mathcal{I}} \|\xi_t\|^2 \\ &\leq 2(1+K_u^2)|\mathcal{I}| \max_{t\in\mathcal{I}} \|x_t\|^2 + 2\sum_{t\in\mathcal{I}} \|\xi_t\|^2.\end{aligned}$$

By a standard Laurent-Massart bound, we have:

$$\mathbf{Pr}\left[\sum_{t\in\mathcal{I}} \|\xi_t\|^2 \geq \sigma_L^2\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}| \ln\left(\frac{1}{\delta}\right)} + 2\ln\left(\frac{1}{\delta}\right)\right)\right] \leq \delta.$$

Using the fact that $\left(|\mathcal{I}| + 2\sqrt{|\mathcal{I}|\ln(\frac{1}{\delta})} + 2\ln(\frac{1}{\delta})\right) \leq 2|\mathcal{I}|$ when $|\mathcal{I}| \geq 16\ln\frac{1}{\delta}$, plugging the above in the bound derived above for $\sum_t \|z_t\|^2$ indicates that

$$\sum_{t\in\mathcal{I}} \|z_t\|^2 \leq 2\left((1+K_u^2)\max_{t\in\mathcal{I}} \|x_t\|^2 + 2\sigma_L^2\right)|\mathcal{I}|$$

holds with probability at least $1-\delta$. The right hand side of the bound given above is a random variable since it involves $\max_{t\in\mathcal{I}} \|x_t\|$, and can instead be interpreted as saying that for any $\bar{x}$, conditioned on the event $\max_{t\in\mathcal{I}} \|x_t\| \leq \bar{x}$, $\sum_{t\in\mathcal{I}} \|z_t\|^2 \leq 2\left((1+K_u^2)\bar{x}^2 + 2\sigma_L^2\right)|\mathcal{I}| =: |\mathcal{I}|\bar{z}$. □

Definition C.7. Define $z_u := \sqrt{2\left((1+K_u^2)x_u^2 + 2\sigma_L^2\right)}$, where x_u is defined in Algorithm 1.

C.5 Bound on the condition number of $\sum_{t\in\mathcal{I}} z_t z_t^\top$

Lemma C.8. *For an arbitrary interval $\mathcal{I}$, denote design matrix $\Upsilon_{\mathcal{I}} = \sum_{t\in\mathcal{I}} z_t z_t^\top$ and its condition number $\kappa = \lambda_{max}(\Upsilon_{\mathcal{I}})/\lambda_{min}(\Upsilon_{\mathcal{I}})$.*

(i) Let $\mathcal{I}$ be an interval within a block $\mathcal{B}_{i,j}$ in Algorithm 1. Define $\bar{x} = \max_{t\in\mathcal{I}} \|x_t\|$, and $\bar{z}$ as in Lemma C.6. If we have

$$|\mathcal{I}| \geq \frac{2000}{9}\left(2(n+d)\log\frac{1}{\delta} + (n+d)\log\frac{\bar{x}^2(1+\|K\|^2)+\sigma_L^2}{\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}}\right),$$

then for $\delta \le 3/100$, it holds with probability at least $1-3\delta$ that the condition number is upper bounded as

$$\kappa \le \frac{\bar{z}^2|\mathcal{I}|}{\frac{9|\mathcal{I}|}{1600}\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}} = \frac{1600\bar{z}^2}{9\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}}.$$

Define $\kappa_{\mathcal{I}}$ be the bound above when $\bar{x} = x_u$:

$$\kappa_{\mathcal{I}} := \frac{1600 z_u^2}{9\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}},$$

whence it follows that $\kappa_{\mathcal{I}} \le C_9\sqrt{|\mathcal{I}|}$ for some problem-dependent constant (independent of T) C_9.

(ii) For any warm-up block $\mathcal{B}_{i,0}$ in in Algorithm 1, with sequentially strongly stablizing policies $\{K_t^{stab}\}$ such that $\left\|K_t^{stab}\right\| \le K_u$, if

$$|\mathcal{B}_{i,0}| \ge \frac{2000}{9}\left(2(n+d)\log\frac{1}{\delta} + (n+d)\log\frac{\bar{x}^2(1+K_u^2)+\sigma_0^2}{\sigma_0^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_0^2+2(K_l^{stab})^2}\right\}}\right),$$

we have for $\delta \le 3/100$,

$$\kappa \le \frac{\bar{z}^2|\mathcal{B}_{i,0}|}{\frac{9|\mathcal{B}_{i,0}|}{1600}\sigma_0^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{v_0^2+2\|K\|^2}\right\}} = \frac{1600\bar{z}^2}{9v_0^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{v_0^2+2\|K\|^2}\right\}}.$$

Define κ_0 to be the bound above when $\bar{x} = x_u$:

$$\kappa_0 := \frac{1600 z_u^2}{9v_0^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{v_0^2+2\|K\|^2}\right\}},$$

from where it follows that $\kappa_0 \le C_{10}\ln T$ for some problem dependent constant C_{10}.

Proof. We need to bound $\lambda_{\min}(\Upsilon_{\mathcal{I}})$ and $\lambda_{\max}(\Upsilon_{\mathcal{I}})$ separately. By direct computation and Lemma C.6, it holds with probability at least $1-\delta$ that

$$\lambda_{\max}(\Upsilon_{\mathcal{I}}) \le \mathrm{Tr}(\Upsilon_{\mathcal{I}}) = \sum_{t\in\mathcal{I}} z_t z_t^\top = \sum_{t\in\mathcal{I}} \|z_t\|^2 \le \bar{z}^2|\mathcal{I}|.$$

In the sequel, we bound $\lambda_{\min}(\Upsilon_{\mathcal{I}})$ from below by specifying the choice of Υ_0 such that $\Upsilon_0 \preceq \Upsilon_{\mathcal{I}}$ with high probability using Lemma C.9. Note that $z_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(\bar{z}_t, \Sigma_t)$, where $\bar{z}_t$ and Σ_t are measurable and

$$\Sigma_t \succeq \begin{bmatrix} \psi^2 I_n & \psi^2 I_n K^\top \\ \psi^2 K I_n & \psi^2 K I_d K^\top + \sigma_t^2 I_d \end{bmatrix}.$$

By Dean et al. [15, Lemma F. 6], we have

$$\lambda_{\min}(\Sigma_t) \ge \sigma_t^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_t^2+2\|K\|^2}\right\} \ge \sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}.$$

Moreover, we have

$$\mathrm{Tr}(\mathbf{E}[\Upsilon_{\mathcal{I}}]) = \mathbf{E}\left[\sum_{t\in\mathcal{I}} \|x_t\|^2 + \|u_t\|^2\right]$$

$$\leq \mathbf{E}\left[\sum_{t\in\mathcal{I}} \|x_t\|^2 + \|K_t\|^2 \|x_t\|^2 + \sigma_L^2\right]$$
$$\leq |\mathcal{I}| \left(\bar{x}^2(1+\|K\|^2) + \sigma_L^2\right).$$

Setting $\mathcal{E} = \Omega$ (the probability space) and

$$\Upsilon_0 = \frac{9|\mathcal{I}|}{1600}\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2 + 2\|K\|^2}\right\} I_{d+n},$$

Lemma C.9 implies if

$$|\mathcal{I}| \geq \frac{2000}{9}\left(2(n+d)\log\frac{100}{3} + (n+d)\log\frac{\bar{x}^2(1+\|K\|^2)+\sigma_L^2}{\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}}\right) \tag{57}$$
$$\geq \frac{2000}{9}\left(2(n+d)\log\frac{1}{\delta} + (n+d)\log\frac{\bar{x}^2(1+\|K\|^2)+\sigma_L^2}{\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}}\right) \tag{58}$$

with $\delta \leq 3/100$, we have

$$\begin{aligned}
\mathbb{P}\left[\Upsilon_T \not\succeq \Upsilon_0\right] &\leq 2\exp(-\frac{9}{2000((n+d)+1)}|\mathcal{I}|) \\
&\leq 2\exp\left(-\frac{9}{2000((n+d)+1)}\frac{2000}{9}\left(2(n+d)\log\frac{1}{\delta}\right)\right) \\
&\leq 2\exp\left(-\frac{9}{2000((n+d)+1)}\frac{2000}{9}\left((n+d+1)\log\frac{1}{\delta}\right)\right) \\
&= 2\delta.
\end{aligned}$$

Then it holds with probability at least $1-2\delta-\delta = 1-3\delta$ that

$$\kappa \leq \frac{\bar{z}^2|\mathcal{I}|}{\frac{9|\mathcal{I}|}{1600}\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}} = \frac{1600\bar{z}^2}{9\sigma_{\mathcal{I}}^2 \min\left\{\frac{1}{2}, \frac{\psi^2}{\sigma_L^2+2\|K\|^2}\right\}}. \tag{59}$$

For a warm-up block $\mathcal{B}_{i,0}$, note that the exploration noise is fixed at $v_0^2 = 1$, and we have $\left\|K_t^{\text{stab}}\right\| \leq K_u$. Plugging these parameters into (57) and (59) yields the corresponding results. □

C.6 Supporting Lemmas

Lemma C.9 (Lemma E.4 in [35]). *Suppose $z_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(\bar{z}_t, \Sigma_t)$, where $\bar{z}_t \in \mathbb{R}^{\widetilde{d}}$ and $\Sigma_t \in \mathbb{R}^{\widetilde{d}\times\widetilde{d}}$ are $\mathcal{F}_{t-1}$-measurable, and $\Sigma_t \succeq \Sigma \succ 0$. Suppose $\mathcal{E}$ is an arbitrary event and suppose $\mathrm{Tr}\left(\mathbf{E}[V_I \mathbb{1}\{\mathcal{E}\}]\right) \leq \Lambda T$ for some constant $\Lambda \geq 0$. Then for*

$$T \geq \frac{2000}{9}(2\widetilde{d}\log(\frac{100}{3}) + \widetilde{d}\log\frac{\Lambda}{\lambda_{\min}(\Sigma)}),$$

let $V_0 \coloneqq \frac{9T}{1600}\Sigma$, it holds that

$$\mathbf{Pr}[\{V_T \not\succeq V_0\} \cap \mathcal{E}] \leq 2\exp\left(-\frac{9}{2000(\widetilde{d}+1)}T\right).$$

Lemma C.10 (Self-Normalized Tail Bound [1]). *Let $\{\eta_t\}_{t\geq 1}$ be a $\mathcal{F}_t$-adapted sequence such that $\eta_t \mid \mathcal{F}_{t-1}$ is σ^2-sub-Gaussian. Define $V_T := \sum_{t=1}^T z_t z_t^\top$. Fix $V_0 \succ 0$, it holds with probability $1-\delta$ that*

$$\left\|\sum_{t=1}^T \mathbf{x}_t \eta_t\right\|^2_{(V_0+V_T)^{-1}} \leq 2\sigma^2 \log\left\{\frac{1}{\delta}\det\left(V_0^{-1/2}\left(V_0+V_T\right)V_0^{-1/2}\right)\right\}.$$

D PROOF OF PROPOSITION 7.2

Consider a non-stationary Markov decision process on state space $\mathcal{S}$ and action space $\mathcal{A}$, with a time-invariant cost function $c(\cdot,\cdot): \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, and time-dependent transition kernel parametrized by $\{\Theta_t\}_{t\in[T]}$. Let J_t^* denote the optimal (minimum) average cost of the MDP corresponding to Θ_t, and let $h_t(\cdot): \mathcal{S}\to\mathbb{R}$ denote the relative value (bias) function. Then for an arbitrary state $s_t\in\mathcal{S}$ and action $a_t\in\mathcal{A}$, we have the inequality:

$$c(s_t,a_t) \geq J_t^* + h_t(s_t) - \mathbf{E}_{\Theta_t}\left[h(s_{t+1}) \mid s_t, a_t\right], \tag{60}$$

where $\mathbf{E}_{\Theta_t}[X]$ denotes the expectation of random variable X under transition kernel parametrized by Θ_t. Summing the above inequality from $t=1$ to T:

$$\begin{aligned}
\sum_{t=1}^T c(s_t,a_t) &\geq \sum_{t=1}^T J_t^* + h_t(s_t) - \mathbf{E}_{\Theta_t}\left[h(s_{t+1}) \mid s_t,a_t\right] \\
&= \sum_{t=1}^T J_t^* + h_1(s_1) - \mathbf{E}_{\Theta_T}\left[h_T(s_{T+1}) \mid s_T, a_T\right] \\
&\quad + \sum_{t=1}^{T-1} h_{t+1}(s_{t+1}) - \mathbf{E}_{\Theta_t}\left[h_t(s_{t+1}) \mid s_t,a_t\right] \\
&= \sum_{t=1}^T J_t^* + h_1(s_1) - \mathbf{E}_{\Theta_T}\left[h_T(s_{T+1}) \mid s_T, a_T\right] \\
&\quad + \sum_{t=1}^{T-1} h_{t+1}(s_{t+1}) - \mathbf{E}_{\Theta_t}\left[h_{t+1}(s_{t+1}) \mid s_t,a_t\right] \\
&\quad + \sum_{t=1}^{T-1} \mathbf{E}_{\Theta_t}\left[h_{t+1}(s_{t+1}) - h_t(s_{t+1}) \mid s_t,a_t\right].
\end{aligned}$$

Taking expectation with respect to the randomization of the policy and the evolution of the non-stationary MDP,

$$\mathbf{E}\left[\sum_{t=1}^T c(s_t,a_t)\right] - \sum_{t=1}^T J_t^* \geq h_1(s_1) - \mathbf{E}[h_T(s_{T+1})] + \sum_{t=1}^{T-1}\mathbf{E}[h_{t+1}(s_{t+1}) - h_t(s_{t+1})]$$

or

$$\sum_{t=1}^T J_t^* - \mathbf{E}\left[\sum_{t=1}^T c(s_t,a_t)\right] \leq -h_1(s_1) + \mathbf{E}[h_T(s_{T+1})] + \sum_{t=1}^{T-1}\mathbf{E}[h_t(s_{t+1}) - h_{t+1}(s_{t+1})].$$

Specializing to the non-stationary LQR setting, $s_t \equiv x_t$, $a_t \equiv u_t$, $c(x_t,u_t) = x_t^\top Q x_t + u_t^\top R u_t$, $h_t(s_t) \equiv x_t^\top P_t^* x_t$:

$$\sum_{t=1}^T J_t^* - \mathbf{E}\left[\sum_{t=1}^T c(x_t,u_t)\right] \leq \mathbf{E}\left[x_{T+1}^\top P_T^* x_{T+1}\right] + \sum_{t=1}^{T-1}\mathbf{E}\left[x_{t+1}^\top\left(P_t^* - P_{t+1}^*\right)x_{t+1}\right]$$

$$\leq \mathbf{E}\left[\|x_{T+1}\|^2\right] \left\|P_T^*\right\| + \sum_{t=1}^{T-1} \mathbf{E}\left[\|x_{t+1}\|^2\right] \left\|P_t^* - P_{t+1}^*\right\|.$$

In Lemma D.1, we prove that under the optimal dynamic policy, $\mathbf{E}\left[\|x_t\|^2\right]$ is bounded from above by a constant depending only on the cost parameters and the sequential stability parameters κ, γ. The perturbation result for the solution of Discrete Algebraic Riccati equation gives $\left\|P_t^* - P_{t+1}^*\right\| \leq \min\{2C_4 \left\|\Theta_t - \Theta_{t+1}\right\|^2, 2P_u\} = O(\Delta_{t+1}^2)$ [35, Theorem 5]. Lemma D.1 does not bound $\mathbf{E}\left[\|x_{T+1}\|^2\right]$, however, following a similar argument as in the Lemma, we can create another policy that has logarithmic regret compared to the optimal policy and has bounded $\mathbf{E}\left[\|x_{T+1}\|^2\right]$. Combining these, we get the desired bound on the additional regret of $O(V_T + \log T)$ with respect to the dynamic optimal policy. □

Lemma D.1. *Under the optimal dynamic policy for the non-stationary LQR problem,*

$$q_{\min}\mathbf{E}\left[\|x_t\|^2\right] \leq M_\Gamma \frac{2\kappa^2}{\gamma} M_x + M_P \frac{2\kappa^2}{\gamma} + \left(\kappa^2 M_x + \frac{2\kappa^2 n\psi^2}{\gamma}\right)\left(\frac{2\kappa^2}{\gamma} M_\Gamma + \kappa^2 M_P\right),$$

where $M_x := \left(\frac{M_\Gamma}{q_{\min}}\right) \frac{2\kappa^2 n\psi^2}{\gamma}$, $M_P := \frac{2n\kappa^2}{\gamma} M_\Gamma$, *and* $M_\Gamma := \max_s \left\|Q + (K_s^{stab})^\top R K_s^{stab}\right\| \leq q_{\max} + r_{\max}\kappa^2$.

Proof. We prove this result by contradiction. We first establish some notation for the optimal dynamic policy. A classical fact is that the optimal dynamic policy for non-stationary LQR is also a linear state feedback policy, given via the following dynamic programming recursion:

$$\begin{aligned}
P_{T+1} &= 0, \\
K_t &= -(R + B_t^\top P_{t+1} B_t) B_t^\top P_{t+1} A_t, \\
P_t &= Q + K_t^\top R K_t + (A_t + B_t K_t)^\top P_{t+1}(A_t + B_t K_t), \\
J_t &= \mathrm{Tr}(W \cdot P_{t+1}).
\end{aligned}$$

Let t be some time such that under the optimal dynamic policy, $\mathbf{E}\left[\|x_t\|^2\right]$ is larger than the bound in the Lemma statement. Define

$$\tau = \max\{s \leq t-1 : \mathbf{E}\left[\|x_s\|^2\right] \leq M_x\}$$

as the last time before t when the expected squared norm of the state under the optimal policy is smaller than M_x. Similarly, define

$$\tau' = \min\{s \geq t : \|P_{s+1}\| \leq M_P\}$$

as the first time including or after t when the norm of P_{s+1} is smaller than M_P. We will show that by deviating to a policy where $K_s' = K_s^{\text{stab}}$ for $s \in \{\tau, \ldots, \tau'\}$ gives a policy with a smaller cost. Let $\{x_s'\}$ denote the state process for this new policy, $\{P_s'\}$ the relative value function matrices, and $J_s' := \mathrm{Tr}(W \cdot P_s')$.

By the definition of the new policy, we must have $P_s' = P_s$ for $s \geq \tau' + 1$. Recall the recursion for the relative value function for LQR:

$$x_t^\top Q x_t + u_t^\top R u_t = x_t^\top P_t x_t + J_t - \mathbf{E}\left[x_{t+1}^T P_{t+1} x_{t+1}\right].$$

We will decompose the cost of the optimal policy into contributions due to the four intervals $\{1, \ldots, \tau-1\}$, $\{\tau, \ldots, t\}$, $\{t+1, \ldots, \tau'\}$ and $\{\tau'+1, \ldots, T\}$. Since both policies agree on the first interval, the total cost is the same, and hence we do not consider it henceforth. For the interval $\{\tau, \ldots, t\}$ we lower bound the cost of the optimal policy as:

$$\mathbf{E}\left[\sum_{s=\tau}^{t} x_s^\top Q x_s + u_s^\top R u_s\right] \geq \sum_{s=\tau}^{t} q_{\min}\mathbf{E}\left[\|x_s\|^2\right]. \tag{61}$$

For the interval $\{t+1\ldots,\tau'\}$:

$$\mathbf{E}\left[\sum_{s=t+1}^{\tau'} x_s^\top Q x_s + u_s^\top R u_s\right] = \mathbf{E}\left[x_{t+1}^\top P_{t+1} x_{t+1}\right] + \sum_{s=t+1}^{\tau'} J_s - \mathbf{E}\left[x_{\tau'+1}^\top P_{\tau'+1} x_{\tau'+1}\right]$$
$$\geq \sum_{s=t+1}^{\tau'+1} \mathrm{Tr}(W\cdot P_s) - \mathbf{E}\left[x_{\tau'+1}^\top P_{\tau'+1} x_{\tau'+1}\right],$$

where we have used $\mathbf{E}\left[x_{t+1}^\top P_{t+1} x_{t+1}\right] \geq \mathbf{E}\left[w_{t+1}^\top P_{t+1} w_{t+1}\right] = \mathrm{Tr}(W\cdot P_{t+1})$. Finally, for the last interval,

$$\mathbf{E}\left[\sum_{s=\tau'+1}^{T} x_s^\top Q x_s + u_s^\top R u_s\right] = \mathbf{E}\left[x_{\tau'+1}^\top P_{\tau'+1} x_{\tau'+1}\right] + \sum_{s=\tau'+1}^{T} J_s.$$

It will be convenient to combine the lower bound for the interval $\{t+1,\ldots,T\}$ as:

$$\mathbf{E}\left[\sum_{s=t+1}^{T} x_s^\top Q x_s + u_s^\top R u_s\right] \geq \sum_{s=t+1}^{\tau'} \mathrm{Tr}(W\cdot P_s) + \sum_{s=\tau'}^{T} J_s$$
$$\geq \sum_{t=t+1}^{\tau'} \psi^2 \|P_s\| + \sum_{s=\tau'}^{T} J_s. \tag{62}$$

We now proceed to upper bound the cost during these intervals for the modified policy $\{K_s'\}$. Denote $u_s' = K_s' x_s'$ as the control at time step s under the new policy. We first summarize the results of Lemma D.2, which bounds $\mathbf{E}\left[\left\|x_s'\right\|^2\right]$ and $\left\|P_s'\right\|$ for $s\in\{\tau+1,\ldots,\tau'\}$:

$$\mathbf{E}\left[\|x_s\|^2\right] \leq \kappa^2\left(1-\frac{\gamma}{2}\right)^{2(s-\tau)}\mathbf{E}\left[\|x_\tau\|^2\right] + \frac{2\kappa^2 n\psi^2}{\gamma},$$
$$\left\|P_s'\right\| \leq \kappa^2\left(1-\frac{\gamma}{2}\right)^{2(\tau'-s+1)}\|P_{\tau'+1}\| + \frac{2\kappa^2}{\gamma}M_\Gamma.$$

For the second interval, $\{\tau,\ldots,t\}$, we can upper bound the cost of the new policy by:

$$\mathbf{E}\left[\sum_{s=\tau}^{t} {x'}_s^\top Q x_s' + {u'}_s^\top R u_s'\right] = \mathbf{E}\left[\sum_{s=\tau}^{t} {x'}_s^\top (Q + {K'}_s^\top R K_s') x_s'\right]$$
$$\leq M_\Gamma \sum_{s=\tau}^{t}\mathbf{E}\left[\left\|x_s'\right\|^2\right]$$
$$\leq M_\Gamma\left((t-\tau)\frac{2\kappa^2 n\psi^2}{\gamma} + \frac{2\kappa^2}{\gamma}\mathbf{E}\left[\|x_\tau\|^2\right]\right), \tag{63}$$

where we have used the bound on $\mathbf{E}\left[\left\|x_s'\right\|^2\right]$ from Lemma D.2. Next, we upper bound the cost for interval $\{t+1,\ldots,T\}$:

$$\mathbf{E}\left[\sum_{s=t+1}^{\tau'} {x'}_s^\top Q x_s' + {u'}_s^\top R u_s'\right] = \mathbf{E}\left[{x'_{t+1}}^\top P'_{t+1} x'_{t+1}\right] + \sum_{s=t+1}^{T} J_s'$$
$$= \mathbf{E}\left[{x'_{t+1}}^\top P'_{t+1} x'_{t+1}\right] + \sum_{s=t+2}^{\tau'} \mathrm{Tr}(W\cdot P_s') + \sum_{s=\tau'}^{T} J_s'$$
$$\leq \mathbf{E}\left[\left\|x'_{t+1}\right\|^2\right]\left\|P'_{t+1}\right\| + n\psi^2\sum_{s=t+2}^{\tau'}\left\|P_s'\right\| + \sum_{s=\tau'}^{T} J_s.$$

Using Lemma D.2 we can bound:

$$\sum_{s=t+2}^{\tau'} \left\|P'_s\right\| \le \sum_{s=t+2}^{\tau'} \left(\kappa^2 \left(1-\frac{\gamma}{2}\right)^{2(\tau'-s+1)} \left\|P_{\tau'+1}\right\| + \frac{2\kappa^2}{\gamma} M_\Gamma\right) \le (\tau'-t-1)\frac{2\kappa^2}{\gamma} M_\Gamma + \frac{2\kappa^2}{\gamma}\left\|P_{\tau'+1}\right\|.$$

Also, using Lemma D.2:

$$\mathbf{E}\left[\left\|x'_{t+1}\right\|^2\right]\left\|P'_{t+1}\right\| \le \left(\kappa^2 \mathbf{E}\left[\|x_\tau\|^2\right] + \frac{2\kappa^2 n\psi^2}{\gamma}\right)\left(\kappa^2\left\|P_{\tau'+1}\right\| + \frac{2\kappa^2}{\gamma} M_\Gamma\right).$$

Putting together,

$$\begin{aligned}\mathbf{E}\left[\sum_{s=t+1}^{\tau'} x'^{\top}_s Q x'_s + u'^{\top}_s R u'_s\right] \le{}& \left(\kappa^2 \mathbf{E}\left[\|x_\tau\|^2\right] + \frac{2\kappa^2 n\psi^2}{\gamma}\right)\left(\kappa^2\left\|P_{\tau'+1}\right\| + \frac{2\kappa^2}{\gamma} M_\Gamma\right) \\ &+ n\psi^2\left((\tau'-t)\frac{2\kappa^2}{\gamma} M_\Gamma + \frac{2\kappa^2}{\gamma}\left\|P_{\tau'+1}\right\|\right) + \sum_{s=\tau'+1}^{T} J_s. \end{aligned} \tag{64}$$

Combining the lower bounds on the optimal policy cost from (61)-(62) and upper bound on the cost of the modified policy from (63)-(64):

$$\begin{aligned}&\mathbf{E}\left[\sum_{s=\tau}^{T} x_s^\top Q x_s + u_s^\top R u_s\right] - \mathbf{E}\left[\sum_{s=\tau}^{T} x'^{\top}_s Q x'_s + u'^{\top}_s R u'_s\right] \\ \ge{}& \sum_{s=\tau}^{t-1}\left(q_{\min}\mathbf{E}\left[\|x_t\|^2\right] - M_\Gamma\frac{2\kappa^2 n\psi^2}{\gamma}\right) \\ &+ \psi^2\sum_{s=t+1}^{\tau'}\left(\|P_s\| - \frac{2\kappa^2 n}{\gamma} M_\Gamma\right) \\ &+ q_{\min}\mathbf{E}\left[\|x_t\|^2\right] - \frac{2\kappa^2}{\gamma}\left\|P_{\tau'+1}\right\| - M_\Gamma\frac{2\kappa^2}{\gamma}\mathbf{E}\left[\|x_\tau\|^2\right] \\ &- \left(\kappa^2\mathbf{E}\left[\|x_\tau\|^2\right] + \frac{2\kappa^2 n\psi^2}{\gamma}\right)\left(\kappa^2\left\|P_{\tau'+1}\right\| + \frac{2\kappa^2}{\gamma} M_\Gamma\right).\end{aligned}$$

Recall that we choose $M_x := \left(\frac{M_\Gamma}{q_{\min}}\right)\frac{2\kappa^2 n\psi^2}{\gamma}$ as the threshold for $\mathbf{E}\left[\|x_\tau\|^2\right]$, and $M_P := \frac{2n\kappa^2}{\gamma} M_\Gamma$ as the threshold of $\|P_{\tau'+1}\|$. This ensures the first two terms above are non-negative. Therefore, if

$$\begin{aligned}q_{\min}\mathbf{E}\left[\|x_t\|^2\right] &\ge \frac{2\kappa^2}{\gamma}\left\|P_{\tau'+1}\right\| + M_\Gamma\frac{2\kappa^2}{\gamma}\mathbf{E}\left[\|x_\tau\|^2\right] + \left(\kappa^2\mathbf{E}\left[\|x_\tau\|^2\right] + \frac{2\kappa^2 n\psi^2}{\gamma}\right)\left(\kappa^2\left\|P_{\tau'+1}\right\| + \frac{2\kappa^2}{\gamma} M_\Gamma\right) \\ &\ge \frac{2\kappa^2}{\gamma} M_P + M_\Gamma\frac{2\kappa^2}{\gamma} M_x + \left(\kappa^2 M_x + \frac{2\kappa^2 n\psi^2}{\gamma}\right)\left(\kappa^2 M_P + \frac{2\kappa^2}{\gamma} M_\Gamma\right),\end{aligned}$$

we get that the cost of the optimal policy is larger than the modified policy, a contradiction. □

Lemma D.2. *For the alternate policy which chooses $K'_s = K_s^{stab}$ for $s \in \{\tau, \ldots, t, \ldots, \tau'\}$, we have:*

- $\mathbf{E}\left[\|x_s\|^2\right] \le \kappa^2\left(1-\frac{\gamma}{2}\right)^{2(s-\tau)}\mathbf{E}\left[\|x_\tau\|^2\right] + \frac{2\kappa^2 n\psi^2}{\gamma}$,
- $\left\|P'_s\right\| \le \kappa^2\left(1-\frac{\gamma}{2}\right)^{2(\tau'-s+1)}\left\|P_{\tau'+1}\right\| + \frac{2\kappa^2}{\gamma} M_\Gamma$,

where $M_\Gamma := \max_s \left\|Q + (K_s^{stab})^\top R K_s^{stab}\right\|$, and $\{x_s\}, \{P_s\}$ denote the optimal policy quantities.

Proof. Recall our notation $\Phi_t := A_t + B_t K_t^{\text{stab}}$, $\Gamma_t = Q + (K_t^{\text{stab}})^\top R K_t^{\text{stab}}$. Denote for $a \le b$:

$$\Phi_{b:a} = \Phi_b \Phi_{b-1} \cdots \Phi_a.$$

We can write:

$$\begin{aligned} x_s &= \Phi_{s-1} x_{s-1} + w_{s-1} \\ &= \Phi_{s-1}\Phi_{s-2} x_{s-2} + \Phi_{s-1} w_{s-2} + w_{s-1} \\ &= \Phi_{s-1:\tau} x_\tau + \sum_{\ell=\tau}^{s-1} \Phi_{s-1:\ell+1} w_\ell, \end{aligned}$$

which gives:

$$\mathbf{E}\left[\|x_s\|^2\right] \le \|\Phi_{s-1:\tau}\|^2 \, \mathbf{E}\left[\|x_\tau\|^2\right] + \sum_{\ell=\tau}^{s-1} \|\Phi_{s-1:\ell+1}\|^2 \, \mathbf{E}\left[\|w_\ell\|^2\right]. \tag{65}$$

By sequential strong stability:

$$\begin{aligned} \|\Phi_{s-1:\ell}\| &= \left\| H_{s-1} L_{s-1} H_{s-1}^{-1} H_{s-2} L_{s-2} H_{s-2}^{-1} \cdots H_\ell L_\ell H_\ell^{-} 1 \right\| \\ &\le \|H_{s-1}\| \, \|L_{s-1}\| \left\| H_{s-1}^{-1} H_{s-2} \right\| \|L_{s-2}\| \cdots \left\| H_{\ell+1}^{-1} H_\ell \right\| \|L_\ell\| \left\| H_\ell^{-} 1 \right\| \\ &\le \kappa (1-\gamma)^{s-\ell} (1+\gamma/2)^{s-\ell-1}, \end{aligned}$$

which using $(1-\gamma)(1+\gamma/2) \le (1-\gamma/2)$ gives

$$\le \kappa (1-\gamma/2)^{s-\ell}.$$

Substituting the above in (65) and using $\mathbf{E}\left[\|w_s\|^2\right] = n\psi^2$:

$$\begin{aligned} \mathbf{E}\left[\|x_s\|^2\right] &\le \kappa^2 (1-\gamma/2)^{2(s-\tau)} \mathbf{E}\left[\|x_\tau\|^2\right] + n\psi^2\kappa^2 \left(1 + (1-\gamma/2)^2 + \cdots + (1-\gamma/2)^{2(s-\tau-1)}\right) \\ &\le \kappa^2 (1-\gamma/2)^{2(s-\tau)} \mathbf{E}\left[\|x_\tau\|^2\right] + n\psi^2\kappa^2 \frac{2}{\gamma}. \end{aligned}$$

This proves the first part. For the second part,

$$\begin{aligned} P'_s &= Q + (K_s^{\text{stab}})^\top R K_s^{\text{stab}} + \Phi_s^\top P'_{s+1} \Phi_s \\ &= \Gamma_s + \Phi_s^\top P'_{s+1} \Phi_s \\ &= \Gamma_s + \Phi_s^\top \Gamma_{s+1} \Phi_s + \Phi_s^\top \Phi_{s+1}^\top P'_{s+2} \Phi_{s+1} \Phi_s \\ &= \Phi_{\tau':s}^\top P_{\tau'+1} \Phi_{\tau':s} + \sum_{\ell=s}^{\tau'} \Phi_{\ell-1:s}^\top \Gamma_\ell \Phi_{\ell-1:s}. \end{aligned}$$

Therefore,

$$\begin{aligned} \left\|P'_s\right\| &\le \left\| \prod_{m=s}^{\tau'} \Phi_{\tau':s} \right\|^2 \|P_{\tau'+1}\| + \sum_{\ell=s}^{\tau'} \left\| \prod_{m=s}^{\ell-1} \Phi_{\ell-1:s} \right\|^2 \|\Gamma_\ell\| \\ &\le \kappa^2 (1-\gamma/2)^{2(\tau'+1-s)} \|P_{\tau'+1}\| + M_\Gamma \kappa^2 \left(1 + (1-\gamma/2)^2 + \cdots + (1-\gamma/2)^{2(\tau'-s)}\right) \\ &\le \kappa^2 (1-\gamma/2)^{2(\tau'+1-s)} \|P_{\tau'+1}\| + M_\Gamma \frac{2\kappa^2}{\gamma}. \end{aligned}$$

□

E PROOF OF THEOREM 7.1

E.1 Proofs for Section 7.2

Proof of Lemma 7.4. Consider a stabilization epoch starting at time τ^{stab} and ending at θ^{stab}. During the epoch, the dynamics of the state evolution is:

$$x_s = \Phi_{s-1} x_{s-1} + w_{s-1},$$

where $\Phi_s := A_s + B_s K_s^{\text{stab}}$. Denote for $a \le b$:

$$\Phi_{b:a} = \Phi_b \Phi_{b-1} \cdots \Phi_a.$$

Then,

$$\begin{aligned}
\|x_s\| &\le \left\|\Phi_{s-1:\tau^{\text{stab}}}\right\| \left\|x_{\tau^{\text{stab}}}\right\| + \left\|\Phi_{s-1:\tau^{\text{stab}}}\right\| \left\|w_{\tau^{\text{stab}}}\right\| + \left\|\Phi_{s-1:\tau^{\text{stab}}+1}\right\| \left\|w_{\tau^{\text{stab}}+1}\right\| + \cdots + \|\Phi_{s-1}\| \|w_{s-1}\| \\
&\le \kappa(1-\gamma/2)^{s-\tau^{\text{stab}}} \left\|x_{\tau^{\text{stab}}}\right\| + \kappa(1-\gamma/2)^{s-\tau^{\text{stab}}} \left\|w_{\tau^{\text{stab}}}\right\| + \kappa(1-\gamma/2)^{s-\tau^{\text{stab}}-1} \left\|w_{\tau^{\text{stab}}+1}\right\| \\
&\qquad + \cdots + \kappa(1-\gamma/2) \|w_{s-1}\| \\
&=: \kappa Y_{s-\tau^{\text{stab}}},
\end{aligned}$$

so that $\left\|x_{t+\tau^{\text{stab}}}\right\| \le \kappa Y_t$ with $Y_0 = \left\|x_{\tau^{\text{stab}}}\right\|$, $W_t = \left\|w_{t+\tau^{\text{stab}}}\right\|$ and

$$Y_{t+1} \le (1-\gamma/2) Y_t + W_t.$$

Now applying Lemma A.3 with $\rho = \rho_0 = (1-\gamma/2)$, and substituting $a = \frac{2\psi\sqrt{n}}{1-\rho_0}$, we get

$$\rho_0 + (1-\rho_0)\frac{\mathbf{E}[\|w_t\|]}{2\psi\sqrt{n}} \le \rho_0 + (1-\rho_0)\frac{\psi\sqrt{n}}{2\psi\sqrt{n}} = \frac{1+\rho_0}{2}.$$

Therefore,

$$\sum_{t=\tau^{\text{stab}}}^{\theta^{\text{stab}}} \|x_t\|^2 \le 2\kappa^2 \frac{\left\|x_{\tau^{\text{stab}}}\right\|^2}{1-\rho_0}.$$

Since

$$\mathbf{E}[c_t \mid \mathcal{F}_{t-1}, \mathcal{G}_{t-1}] = x_t^\top Q x_t + [(A_t + B_t K_t) x_t]^\top R [(A_t + B_t K_t) x_t] + \sigma_t^2 \operatorname{Tr}(R) = O(\|x_t\|^2 + 1),$$

the total cost during the stabilization epoch is bounded by $O(\left\|x_{\tau^{\text{stab}}}\right\|^2)$. We next show that this is $O(\ln T)$. Note that we have $\left\|x_{\tau^{\text{stab}}-1}\right\| \le x_u = 2\kappa e^{C_{ss}} \left(\frac{\sqrt{8(n+d)}\beta}{\sqrt{1-\rho_0}} \sqrt{\log T} + \frac{(n+d)B}{1-\rho_0} \right)$. Therefore,

$$\mathbf{E}\left[\left\|x_{\tau^{\text{stab}}}\right\|^2\right] \le \mathbf{E}\left[\sup_{\substack{y_1, \ldots y_{T-1} \in \mathbb{R}^n: \\ \|y_t\| \le x_u \forall t \in [T-1]}} \max_t \|(A_t + B_t K_t) y_t + B_t \sigma_t \eta_t + w_t\|^2 \right].$$

Since $\max_t \max\{\|A_t\|, \|B_t\|, \|K_t\|\}$ are bounded, for some problem dependent constant C_{13}

$$\begin{aligned}
\mathbf{E}\left[\left\|x_{\tau^{\text{stab}}}\right\|^2\right] \le C_{13} \Big(&\mathbf{E}\left[\max_t \|w_t\|^2 + \max_t \|\eta_t\|^2\right] + x_u \mathbf{E}\left[\max_t \|w_t\| + \max_t \|\eta_t\|\right] \\
&+ \mathbf{E}\left[\max_t \|w_t\|\right] \cdot \mathbf{E}\left[\max_t \|\eta_t\|\right] \Big).
\end{aligned}$$

Using Lemma A.2, the expression above is $O(n + d + \ln T)$.

Proof of Lemma 7.5. Since $\mathcal{E}_i$ is an exploration epoch, we have $\left\|x_{\tau_i}\right\| \leq x_u$. During $\mathcal{B}_{i,0}$ the dynamics are given by:

$$x_{t+1} = \Phi_t x_t + \nu_0 \Xi_t \eta_t + w_t,$$

where $\Phi_t := A_t + B_t K_t^{\text{stab}}$ and $\Xi_t := B_t K_t^{\text{stab}}$. Denote for $a \leq b$:

$$\Phi_{b:a} = \Phi_b \Phi_{b-1} \cdots \Phi_a.$$

By our assumption that η_t and w_t are a sequence of independent mean 0 Gaussian random vectors:

$$\begin{aligned}
\mathbf{E}\left[\|x_t\|^2 \mid \mathcal{F}_{\tau_i-1}, \mathcal{G}_{\tau_i-1}\right] &= \left\|\Phi_{t:\tau_i} x_{\tau_i}\right\|^2 + \sum_{s=\tau_i}^{t-1} \nu_0^2 \mathbf{E}\left[\|\Phi_{t:s+1}\Xi_s \eta_s\|^2 \mid \mathcal{F}_{\tau_i-1}, \mathcal{G}_{\tau_i-1}\right] \\
&\quad + \sum_{s=\tau_i}^{t-1} \mathbf{E}\left[\|\Phi_{t:s+1} w_s\|^2 \mid \mathcal{F}_{\tau_i-1}, \mathcal{G}_{\tau_i-1}\right] \\
&\leq \kappa^2 \rho_0^{2(t-\tau_i)} \left\|x_{\tau_i}\right\|^2 + \kappa^2 n \frac{\nu_0^2 C_6 + \psi^2}{1-\rho_0^2},
\end{aligned}$$

where $C_6 := \max_{t\in[T]} \left\|B_t K_t^{\text{stab}}\right\|^2$. This further gives the total during $\mathcal{B}_{i,0}$

$$as\mathbf{E}\left[\sum_{t=\tau_i}^{\tau_i+L-1} \|x_t\|^2 \mid \mathcal{F}_{\tau_i-1}, \mathcal{G}_{\tau_i-1}\right] \leq \kappa^2 \frac{x_u^2}{1-\rho_0^2} + L\kappa^2 n \cdot \frac{\nu_0^2 C_6 + \psi^2}{1-\rho_0^2}.$$

Finally, again using the fact that $\mathbf{E}\left[c_t \mid \mathcal{F}_{\tau_i-1}\right] = \widetilde{O}(\mathbf{E}\left[\|x_t\|^2 \mid \mathcal{F}_{\tau_i-1}\right] + 1)$, and the definition of $L := \frac{16(n+d)\log^3 T}{1-\rho_0}$, the bound in the lemma statement follows.

Proof of Lemma 7.6. We first prove a lemma that gives a regret decomposition for good intervals.

Lemma E.1. *For some epoch $\mathcal{E}_i$, a block $\mathcal{B}_{i,j}$ in epoch $\mathcal{E}_i$, and a good interval $\mathcal{I}_{i,j,k}^{good} = [\tau, \theta]$ in block $\mathcal{B}_{i,j}$, the following identity holds:*

$$\begin{aligned}
\mathcal{R}^{\pi}(\mathcal{I}_{i,j,k}^{good}) &:= \sum_{t\in\mathcal{I}_{i,j,k}^{good}} x_t^\top Q x_t + u_t^\top R u_t - J_t^* \\
&= \sum_{t=\tau}^{\theta} J_t(K_t) - J_t^* \\
&\quad + x_\tau^\top P_\tau(K_\tau) x_\tau - x_{\theta+1}^\top P_\theta(K_\theta) x_{\theta+1} \\
&\quad + \sum_{t=\tau}^{\theta-1} x_{t+1}^\top \left(P_{t+1}(K_{t+1}) - P_t(K_t)\right) x_{t+1} \\
&\quad + \sum_{t=\tau}^{\theta} \left(x_{t+1}^\top P_t(K_t) x_{t+1} - \mathbf{E}\left[x_{t+1}^\top P_t(K_t) x_{t+1} \mid x_t, \sigma_t\right]\right) \\
&\quad + \sum_{t=\tau}^{\theta} \sigma_t^2 \operatorname{Tr}\left(R + B_t^\top P_t(K_t) B_t\right).
\end{aligned}$$

Proof. Note that for an interval lying within block $\mathcal{B}_{i,j}$, the policy $K_t = K^*(\widehat{\Theta}_{i,j-1})$ is fixed, however for generality, we use K_t. For dynamics given by Θ_t, and control policy $u_t = K_t x_t + \sigma_t \eta_t$

with $\eta_t \sim N(0, I_n)$, we have the following Bellman recursion:

$$x_t^\top P_t(K_t) x_t = x_t^\top Q x_t + u_t^\top R u_t - J_t(K_t) - \sigma_t^2 \operatorname{Tr}\left(R + B_t^\top P_t(K_t) B_t\right) + \mathbf{E}\left[x_{t+1}^\top P_t(K_t) x_{t+1} \mid x_t, \sigma_t\right].$$

Rearranging terms, we get,

$$\begin{aligned}
&x_t^\top Q x_t + u_t^\top R u_t - J_t^* \\
&\quad = J_t(K_t) - J_t^* + x_t^\top P_t(K_t) x_t - \mathbf{E}\left[x_{t+1}^\top P_t(K_t) x_{t+1} \mid x_t, \sigma_t\right] + \sigma_t^2 \operatorname{Tr}\left(R + B_t^\top P_t(K_t) B_t\right) \\
&\quad = J_t(K_t) - J_t^* + \left(x_t^\top P_t(K_t) x_t - x_{t+1}^\top P_{t+1}(K_{t+1}) x_{t+1}\right) \\
&\qquad + \left(x_{t+1}^\top P_{t+1}(K_{t+1}) x_{t+1} - x_{t+1}^\top P_t(K_t) x_{t+1}\right) \\
&\qquad + \left(x_{t+1}^\top P_t(K_t) x_{t+1} - \mathbf{E}\left[x_{t+1}^\top P_t(K_t) x_{t+1} \mid x_t, \sigma_t\right]\right) + \sigma_t^2 \operatorname{Tr}\left(R + B_t^\top P_t(K_t) B_t\right)
\end{aligned}$$

Summing the above for the entire interval gives the identity in the lemma. □

With Lemma E.1, after taking expectation and using Lemma 3.1, we prove Lemma 7.6 below.

Proof. The second expression follows from the first by noting the definition of good intervals: for all $t \in \mathcal{I}_{i,j,k}^{\text{good}}$, $\left\|\widehat{\Theta}_{i,j-1} - \Theta_t\right\| \le C_3$ and applying Lemma 3.1.

To arrive at the first expression, we go through the expression in Lemma E.1 line-by-line. The expression in the second line is bounded by $\|x_\tau\|^2 \|P_\tau(K_\tau)\|$, which is $\widetilde{O}\left(\frac{n+d+\log T}{1-\rho_0}\right)$. For the expression in the third line, noting that $K_t = K_{t+1}$, and that $\|P_t(K) - P_{t+1}(K)\| \le C_{12} \|\Theta_t - \Theta_{t+1}\|$ for a stabilizing controller K, and a constant C_{12}, the sum is bounded by $C_{12} \sum_{t=\tau}^{\theta} \|x_t\|^2 \|\Theta_t - \Theta_{t+1}\|$, which is $\widetilde{O}\left(\frac{n+d+\log T}{1-\rho_0} \Delta_{i,j,k}\right)$. The expression in the fourth line is a mean 0 random variable and hence vanishes when we take the expectation. For the expression in the last line, note that in block $\mathcal{B}_{i,j}$, for each $m = 0, 1, \ldots, j-1$ we start an exploration phase of scale m (duration $L \cdot 2^m$) with probability $\frac{1}{L\sqrt{2^j}\sqrt{2^m}}$ at each time t, and during an exploration of phase m, we choose $\sigma_t^2 = \sqrt{\frac{C_0}{L 2^m}}$. We will upper bound $\mathbf{E}\left[\sum_{t \in \mathcal{I}_{i,j,k}^{\text{good}}} \sigma_t^2\right]$ by allocating the entire exploration variance to the time t at which an exploration phase begins. For a given time t, this gives the expected contribution due to scale m as $\frac{1}{L\sqrt{2^j}\sqrt{2^m}} \times \sqrt{C_0 2^m / L} = \frac{C_0^{1/2}}{L^{3/2}\sqrt{2^j}}$. Summing over m gives $\frac{j C_0^{1/2}}{L^{3/2}\sqrt{2^j}}$, and multiplying by $\left|\mathcal{I}_{i,j,k}^{\text{good}}\right|$ finally gives the expression in the Lemma. □

E.2 Proofs for Section 7.3

Proof of Lemma 7.7. Consider the block $\mathcal{B}_{i,j} = [s_{i,j}, s_{i,j} + 2^j L - 1]$ in $\mathcal{E}_i$. We first show that no restart is triggered by EndOfBlockTest(i, j). Let $t = \tau_i + 2^j L - 1$, then $\Delta_{\mathcal{B}_{i,j}} \le \Delta_{[\tau_i, t]} \le (t - \tau_i + 1)^{-1/4} \le |\mathcal{B}_{i,j}|^{-1/4}$. Conditioning on Event $\mathcal{E}$, we have

$$\left\|\Theta_{s_{i,j}} - \widehat{\Theta}_{i,j}\right\|_F \le \bar{C}_{\text{bias}} \Delta_{\mathcal{B}_{i,j}} + \bar{C}_{\text{var}} |\mathcal{B}_{i,j}|^{-1/4}.$$

Similarly, we also have

$$\left\|\Theta_{s_{i,j-1}} - \widehat{\Theta}_{i,j-1}\right\|_F \le \bar{C}_{\text{bias}} \Delta_{\mathcal{B}_{i,j-1}} + \bar{C}_{\text{var}} |\mathcal{B}_{i,j-1}|^{-1/4}$$

and

$$\begin{aligned}
\left\|\Theta_{s_{i,j-1}} - \widehat{\Theta}_{[\tau_i, t]}\right\|_F &\le \Delta_{[\tau_i, t]} + C_1 |t - \tau_i + 1|^{-1/4}, \\
\left\|\Theta_{s_{i,j}} - \widehat{\Theta}_{[\tau_i, t]}\right\|_F &\le \Delta_{[\tau_i, t]} + C_1 |t - \tau_i + 1|^{-1/4}.
\end{aligned}$$

Then

$$\begin{aligned}\left\|\widehat{\Theta}_{i,j}-\widehat{\Theta}_{i,j-1}\right\|_F &\le \left\|\Theta_{s_{i,j}}-\widehat{\Theta}_{i,j}\right\|_F + \left\|\Theta_{s_{i,j}}-\Theta_{s_{i,j-1}}\right\|_F + \left\|\Theta_{s_{i,j-1}}-\widehat{\Theta}_{i,j-1}\right\|_F \\ &\le \bar{C}_{\text{bias}}\Delta_{\mathcal{B}_{i,j}} + \bar{C}_{\text{var}}|\mathcal{B}_{i,j}|^{-1/4} + \Delta_{\tau_i,t} + \bar{C}_{\text{bias}}\Delta_{\mathcal{B}_{i,j-1}} + \bar{C}_{\text{var}}|\mathcal{B}_{i,j-1}|^{-1/4} \\ &\le (1+\bar{C}_{\text{bias}})\Delta_{\tau_i,t} + 2\bar{C}_{\text{var}}|\mathcal{B}_{i,j-1}|^{-1/4} \\ &\le (1+\bar{C}_{\text{bias}}+2\bar{C}_{\text{var}})|\mathcal{B}_{i,j-1}|^{-1/4}.\end{aligned}$$

As a result, $\left\|\widehat{\Theta}_{i,j}-\widehat{\Theta}_{i,j-1}\right\|_F^2 \le (1+\bar{C}_{\text{bias}}+2\bar{C}_{\text{var}})^2|\mathcal{B}_{i,j-1}|^{-1/2}$ and $\textsc{EndOfBlockTest}(i,j) = \textit{Pass}$. Similarly, for any exploration interval $\mathcal{I} = [s,e] \subset [\tau_i,t]$ with index $m \le j-1$, note that $\Delta_{\mathcal{I}} \le \Delta_{[\tau_i,t]} \le (t-\tau_i+1)^{-1/4} \le |\mathcal{I}|^{-1/4}$. Then

$$\begin{aligned}\left\|\widehat{\Theta}_{i,j,(m,s)}-\widehat{\Theta}_{i,j-1}\right\|_F &\le \left\|\widehat{\Theta}_{i,j,(m,s)}-\Theta_s\right\|_F + \left\|\Theta_s-\Theta_{s_{i,j-1}}\right\|_F + \left\|\Theta_{s_{i,j-1}}-\widehat{\Theta}_{i,j-1}\right\|_F \\ &\le \bar{C}_{\text{bias}}\Delta_{\mathcal{I}} + \bar{C}_{\text{var}}|\mathcal{I}|^{-1/4} + \Delta_{\tau_i,t} + \bar{C}_{\text{bias}}\Delta_{\mathcal{B}_{i,j-1}} + \bar{C}_{\text{var}}|\mathcal{B}_{i,j-1}|^{-1/4} \\ &\le (1+\bar{C}_{\text{bias}})\Delta_{\tau_i,t} + 2\bar{C}_{\text{var}}|\mathcal{I}|^{-1/4} \\ &\le (1+\bar{C}_{\text{bias}}+2\bar{C}_{\text{var}})|\mathcal{I}|^{-1/4}.\end{aligned}$$

Then $\left\|\widehat{\Theta}_{i,j,(m,s)}-\widehat{\Theta}_{i,j-1}\right\|_F^2 \le (1+\bar{C}_{\text{bias}}+2\bar{C}_{\text{var}})^2|\mathcal{I}|^{-1/2}$ and $\textsc{EndOfExplorationTest}(i,j,m,s) = \textit{Pass}$.

Proof of Corollary 7.8. By Lemma 7.7, to end an epoch $\mathcal{E}_i$ due to detection of nonstationarity, we need $\Delta_{[\tau_i,t]} \ge \sqrt{\frac{C_0}{|\mathcal{E}_i|^{1/2}}}$. Then

$$\Delta_T \ge \sum_{i=1}^{E}\Delta_{[\tau_i,t]} \ge \sum_{i=1}^{E}\sqrt{\frac{C_0}{|\mathcal{E}_i|^{1/2}}}$$

or

$$\sum_{i=1}^{E}|\mathcal{E}_i|^{-\frac{1}{4}} \le C_0^{-\frac{1}{2}}\Delta_T.$$

By Hölder's inequality,

$$\begin{aligned}E &\le \left(\sum_{i=1}^{E}|\mathcal{E}_i|^{-\frac{1}{4}}\right)^{\frac{4}{5}}\left(\sum_{i=1}^{E}|\mathcal{E}_i|\right)^{\frac{1}{5}} \\ &\le \left(C_0^{-\frac{1}{2}}\Delta_T\right)^{\frac{4}{5}}(T)^{\frac{1}{5}} \\ &= C_0^{-\frac{2}{5}}\Delta_T^{\frac{4}{5}}T^{\frac{1}{5}}.\end{aligned}$$

Proof of Lemma 7.9. By Assumption 4.3, a control based on an estimate $\widehat{\Theta}_t$ of Θ_t such that $\left\|\Theta_t-\widehat{\Theta}_t\right\|_F \le C_3/2$ guarantees that $K^*(\widehat{\Theta}_t)$ is in fact (κ,γ,ν) sequentially strongly-stable for the epoch $\mathcal{E}_i$ and parameters κ,γ,ν specified in Lemma 4.6. We first show that under the assumption that $\{K_t\}$ is a (κ,γ,ν) sequentially strongly-stable sequence of controllers for the non-stationary dynamics in the interval $[s,e]$ $(1 \le s \le e \le T)$, then Lemma A.4 implies that with high probability $\max_{t\in[s,e]}\|x_t\| \le x_u$.

The LQR dynamics are given by:

$$x_{t+1} = (A_t + B_tK_t)x_t + \sigma_tB_t\eta_t + w_t.$$

Under the independence assumptions on $\{\eta_t\}, \{w_t\}$, we can use the analysis approach in Lemma 7.4 to show that $\|x_t\| \le \kappa e^{C_{ss} V_{[s:t-1]}} Y_t$ where Y_t obeys $Y_{s-1} = \|x_{s-1}\|$ and

$$Y_{t+1} \le (1-\gamma) Y_t + \sum_{i=1}^{d} \beta_{i,t} |\widehat{\eta}_{i,t}| + \psi \sum_{j=1}^{n} |\widehat{w}_{i,t}|.$$

In the above, $\beta_{i,t}$ are the singular values of B_t, and $\widehat{\eta}_{i,t}, \widehat{w}_{i,t}$ are independent $\mathcal{N}(0,1)$ random variables. We have used the fact that $\sigma_t \le 1$ for all t. Denoting:

$$\beta = \max\left\{\psi, \max_{i,t} \beta_{i,t}\right\}$$

and applying Lemma A.4,

$$\mathbf{Pr}\left[\max_{t\in[s,e]} Y_t \ge Y_{s-1} + \left(\frac{\sqrt{8(n+d)}\beta}{\sqrt{1-\rho_0}}\sqrt{\log T} + \frac{(n+d)B}{1-\rho_0}\right)\right] \le \frac{|e-s+1|}{T^4}, \tag{66}$$

or,

$$\mathbf{Pr}\left[\max_{t\in[s,e]} \|x_t\| \, e^{-C_{ss} V_{[s:e-1]}}/\kappa \ge \|x_{s-1}\| + \left(\frac{\sqrt{8(n+d)}\beta}{\sqrt{1-\rho_0}}\sqrt{\log T} + \frac{(n+d)B}{1-\rho_0}\right)\right] \le \frac{|e-s+1|}{T^4}. \tag{67}$$

Note that we start an epoch with $\|x_{\tau_i}\| \le x_u$, and then use stabilizing controls for L time steps. Using (A.4),

$$\mathbf{Pr}\left[\|x_{\tau_i+L}\| \ge \kappa^2\left(\rho_0^L x_u + \left(\frac{\sqrt{6(n+d)}\beta}{\sqrt{1-\rho_0}}\sqrt{\log T} + \frac{(n+d)B}{1-\rho_0}\right)\right)\right] \le \frac{1}{T^3}. \tag{68}$$

With our choice of L, $\rho_0^L x_u = o(1)$.

Lemma 6.2 and Lemma 6.1 prove that under Event 1 the OLS estimate $\widehat{\Theta}_{i,j}$ based on block $j \ge 0$ of epoch $\mathcal{I}_i$ indeed satisfies $\left\|\Theta_t - \widehat{\Theta}_t\right\|_F^2 \le C_3$. Thus it holds that the controllers $\{K_t\}$ are indeed (κ, γ, ν) sequentially strongly-stable for $t \in [\tau_i + L, \theta_i]$. Combining (68) and (67), for epoch $\mathcal{E}_i = [\tau_i, \theta_i]$ we get

$$\mathbf{Pr}\left[\max_{t\in[\tau_i+L,\theta_i]} \|x_t\| \ge 2\kappa e^{C_{ss}}\left(\frac{\sqrt{8(n+d)}\beta}{\sqrt{1-\rho_0}}\sqrt{\log T} + \frac{(n+d)B}{1-\rho_0}\right)\right] \le \frac{2}{T^3}. \tag{69}$$

Therefore, with high probability, a restart of the epoch based on instability detection does not happen.

E.3 Proofs for Section 7.4

Lemma E.2. *Assume Event 1 holds. Let $\mathcal{I} = [s,e]$ be an interval in $\mathcal{B}_{i,j}$ satisfying $\Delta_{\mathcal{I}}^2 \le \alpha_{\mathcal{I}} = \frac{1}{\sqrt{|\mathcal{I}|}}$ and $\varepsilon_{\mathcal{I}} = \left\|\widehat{\Theta}_{i,j-1} - \Theta_s\right\|_F^2 \ge C_5 \cdot \frac{1}{\sqrt{\mathcal{I}}}$, where we define $C_5 := (2 + 2\bar{C}_{bias} + 3\bar{C}_{var})^2$. Define $\widetilde{\varepsilon}_{\mathcal{I}} := \min\{\varepsilon_{\mathcal{I}}, C_3\}$. Then, there exists an index m, such that (1) $C_5 \cdot \frac{1}{\sqrt{2^{m+1}L}} \le \widetilde{\varepsilon}_{\mathcal{I}} \le C_5 \cdot \frac{1}{\sqrt{2^m L}}$, (2) $2^m L \le |\mathcal{I}|$, and (3) if an exploration phase with index m starts at some time $\widetilde{s}$ within the interval $[s, e - 2^m L]$, then the algorithm starts a new epoch at the end of the exploration phase.*

Proof. By our assumption, $\widetilde{\varepsilon}_{\mathcal{I}} \le \frac{C_5}{\sqrt{L}}$. Note that $\mathcal{I} \subset \mathcal{J}$, then $\widetilde{\varepsilon}_{\mathcal{I}} \ge C_5 \cdot \frac{1}{\sqrt{\mathcal{I}}} \ge C_5 \cdot \frac{1}{\sqrt{2^j L}}$. Then there exist a index $m \in [j]$ satisfying *(1). (2)* is implied by $C_5 \cdot \frac{1}{\sqrt{|\mathcal{I}|}} \le \widetilde{\varepsilon}_{\mathcal{I}} \le C_5 \cdot \frac{1}{\sqrt{2^m L}}$.

To prove *(3)*, let $\widetilde{s} \in [s, e - 2^m L]$ be the starting time of $\mathcal{I}$, condition on Event $\mathcal{E}$ and note that $[\widetilde{s}, \widetilde{s} + 2^m L] \subset I$. We have

$$\left\|\widehat{\Theta}_{i,j,(m,\widetilde{s})} - \Theta_{\widetilde{s}}\right\|_F \leq \bar{C}_{\text{bias}} \Delta_{\mathcal{I}} + \bar{C}_{\text{var}} |\mathcal{I}|^{-\frac{1}{2}}.$$

By direct computation,

$$\begin{aligned}
\left\|\widehat{\Theta}_{i,j-1} - \widehat{\Theta}_{i,j,(m,s)}\right\|_F &\geq \left\|\widehat{\Theta}_{i,j-1} - \Theta_{\widetilde{s}}\right\|_F - \|\Theta_{\widetilde{s}} - \Theta_s\|_F - \left\|\Theta_s - \widehat{\Theta}_{i,j,(m,\widetilde{s})}\right\|_F \\
&\geq \sqrt{C_5} \cdot \mathcal{I}^{-\frac{1}{4}} - \Delta_{\mathcal{I}} - \bar{C}_{\text{bias}} \Delta_{\mathcal{I}} - \bar{C}_{\text{var}} |\mathcal{I}|^{-\frac{1}{2}} \\
&\geq \sqrt{C_5} \cdot \mathcal{I}^{-\frac{1}{4}} - (1 + \bar{C}_{\text{bias}}) \mathcal{I}^{-\frac{1}{4}} - C_{\text{var}} |\mathcal{I}|^{-\frac{1}{2}} \\
&\geq (\sqrt{C_5} - 1 - \bar{C}_{\text{bias}} - \bar{C}_{\text{var}}) \mathcal{I}^{-\frac{1}{4}}.
\end{aligned}$$

Hence

$$\left\|\widehat{\Theta}_{i,j-1} - \widehat{\Theta}_{i,j,(m,\widetilde{s})}\right\|_F^2 \geq (\sqrt{C_5} - 1 - \bar{C}_{\text{bias}} - \bar{C}_{\text{var}})^2 \mathcal{I}^{-\frac{1}{2}} \geq (1 + \bar{C}_{\text{bias}} + 2\bar{C}_{\text{var}})^2 \mathcal{I}^{-\frac{1}{2}}$$

and EndOfExplorationTest$(i, j, m, s) = \textit{Fail}$. □

Proof of Lemma 7.11. Starting with the definition in (12),

$$\begin{aligned}
\mathcal{L}(I) &:= \sum_{t \in \mathcal{I}} \min \left\{ C_4 \left\|\widehat{\Theta}_{i,j-1} - \Theta_t\right\|_F^2, C_3 \right\} \\
&\leq C_4 \sum_{t \in \mathcal{I}} \left\|\Theta_t - \widehat{\Theta}_{i,j-1}\right\|_F^2 \\
&\leq 2C_4 |\mathcal{I}| \left\|\widehat{\Theta}_{i,j-1} - \Theta_s\right\|_F^2 + 2C_4 \sum_{t \in \mathcal{I}} \|\Theta_t - \Theta_s\|_F^2 \\
&\leq 2C_4 |\mathcal{I}| \left((\alpha_{\mathcal{I}} + \varepsilon_{\mathcal{I}} \mathbb{1}\{\varepsilon_{\mathcal{I}} \geq \alpha_{\mathcal{I}}\}) + \Delta_{\mathcal{I}}^2 \right).
\end{aligned}$$

Proof of Lemma 7.12. We create the partition using Algorithm 4, where we check the truncating condition *current interval ends and a new interval is created at time* $t \in \mathcal{J}$ *whenever* $\Delta_{[s_k,t]} \leq \sqrt{\frac{\log |J|}{(t-s_k)^{1/2}+1}}$ and $\Delta_{[s_k,t+1]} > \sqrt{\frac{\log |J|}{(t-s_k)^{1/2}+2}}$ at each time $t \in \mathcal{J}$.

Algorithm 4: Creating Partition

Input: *an block* $\mathcal{J} = [s, e]$.;
Initialize: *Set* $k = 1$; $s_1 = s$; $t = s$.;
while $t \leq e$ **do**
  if $\Delta_{[s_k,t]} \leq \sqrt{\frac{\log |J|}{(t-s_k)^{1/2}+1}}$ *and* $\Delta_{[s_k,t+1]} > \sqrt{\frac{\log |J|}{(t-s_k)^{1/2}+2}}$ **then**
    Let $e_k \leftarrow t$; $\mathcal{I}_k \leftarrow [s_k, e_k]$; $k \leftarrow k + 1$.
  end
  $t \leftarrow t + 1$
end
if $s_k \leq e$ **then**
  $e_k \leftarrow e$; $\mathcal{I}_k \leftarrow [s_k, e_k]$.
end

To calculate an upper bound for the number of intervals Γ, consider the inequality

$$\Delta_{[s,e]} \geq \Delta_{[s_1,e_1+1]} + \Delta_{[s_2,e_2+1]} + \ldots + \Delta_{[s_{\Gamma-1},e_{\Gamma-1}+1]} \geq \sum_{k=1}^{\Gamma-1} \sqrt{\frac{\log|J|}{(e_k - s_k)^{1/2} + 2}} = \sum_{k=1}^{\Gamma-1} \sqrt{\frac{\log|J|}{\mathcal{I}_k^{1/2} + 1}}.$$

On the other hand, by Holder's inequality,

$$\left(\sum_{k=1}^{\Gamma-1} \sqrt{\frac{\log|J|}{|\mathcal{I}_k|^{1/2} + 1}}\right)^{\frac{2}{3}} \left(\sum_{k=1}^{\Gamma-1} \left(|\mathcal{I}_k|^{1/2} + 1\right)\right)^{\frac{1}{3}} \geq (\Gamma - 1)\,(\log|J|)^{\frac{1}{3}}.$$

Combining the two inequalities, we have

$$\Gamma - 1 \leq (\log|J|)^{-\frac{1}{3}} \left(\sum_{k=1}^{\Gamma-1} \left(|\mathcal{I}_k|^{1/2} + 1\right)\right)^{\frac{1}{3}} \Delta_{[s,e]}^{\frac{2}{3}} \leq O\left((\log|J|)^{-\frac{2}{5}} |\mathcal{J}|^{\frac{1}{5}} \Delta_{[s,e]}^{\frac{4}{5}} + 1\right).$$

To prove the upper bound using $S_{\mathcal{J}}$, recall the condition $\Delta_{[s_k,t+1]} > \sqrt{\frac{\log|J|}{(t-s_k)^{1/2}+2}}$. Each distribution switch only creates one interval, then we have $\Gamma - 1 \leq S_{\mathcal{J}} - 1$.

Proof of Lemma 7.13. In Section 7.4, we sketched the proof for an upper bound for regret for block J, where we only considered the first $\Gamma - 1$ complete intervals. Here we show the omitted details in the proof. Following the technique in [8], we define $\mathcal{J}' := [\tau_i, \tau_i + 2^j L]$ to be the block that differs from $\mathcal{J}$ only in that $\mathcal{J}$ is assumed not triggering the restart. Note that following the same partitioning procedure, we have $\mathcal{J}' = \mathcal{I}_1' \cup \mathcal{I}_2' \cup \cdots \cup \mathcal{I}_{\Gamma'}'$. We can check that $\Gamma \leq \Gamma'$, $\mathcal{I}_k' = \mathcal{I}_k$ for $k = 1, 2, \ldots, \Gamma - 1$. Moreover, let $\mathcal{I}_\Gamma = [s_\Gamma, e_\Gamma]$ and $\mathcal{I}_\Gamma' = [s_\Gamma', e_\Gamma']$. We have $s_\Gamma = s_\Gamma'$ and $e_\Gamma \leq e_\Gamma'$.

By direct computation, we bound the first term in (13) by

$$\begin{aligned}
\sum_{k=1}^{\Gamma} |\mathcal{I}_k| \alpha_{\mathcal{I}_k} &= \sum_{k=1}^{\Gamma-1} |\mathcal{I}_k'| \alpha_{\mathcal{I}_k'} + |\mathcal{I}_\Gamma| \alpha_{\mathcal{I}_\Gamma} \\
&\leq \sum_{k=1}^{\Gamma-1} |\mathcal{I}_k'| \alpha_{\mathcal{I}_k'} + |\mathcal{I}_\Gamma'| \alpha_{\mathcal{I}_\Gamma'} \\
&\leq \sum_{k=1}^{\Gamma} \sqrt{|\mathcal{I}_k'|} \log|\mathcal{I}_k'| \\
&\leq \sqrt{\Gamma \sum_{k=1}^{\Gamma} |\mathcal{I}_k'|} \\
&\leq O\left(\min\left\{|\mathcal{J}|^{\frac{3}{5}} \Delta_{\mathcal{J}}^{\frac{2}{5}}, \sqrt{S\mathcal{J}}\right\}\right),
\end{aligned}$$

where the last inequality comes from applying Cauchy-Schwarz inequality and plugging in

$$\Gamma = O\left(\min\left\{S_J, (\log|J|)^{-\frac{2}{5}} \Delta_{\mathcal{J}}^{\frac{4}{5}} |\mathcal{J}|^{\frac{1}{5}} + 1\right\}\right).$$

In the following, we upper bound the second term in (13). Note that $\mathcal{I}_\Gamma \subset \mathcal{I}_\Gamma'$, we only need to bound $\sum_{k=1}^{\Gamma} |\mathcal{I}_k'| \varepsilon_{\mathcal{I}_k'} \mathbb{1}\{\varepsilon_{\mathcal{I}_k'} \geq \alpha_{\mathcal{I}_k'}\}$. We follow [8] and prove the following adapted lemma.

Lemma E.3. *With probability at least $1 - \delta$, it holds that*

$$\sum_{k=1}^{\Gamma} |\mathcal{I}_k'| \varepsilon_{\mathcal{I}_k'} \mathbb{1}\{\varepsilon_{\mathcal{I}_k'} \geq \alpha_{\mathcal{I}_k'}\} \leq O\left(\min\left\{|\mathcal{J}|^{\frac{3}{5}} \Delta_{\mathcal{J}}^{\frac{2}{5}}, \sqrt{S\mathcal{J}}\right\}\right).$$

Proof. Define $\mathcal{M} = \{k \in [\Gamma] | \varepsilon_{I'_k} \geq \alpha_{I'_k}\}$. Let m_k be the index defined in Lemma E.2. By definition,

$$\begin{aligned}\sum_{k=1}^{\Gamma} |I'_k| \varepsilon_{I'_k} \mathbb{1}\{\varepsilon_{I'_k} \geq \alpha_{I'_k}\} &= \sum_{k\in\mathcal{M}} |I'_k| \varepsilon_{I'_k} \\ &\leq \sum_{k\in\mathcal{M}} (|I'_k| - 2^{m_k} L)\varepsilon_{I'_k} + \sum_{k\in\mathcal{M}} 2^{m_k} L \times \varepsilon_{I'_k}.\end{aligned}$$

Following a similar derivation as in Chen et al. [8, Lemma 26], the first term is bounded by $O\left(\sqrt{|\mathcal{J}|}\log T\right)$ with probability at least $1-\delta$. Specifically,

$$\begin{aligned}\sum_{k\in\mathcal{M}} (|I'_k| - 2^{m_k} L)\varepsilon_{I'_k} &= \sum_{k\in\mathcal{M}} \sum_{t\in[s_k+2^{m_k}L, e_k]} \varepsilon_{I'_k} \\ &\leq \sum_{k\in\mathcal{M}} \sum_{t\in[s_k+2^{m_k}L, e_k]} C_5 \cdot \frac{1}{\sqrt{2^{m_k} L}} \\ &= \sum_{k\in\mathcal{M}} \sum_{t\in[s'_k+2^{m_k}L, e'_k]} C_5 \cdot \frac{1}{\sqrt{2^{m_k} L}} \mathbb{1}\{t \leq e_\Gamma\} \\ &= \varphi(e_\Gamma),\end{aligned}$$

where we define $\varphi(\tau) = \sum_{k\in\mathcal{M}} \sum_{t\in[s'_k+2^{m_k}L, e'_k]} C_5 \cdot \frac{1}{\sqrt{2^{m_k}L}} \mathbb{1}\{t \leq \tau\}$. By definition, we have $\mathbf{Pr}[\phi(e_\Gamma) > \phi(\tau)] \leq \mathbf{Pr}[e_\Gamma > \tau]$. By Lemma E.2, if the algorithm has not been restarted till time τ, for all k such that $e'_k \leq \tau$, the algorithm must have missed all opportunities to start an exploration phase with index m_k. And for the k with $\tau \in \left[s'_k, e'_k\right]$ the algorithm must have missed all opportunities to start a exploration phase with index in $\left[s'_k, \tau - 2^{m_k}L\right]$. Define $p_m = \frac{1}{L} 2^{-j/2} 2^{-m/2}$. Hence, we have

$$\begin{aligned}\mathbf{Pr}[e_\Gamma > \tau] &\leq \prod_{k\in\mathcal{M}} \prod_{t\in\left[s'_k, e'_k - 2^{m_k}L\right]} \left(1 - p_{m_k} \mathbf{1}\left\{t \leq \tau - 2^{m_k}L\right\}\right) \\ &\leq \prod_{k\in\mathcal{M}} \prod_{t\in\left[s'_k+2^{m_k}L, e'_k\right]} \left(1 - q_{m_k} \mathbf{1}\{t \leq \tau\}\right) \\ &\leq \prod_{k\in\mathcal{M}} \prod_{t\in\left[s'_k+2^{m_k}L, e'_k\right]} \exp\left(-q_{m_k} \mathbf{1}\{t \leq \tau\}\right) \\ &\leq \prod_{k\in\mathcal{M}} \prod_{t\in\left[s'_k+2^{m_k}L, e'_k\right]} \left(1 - q_{m_k} \mathbf{1}\{t \leq \tau\}\right) \\ &\leq \exp\left(-\sum_{k\in\mathcal{M}} \sum_{t\in\left[s'_k+2^{m_k}L, e'_k\right]} q_{m_k} \mathbf{1}\{t \leq \tau\}\right) \\ &= \exp\left(-\frac{\varphi(\tau)}{C_5\sqrt{2^j L}} \mathbf{1}\{t \leq \tau\}\right) \\ &= \exp\left(-\frac{\varphi(\tau)}{C_5\sqrt{|\mathcal{J}|}} \mathbf{1}\{t \leq \tau\}\right).\end{aligned}$$

Define $z = \left(\frac{1}{\sqrt{L|\mathcal{J}|}} + \log(1/\delta)\right) C_5\sqrt{|\mathcal{J}|}$ and pick τ such that $\varphi(\tau) \le z \le \varphi(\tau+1)$. If no such z exists, then $\mathbf{Pr}[\phi(e_\Gamma) > \phi(\tau)] = 0$. Then we have $\varphi(\tau) > \varphi(\tau+1) - \frac{C_5}{\sqrt{L}} > z - \frac{C_5}{\sqrt{L}}$ and

$$\mathbf{Pr}[\phi(e_\Gamma) > z] \le \mathbf{Pr}[\phi(e_\Gamma) > \phi(\tau)] \le \exp\left(-\frac{z}{C_5\sqrt{|\mathcal{J}|}} + \frac{C_5}{C_5\sqrt{|\mathcal{J}|}\sqrt{L}}\right) = \delta.$$

Hence, $\phi(e_\Gamma) \le \left(\frac{1}{\sqrt{L|\mathcal{J}|}} + \log(1/\delta)\right) C_5\sqrt{|\mathcal{J}|}$ with probability at least $1-\delta$.

The second term is bounded as

$$\begin{aligned}\sum_{k\in\mathcal{M}} 2^{m_k} L \times \varepsilon_{\mathcal{I}'_k} &\le \sum_{k\in\mathcal{M}} 2^{m_k} L \times C_5 \frac{1}{\sqrt{2^{m_k} L}} \\ &= \sum_{k\in\mathcal{M}} C_5\sqrt{|\mathcal{I}_k|} \\ &\le C_5\sqrt{\Gamma \sum_{k\in\mathcal{M}} |\mathcal{I}_k|}.\end{aligned}$$

Plugging in $\Gamma = O\left(\min\left\{S_J, (\log|J|)^{-\frac{2}{5}} \Delta_{\mathcal{J}}^{\frac{4}{5}} |\mathcal{J}|^{\frac{1}{5}} + 1\right\}\right)$ concludes the proof. □

F PROOF OF THEOREM 8.3

Our goal is to prove that for the randomized instance described in Section 8, algorithm RestartLQR with the optimally tuned window size W and exploration noise σ yields regret $\Omega(V_T^{1/3}T^{2/3})$.

We first begin by noting that using the sequence of controllers $K_t = K^*(\Theta_t)$ incurs a total cost of at most $\sum_{t\in[T]} J_t^* + O(S)$, where S denotes the number of switches in the hypothesis. This is because for an interval $\{\tau_1, \tau_1+1, \ldots, \tau_2\}$ where the dynamics remain fixed at θ with optimal parameters p^*, k^*, J^*, we have

$$\sum_{t=\tau_1}^{\tau_2} x_t^2 + u_t^2 = \sum_{t=\tau_1}^{\tau_2} x_t^2 + (k^* x_t)^2 = \sum_{t=\tau_1}^{\tau_2} J^* + p^* x_{\tau_1}^2 - p^* \mathbf{E}\left[x_{\tau_2+1}^2\right] \le \sum_{t=\tau_1}^{\tau_2} J^* + p^* x_{\tau_1}^2.$$

Furthermore, since the optimal controllers yield $|a + b_t k_t^*| = \left|a\frac{1-b_t^2 p_t^*}{1+b_t^2 p_t^*}\right| \le a = \frac{1}{\sqrt{5}}$, $\mathbf{E}[x_t]$ is bounded for all $t \in [T]$.

We next show that the loss for the optimally tuned RestartLQR algorithm is at least $\sum_t J_t^* + \Omega\left(V_T^{1/3}T^{2/3}\right)$. We will use the following lemma from [7].

Lemma F.1 (Lemma 14 in [7]). *Let $I = \{\tau_1, \ldots, \tau_2\}$ be an interval with dynamics $a = 1/\sqrt{5}$, $b_t = b$ with $|b| \le 0.05$, $\mathbf{E}\left[w_t^2\right] = \psi^2$, and optimal policy parameters k^*, J^*. Then for an arbitrary admissible control policy $\{u_t\}$,*

$$\mathbf{E}\left[\sum_{t\in I} x_t^2 + u_t^2\right] - |I|J^* \ge 0.99\mathbf{E}\left[\sum_{t\in I}(u_t - k^* x_t)^2\right] - 4\psi^2, \tag{70}$$

as well as:

$$\mathbf{E}\left[\sum_{t\in I} x_t^2 + u_t^2\right] - |I|J^* \ge \frac{1}{3}\mathbf{E}\left[\sum_{t\in I} u_t^2\right] - \frac{5}{6}\psi^2(k^*)^2|I|. \tag{71}$$

We begin by defining the random variables that specify the instance. Let $\{\mu_t\}$ $(t = 1, 2, \ldots, T)$ be the sequence specifying the *magnitude* of changes in b_t, defined so that $\mu_1 = \epsilon$, and $\mu_2, \ldots, \mu_T$ are i.i.d. random variables with the following distribution:

$$\mu_t = \begin{cases} 0.05 & \text{with probability } \frac{V_T}{2T}, \\ \epsilon & \text{with probability } \left(\frac{V_T}{T}\right)^{5/6}, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\epsilon = 0.05 \cdot (V_T/T)^{1/6}.$$

Let $\{\chi_t\}$ be the sequence specifying the *sign* of changes in b_t, defined so that $\chi_1 = 1$ and χ_t for $t \ge 2$ are i.i.d. Rademacher random variables (i.e., ± 1 with equal probability). Given the above, the sequence b_t is defined as

$$b_t = b_{t-1} \cdot \mathbb{1}_{\mu_t = 0} + \mu_t \cdot \chi_t.$$

Let

$$\mathcal{H}_t = \{w_s, \eta_s, \sigma_s, \mu_s, \chi_s\}_{s=1}^t$$

denote the history of the dynamics and instances until time t. Recall that the RestartLQR(W) family of algorithms partition the horizon into contiguous non-overlapping windows of size W. We will use $I_i = [W \cdot (i-1) + 1, \ldots, W \cdot i]$ to denote the i-th window. The control for $t \in I_i$ is chosen as $\widehat{k}_{(i)} + \sigma_t \eta_t$ where η_t are i.i.d. $\mathcal{N}(0,1)$ and σ_t is an arbitrary adapted sequence of exploration energy

injected by the algorithm. With some abuse of notation, we will use $k^*(b)$ to denote the optimal linear feedback controller as a function of b (with $a = 1/\sqrt{5}$ and $w_t \sim \mathcal{N}(0, \psi^2)$ implicit), and note that $k^*(b) = -k^*(-b)$.

We will partition our windows into three sets:

(1) $\mathcal{I}_1$: windows i which have at least one $\mu_t = 0.05$ for $t \in I_i$; let $\tau_1(i) \in I_i$ be the first time such that $\mu_{\tau_1(i)} \neq 0$,
(2) $\mathcal{I}_\epsilon$: *pairs* of contiguous windows $(i, i+1)$ with $\mu_t = 0$ for all $t \in I_i \cup I_{i+1}$, and $|b_t| = \epsilon$,
(3) $\mathcal{I}_2$: the remaining windows.

Note that this partition is not unique. In particular, there could be many ways to pair up contiguous windows with small b_t and no change of dynamics to create the second set. We pick any such maximal partition.

We can use (70) and (71) to express the total cost of the algorithm as:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t\in[T]} x_t^2 + u_t^2\right] - \sum_{t\in[T]} J_t^* \\
&\qquad \geq\ 0.99\mathbf{E}\left[\sum_{t\in T}(u_t - k_t^* x_t)^2\right] - 4\psi^2 S \\
&\qquad \geq \sum_{i\in\mathcal{I}_1} 0.99\mathbf{E}\left[\sum_{t=\tau_1(i)}^{i\cdot W}(u_t - k_t^* x_t)^2\right] \\
&\qquad\quad + \sum_{(i,i+1)\in\mathcal{I}_\epsilon}\left(\frac{1}{3}\mathbf{E}\left[\sum_{t\in I_i} u_t^2\right] - \frac{5}{6}\psi^2(k^*(\epsilon))^2 W + 0.99\mathbf{E}\left[\sum_{t\in I_{i+1}}(u_t - k_t^* x_t)^2\right]\right) - \psi^2 S. \qquad (72)
\end{aligned}$$

Begin by considering the event $\mathcal{E}_{i,1} := \{i \in I_1\}$. Conditioning on this event, $\tau_1(i)$ is uniformly distributed in I_i. Furthermore, the sign of $b_{\tau_1(i)}$ is ± 1 with equal probability. We thus bound the contribution to regret due to windows in $\mathcal{I}_1$ as:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=\tau_1(i)}^{i\cdot W}(u_t - k_t^* x_t)^2 \,\middle|\, \mathcal{E}_{i,1}, \mathcal{H}_t\right] \\
&= \mathbf{E}\left[\sum_{t=\tau_1(i)}^{i\cdot W}(\sigma_t\eta_t + (\widehat{k}_{(i)} - k_t^*)x_t)^2 \,\middle|\, \mathcal{E}_{i,1}, \mathcal{H}_t\right] \\
&\geq \mathbf{E}\left[\sum_{t=\tau_1(i)}^{i\cdot W}(\widehat{k}_{(i)} - k_t^*)^2 x_t^2 \,\middle|\, \mathcal{E}_{i,1}, \mathcal{H}_t\right] \\
&\geq \psi^2\mathbf{E}\left[\sum_{t=\tau_1(i)}^{i\cdot W}(\widehat{k}_{(i)} - k_t^*)^2 \,\middle|\, \mathcal{E}_{i,1}, \mathcal{H}_t\right] \\
&\geq \psi^2\mathbf{E}\left[\sum_{t=(i-1)W+1}^{i\cdot W}\mathbb{1}_{\{t\geq\tau_1(i),|b_t|=1\}}(\widehat{k}_{(i)} - k_t^*)^2 \,\middle|\, \mathcal{E}_{i,1}, \mathcal{H}_t\right] \\
&\geq \psi^2\mathbf{E}[|t : t \geq \tau_1(i), |b_t| = 1|]\mathbf{E}\left[(\widehat{k}_{(i)} - k^*(b))^2 \,\middle|\, \mathcal{E}_{i,1}, \mathcal{H}_t\right],
\end{aligned}$$

where b denotes a random variable that is ± 0.05 with equal probability;

$$\begin{aligned}
&= \psi^2 \cdot \mathbf{E}[|t : t \geq \tau_1(i), |b_t| = 0.05|] \cdot \left(\frac{1}{2}(\widehat{k}_{(i)} - k^*(0.05))^2 + \frac{1}{2}(\widehat{k}_{(i)} + k^*(0.05))^2 \right) \\
&\geq \psi^2 \cdot \mathbf{E}[|t : t \geq \tau_1(i), |b_t| = 0.05|] \cdot k^*(0.05)^2 \\
&\geq \psi^2 \cdot \mathbf{E}[|t : t \geq \tau_1(i), |b_t| = 0.05|] \cdot \frac{1}{4000}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i \in \mathcal{I}_1} \mathbf{E}\left[\sum_{t=\tau_1(i)}^{i \cdot W} (u_t - k_t^* x_t)^2 \middle| \mathcal{E}_{i,1}, \mathcal{H}_t \right] &= \sum_i \mathbf{E}\left[\sum_{t=\tau_1(i)}^{i \cdot W} (u_t - k_t^* x_t)^2 \middle| \mathcal{E}_{i,1}, \mathcal{H}_t \right] \mathbf{E}\left[\mathbb{1}_{\mathcal{E}_{i,1}} \right] \\
&\geq \sum_i \psi^2 \cdot \mathbf{E}[|t : t \geq \tau_1(i), |b_t| = 0.05|] \cdot \frac{1}{4000} \mathbf{E}\left[\mathcal{E}_{i,1} \right] \\
&= \frac{\psi^2}{4000} \sum_i \sum_{t \in I_i} \mathbf{Pr}[\mu_t = 0.05] \mathbf{E}\Big[\min\{ i \cdot W - t + 1, \mathrm{Geom}(V_T/T + (V_T/4T)^{5/6}) \} \Big],
\end{aligned}$$

where $\mathrm{Geom}(p)$ denotes a Geometric random variable with success probability p. For any non-negative integer-valued random variable X with median $\widehat{X}$ and non-negative integer a, we have the identity,

$$\mathbf{E}[\min\{X, a\}] = \sum_{x=1}^{a} \mathbf{Pr}[X \geq x] \geq \sum_{x=1}^{\min\{a, \widehat{X}\}} \mathbf{Pr}[X \geq x] \geq \frac{\min\{\widehat{X}, a\}}{2}.$$

For $X \sim \mathrm{Geom}(p)$, we have $\widehat{X} \geq \frac{1}{5p}$, which finally gives,

$$\begin{aligned}
\sum_{i \in \mathcal{I}_1} \mathbf{E}\left[\sum_{t=\tau_1(i)}^{i \cdot W} (u_t - k_t^* x_t)^2 \middle| \mathcal{E}_{i,1}, \mathcal{H}_t \right] &\geq \frac{\psi^2}{4000} \sum_i \sum_{t \in I_i} \frac{V_T}{2T} \frac{\min\{W, 0.1(4T/V_T)^{5/6}\}}{2} \\
&= \frac{\psi^2}{8000} V_T \min\left\{ W, 0.1(4T/V_T)^{5/6} \right\}. \qquad (73)
\end{aligned}$$

Note that if $W = \Omega\left((T/V_T)^{2/3}\right)$, then (73) already gives the regret lower bound of the Theorem. Therefore, henceforth we will assume $W = O\left((T/V_T)^{2/3}\right)$.

Next, we turn to windows in $\mathcal{I}_\epsilon$. Specifically, pick a pair $(i, i+1)$, and our goal is to bound

$$\frac{1}{3}\mathbf{E}\left[\sum_{t \in I_i} u_t^2 \right] - \frac{5}{6}\psi^2 (k^*(\epsilon))^2 W + 0.99\mathbf{E}\left[\sum_{t \in I_{i+1}} (u_t - k_t^* x_t)^2 \right].$$

We next invoke yet another useful lemma from [7].

Lemma F.2 (Lemma 15 in [7]). *Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the probability laws of $\{x_t\}_{t \in I_i}$ under $b_t = +\epsilon$ and $b_t = -\epsilon$ ($\forall t \in I_i$), respectively. Then, the total variation distance between these is upper bounded as*

$$TV(\mathbb{P}_+, \mathbb{P}_-) \leq \frac{\epsilon}{\psi} \sqrt{\mathbf{E}\left[\sum_{t \in I_i} u_t^2 \right]}.$$

We will use the notation of the above lemma for the rest of the proof to bound the regret due to windows $(i, i+1)$. As before, we bound the regret in the window I_{i+1} by:

$$0.99\mathbf{E}\left[\sum_{t\in I_{i+1}}(u_t - k_t^* x_t)^2\right] \geq 0.99\psi^2 W\mathbf{E}\left[\left(\widehat{k}_{(i+1)} - k_t^*\right)^2\right]$$
$$= 0.99\psi^2 W\left(\frac{1}{2}\mathbf{E}_+\left[\left(\widehat{k}_{(i+1)} - k^*(\epsilon)\right)^2\right] + \frac{1}{2}\mathbf{E}_-\left[\left(\widehat{k}_{(i+1)} + k^*(\epsilon)\right)^2\right]\right).$$

Let F_+, F_- denote the distribution of $\widehat{k}_{(i+1)}$ under $\mathbb{P}_+, \mathbb{P}_-$, respectively, and let $g_+(k) := \left(\widehat{k}_{(i+1)} - k^*(\epsilon)\right)^2$ and $g_-(k) := \left(\widehat{k}_{(i+1)} + k^*(\epsilon)\right)^2$. Note that g_+, g_- are non-negative and

$$\frac{1}{2}\left(g_+(k) + g_-(k)\right) \geq k^*(\epsilon)^2.$$

Therefore,

$$\frac{1}{2}\mathbf{E}_+\left[\left(\widehat{k}_{(i+1)} - k^*(\epsilon)\right)^2\right] + \frac{1}{2}\mathbf{E}_-\left[\left(\widehat{k}_{(i+1)} + k^*(\epsilon)\right)^2\right]$$
$$= \frac{1}{2}\int_{\mathfrak{R}} g_+(k)dF_+(k) + \frac{1}{2}\int_{\mathfrak{R}} g_-(k)dF_-(k)$$
$$= \sup_{F\in\Gamma(F_+,F_+)}\int_{\mathfrak{R}\times\mathfrak{R}}\frac{1}{2}\left(g_+(k_1) + g_-(k_2)\right)dF(k_1,k_2)$$

where $\Gamma(F_+, F_-)$ denotes the set of stochastic couplings of measures F_+, F_-,

$$\geq \sup_{F\in\Gamma(F_+,F_+)}\int_{\mathfrak{R}\times\mathfrak{R}}\frac{1}{2}\left(g_+(k_1) + g_-(k_2)\right)\mathbb{1}_{\{k_1=k_2\}}\,dF(k_1,k_2)$$
$$\geq \sup_{F\in\Gamma(F_+,F_+)}\int_{\mathfrak{R}\times\mathfrak{R}} k^*(\epsilon)^2 dF(k_1,k_2)$$
$$\geq k^*(\epsilon)^2(1 - TV(\mathbb{P}_+, \mathbb{P}_-)).$$

We therefore have,

$$\frac{1}{3}\mathbf{E}\left[\sum_{t\in I_i} u_t^2\right] - \frac{5}{6}\psi^2(k^*(\epsilon))^2 W + 0.99\mathbf{E}\left[\sum_{t\in I_{i+1}}(u_t - k_t^* x_t)^2\right]$$
$$\geq \frac{\psi^2 TV(\mathbb{P}_+, \mathbb{P}_-)^2}{3\epsilon^2} + \psi^2 k^*(\epsilon)^2 W\left(0.99(1 - TV(\mathbb{P}_+, \mathbb{P}_-)) - \frac{5}{6}\right)$$
$$\geq \min\left\{\frac{\psi^2}{300\epsilon^2}, \frac{\psi^2 k^*(\epsilon)^2 W}{20}\right\}$$
$$\geq \min\left\{\frac{\psi^2}{300\epsilon^2}, \frac{\psi^2\epsilon^2 W}{200}\right\}. \tag{74}$$

Since we are assuming $W = O\left((T/V_T)^{2/3}\right) = o\left((T/V_T)^{5/6}\right)$ (the mean duration between switches in b_t), the expected number of pairs $(i, i+1)$ in any maximal choice of $\mathcal{I}_\epsilon$ is $\Omega\left(T/W\right)$, which gives the total regret contribution due to intervals in $\mathcal{I}_\epsilon$ of at least

$$T\cdot\min\left\{\frac{\psi^2}{300W\epsilon^2}, \frac{\psi^2\epsilon^2}{200}\right\}.$$

The expression above is decreasing in W, and for $W = O\left((T/V_T)^{2/3}\right)$ is $\Omega\left(V_T^{1/3}T^{2/3}\right)$. □

Received October 2021; revised December 2021; accepted January 2022