Efficient Reinforcement Learning for High Dimensional Linear Quadratic Systems

Morteza Ibrahimi
Stanford University
Stanford, CA 94305
ibrahimi@stanford.edu

Adel Javanmard
Stanford University
Stanford, CA 94305
adelj@stanford.edu

Benjamin Van Roy
Stanford University
Stanford, CA 94305
bvr@stanford.edu

Abstract

We study the problem of adaptive control of a high dimensional linear quadratic (LQ) system. Previous work established the asymptotic convergence to an optimal controller for various adaptive control schemes. More recently, for the average cost LQ problem, a regret bound of $O(\sqrt{T})$ was shown, apart form logarithmic factors. However, this bound scales exponentially with p, the dimension of the state space. In this work we consider the case where the matrices describing the dynamic of the LQ system are sparse and their dimensions are large. We present an adaptive control scheme that achieves a regret bound of $O(p\sqrt{T})$, apart from logarithmic factors. In particular, our algorithm has an average cost of $(1+\epsilon)$ times the optimum cost after $T = \mathrm{polylog}(p)O(1/\epsilon^2)$. This is in comparison to previous work on the dense dynamics where the algorithm requires time that scales exponentially with dimension in order to achieve regret of ϵ times the optimal cost.

We believe that our result has prominent applications in the emerging area of computational advertising, in particular targeted online advertising and advertising in social networks.

1 Introduction

In this paper we address the problem of adaptive control of a high dimensional linear quadratic (LQ) system. Formally, the dynamics of a linear quadratic system are given by

$$\begin{aligned} x(t+1) &= A^0 x(t) + B^0 u(t) + w(t+1), \\ c(t) &= x(t)^T Q x(t) + u(t)^T R u(t), \end{aligned} \tag{1}$$

where $u(t) \in \mathbb{R}^r$ is the control (action) at time t, $x(t) \in \mathbb{R}^p$ is the state at time t, $c(t) \in \mathbb{R}$ is the cost at time t, and $\{w(t+1)\}_{t\geq 0}$ is a sequence of random vectors in $\mathbb{R}^p$ with i.i.d. standard Normal entries. The matrices $Q \in \mathbb{R}^{p\times p}$ and $R \in \mathbb{R}^{r\times r}$ are positive semi-definite (PSD) matrices that determine the cost at each step. The evolution of the system is described through the matrices $A^0 \in \mathbb{R}^{p\times p}$ and $B^0 \in R^{p\times r}$. Finally by high dimensional system we mean the case where $p, r \gg 1$.

A celebrated fundamental theorem in control theory asserts that the above LQ system can be optimally controlled by a simple linear feedback if the pair (A^0, B^0) is controllable and the pair $(A^0, Q^{1/2})$ is observable. The optimal controller can be explicitly computed from the matrices describing the dynamics and the cost. Throughout this paper we assume that controllability and observability conditions hold.

When the matrix $\Theta^0 \equiv [A^0, B^0]$ is unknown, the task is that of adaptive control, where the system is to be learned and controlled at the same time. Early works on the adaptive control of LQ systems relied on the *certainty equivalence principle* [2]. In this scheme at each time t the unknown parameter Θ^0 is estimated based on the observations collected so far and the optimal controller for the

estimated system is applied. Such controllers are shown to converge to an optimal controller in the case of minimum variance cost, however, in general they may converge to a suboptimal controller [11]. Subsequently, it has been shown that introducing random exploration by adding noise to the control signal, e.g., [14], solves the problem of converging to suboptimal estimates.

All the aforementioned work have been concerned with the asymptotic convergence of the controller to an optimal controller. In order to achieve regret bounds, cost-biased parameter estimation [12, 8, 1], in particular the optimism in the face of uncertainty (OFU) principle [13] has been shown to be effective. In this method a *confidence set* S is found such that $\Theta^0 \in S$ with high probability. The system is then controlled using the *most optimistic* parameter estimates, i.e., $\widehat{\Theta} \in S$ with the smallest optimum cost. The asymptotic convergence of the average cost of OFU for the LQR problem was shown in [6]. This asymptotic result was extended in [1] by providing a bound for the cumulative regret. Assume $x(0) = 0$ and for a control policy π define the average cost

$$J_\pi = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}[c_t]\,. \tag{2}$$

Further, define the cumulative regret as

$$R(T) = \sum_{t=0}^{T} (c_\pi(t) - J_*)\,, \tag{3}$$

where $c_\pi(t)$ is the cost of control policy π at time t and $J_* = J(\Theta_0)$ is the optimal average cost. The algorithm proposed in [1] is shown to have cumulative regret $\tilde{O}(\sqrt{T})$ where $\tilde{O}$ is hiding the logarithmic factors. While no lower bound was provided for the regret, comparison with the multi-armed bandit problem where a lower bound of $O(\sqrt{T})$ was shown for the general case [9], suggests that this scaling with time for the cumulative regret is optimal.

The focus of [1] was on scaling of the regret with time horizon T. However, the regret of the proposed algorithm scales poorly with dimension. More specifically, the analysis in [1] proves a regret bound of $R(T) < Cp^{p+r+2}\sqrt{T}$. The current paper focuses on (many) applications where the state and control dimensions are much larger than the time horizon of interest. A powerful reinforcement learning algorithm for these applications should have regret which depends gracefully on dimension. In general, there is little to be achieved when $T < p$ as the number of *degrees of freedom* ($pr + p^2$) is larger than the number of observations (Tp) and any estimator can be arbitrary inaccurate. However, when there is prior knowledge about the unknown parameters A^0, B^0, *e.g.*, when A^0, B^0 are *sparse*, accurate estimation can be feasible. In particular, [3] proved that under suitable conditions the unknown parameters of a noise driven system (i.e., no control) whose dynamics are modeled by linear stochastic differential equations can be estimated accurately with as few as $O(\log(p))$ samples. However, the result of [3] is not directly applicable here since for a general feedback gain L even if A^0 and B^0 are sparse, the closed loop gain $A^0 - B^0L$ need not be sparse. Furthermore, system dynamics would be correlated with past observations through the estimated gain matrix L. Finally, there is no notion of cost in [3] while here we have to obtain bounds on cost and its scaling with p. In this work we extend the result of [3] by showing that under suitable conditions, unknown parameters of sparse high dimensional LQ systems can be accurately estimated with as few as $O(\log(p + r))$ observations. Equipped with this efficient learning method, we show that sparse high dimensional LQ systems can be adaptively controlled with regret $\tilde{O}(p\sqrt{T})$.

To put this result in perspective note that even when $x(t) = 0$, the expected cost at time $t + 1$ is $\Omega(p)$ due to the noise. Therefore, the cumulative cost at time T is bounded as $\Omega(pT)$. Comparing this to our regret bound, we see that for $T = \mathrm{polylog}(p)O(\frac{1}{\epsilon^2})$, the cumulative cost of our algorithm is bounded by $(1 + \epsilon)$ times the optimum cumulative cost. In other words, our algorithm performs close to optimal after $\mathrm{polylog}(p)$ steps. This is in contrast with the result of [1] where the algorithm needs $\Omega(p^{2p})$ steps in order to achieve regret of ϵ times the optimal cost.

Sparse high dimensional LQ systems appear in many engineering applications. Here we are particularly motivated by an emerging field of applications in marketing and advertising. The use of dynamical optimal control models in advertising has a history of at least four decades, cf. [17, 10] for a survey. In these models, often a partial differential equation is used to describe how advertising expenditure over time translates into sales. The basic problem is to find the advertising expenditure that maximizes the net profit. The focus of these works is to model the temporal dynamics of

the advertising expenditure (the control variable) and the variables of interest (sales, goodwill level, etc.). There also exists a rich literature studying the *spatial* interdependence of consumers' and firms' behavior to devise marketing schemes [7]. In these models space can be generalized beyond geographies to include notions like demographies and psychometry.

Combination of spatial interdependence and temporal dynamics models for optimal advertising was also considered [16, 15]. A simple temporal dynamics model is extended in [15] by allowing state and control variables to have spatial dependence and introducing a diffusive component in the controlled PDE which describes the spatial dynamics. The controlled PDE is then showed to be equivalent to an abstract linear control system of the form

$$\frac{\mathrm{dx(t)}}{\mathrm{dt}} = Ax(t) + Bu(t). \tag{4}$$

Both [15] and [7] are concerned with the optimal control and the interactions are either dictated by the model or assumed known. Our work deals with a discrete and noisy version of (4) where the dynamics is to be estimated but is known to be sparse. In the model considered in [15] the state variable x lives in an infinite dimensional space. Spatial models in marketing [7] usually consider state variables which have a large number of dimensions, e.g., number of zip codes in the US ($\sim$ 50K). High dimensional state space and control is a recurring theme in these applications.

In particular, with the modern social networks customers are classified in a highly granular way, potentially with each customer representing his own class. With the number of classes and complexity of their interactions, its unlikely that we could formulate an effective model a priori for how classes interact. Further, the nature of these interactions change over time with the changing landscape of Internet services and information available to customers. This makes it important to efficiently learn from real-time data about the nature of these interactions.

Notation: We bundle the unknown parameters into one variable $\Theta^0 = [A^0, B^0] \in \mathbb{R}^{p\times q}$ where $q = p + r$ and call it the interaction matrix. For $v \in \mathbb{R}^n$, $M \in \mathbb{R}^{m\times n}$ and $p \geq 1$, we denote by $\|v\|_p$ the standard p-norm and by $\|M\|_p$ the corresponding operator norm. For $1 \leq i \leq m$, M_i represents the i^{th} row of matrix M. For $S \subseteq [m], J \subseteq [n]$, M_{SJ} is the submatrix of M formed by the rows in S and columns in J. For a set S denote by $|S|$ its cardinality. For an integer n denote by $[n]$ the set $\{1, \ldots, n\}$.

2 Algorithm

Our algorithm employs the *Optimism in the Face of Uncertainty* (OFU) principle in an episodic fashion. At the beginning of episode i the algorithm constructs a *confidence set* $\Omega^{(i)}$ which is guaranteed to include the unknown parameter Θ^0 with high probability. The algorithm then chooses $\widetilde{\Theta}^{(i)} \in \Omega^{(i)}$ that has the smallest expected cost as the estimated parameter for episode i and applies the optimal control for the estimated parameter during episode i.

The confidence set is constructed using observations from the last episode only but the length of episodes are chosen to increase geometrically allowing for more accurate estimates and shrinkage of the confidence set by a constant factor at each episode. The details of each step and the pseudo code for the algorithm follows.

Constructing confidence set: Define τ_i to be the start of episode i with $\tau_0 = 0$. Let $L^{(i)}$ be the controller that has been chosen for episode i. For $t \in [\tau_i, \tau_{i+1})$ the system is controlled by $u(t) = -L^{(i)}x(t)$ and the system dynamics can be written as $x(t+1) = (A^0 - B^0L^{(i)})x(t) + w(t+1)$. At the beginning of episode $i+1$, first an initial estimate $\widehat{\Theta}$ is obtained by solving the following convex optimization problem for each row $\Theta_u \in \mathbb{R}^q$ separately:

$$\widehat{\Theta}_u^{(i+1)} \in \text{argmin } \mathcal{L}(\Theta_u) + \lambda\|\Theta_u\|_1, \tag{5}$$

where

$$\mathcal{L}(\Theta_u) = \frac{1}{2\Delta\tau_{i+1}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \{x_u(t+1) - \Theta_u\widetilde{L}^{(i)}x(t)\}^2, \quad \Delta\tau_{i+1} = \tau_{i+1} - \tau_i, \tag{6}$$

ALGORITHM: Reinforcement learning algorithm for LQ systems.

Input: Precision ϵ, failure probability 4δ, initial $(\rho, C_{\min}, \alpha)$ identifiable controller $L^{(0)}$, $\ell(\Theta^0, \epsilon)$
Output: Series of estimates $\widetilde{\Theta}^{(i)}$, confidence sets $\Omega^{(i)}$ and controllers $L^{(i)}$

1: Let $\ell_0 = \max(1, \max_{j\in[r]} \|L_j^{(0)}\|_2)$, and

$$n_0 = \frac{4\cdot 10^3\, k^2 \ell_0^2}{\alpha(1-\rho)C_{\min}^2}\left(\frac{1}{\epsilon^2} + \frac{k}{(1-\rho)^2}\right)\log(\frac{4kq}{\delta}),$$
$$n_1 = \frac{4\cdot 10^3\, k^2 \ell(\Theta^0,\epsilon)^2}{(1-\rho)C_{\min}^2}\left(\frac{1}{\epsilon^2} + \frac{k}{(1-\rho)^2}\right)\log(\frac{4kq}{\delta}).$$

Let $\Delta\tau_0 = n_0$, $\Delta\tau_i = 4^i(1 + i/\log(q/\delta))n_1$ for $i \geq 1$, and $\tau_i = \sum_{j=0}^{i} \Delta\tau_j$.
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Apply the control $u(t) = -L^{(i)}x(t)$ until $\tau_{i+1} - 1$ and observe the trace $\{x(t)\}_{\tau_i \leq t < \tau_{i+1}}$.
4: Calculate the estimate $\widehat{\Theta}^{(i+1)}$ from (5) and construct the confidence set $\Omega^{(i+1)}$.
5: Calculate $\widetilde{\Theta}^{(i+1)}$ from (9) and set $L^{(i+1)} \leftarrow L(\widetilde{\Theta}^{(i+1)})$.

and $\widetilde{L}^{(i)} = [I, -L^{(i)\mathsf{T}}]^{\mathsf{T}}$. The estimator $\widehat{\Theta}_u$ is known as the LASSO estimator. The first term in the cost function is the normalized negative log likelihood which measures the fidelity to the observations while the second term imposes the sparsity constraint on Θ_u. λ is the regularization parameter.

For $\Theta^{(1)}, \Theta^{(2)} \in \mathbb{R}^{p\times q}$ define the distance $\mathsf{d}(\Theta^{(1)}, \Theta^{(2)})$ as

$$\mathsf{d}(\Theta^{(1)}, \Theta^{(2)}) = \max_{u\in[p]} \|\Theta_u^{(1)} - \Theta_u^{(2)}\|_2, \tag{7}$$

where Θ_u is the u^{th} row of the matrix Θ. It is worth noting that for k-sparse matrices with k constant, this distance does not scale with p or q. In particular, if the absolute value of the elements of $\Theta^{(1)}$ and $\Theta^{(2)}$ are bounded by $\Theta_{\max}$ then $\mathsf{d}(\Theta^{(1)}, \Theta^{(2)}) \leq 2\sqrt{k}\Theta_{\max}$.

Having the estimator $\widehat{\Theta}^{(i)}$ the algorithm constructs the confidence set for episode i as

$$\Omega^{(i)} = \{\Theta \in \mathbb{R}^{p\times q} \mid \mathsf{d}(\Theta, \widehat{\Theta}^{(i)}) \leq 2^{-i}\epsilon\}, \tag{8}$$

where $\epsilon > 0$ is an input parameter to the algorithm. For any fixed $\delta > 0$, by choosing τ_i judiciously we ensure that with probability at least $1 - \delta$, $\Theta^0 \in \Omega^{(i)}$, for all $i \geq 1$. (see Theorem 3.2).

Design of the controller: Let $J(\Theta)$ be the minimum expected cost if the interaction matrix is $\Theta = [A, B]$ and denote by $L(\Theta)$ the optimal controller that achieves the expected cost $J(\Theta)$. The algorithm implements OFU principle by choosing, at the beginning of episode i, an estimate $\widetilde{\Theta}^{(i)} \in \Omega^{(i)}$ such that

$$\widetilde{\Theta}^{(i)} \in \operatorname*{argmin}_{\Theta\in\Omega^{(i)}} J(\Theta). \tag{9}$$

The optimal control corresponding to $\widetilde{\Theta}^{(i)}$ is then applied during episode i, i.e., $u(t) = -L(\widetilde{\Theta}^{(i)})x(t)$ for $t \in [\tau_i, \tau_{i+1})$. Recall that for $\Theta = [A, B]$, the optimal controller is given through the following relations

$$K(\Theta) = Q + A^{\mathsf{T}}K(\Theta)A - A^{\mathsf{T}}K(\Theta)B(B^{\mathsf{T}}K(\Theta)B + R)^{-1}B^{\mathsf{T}}K(\Theta)A\,, \quad \text{(Riccati equation)}$$
$$L(\Theta) = (B^{\mathsf{T}}K(\Theta)B + R)^{-1}B^{\mathsf{T}}K(\Theta)A\,.$$

The pseudo code for the algorithm is summarized in the table.

3 Main Results

In this section we present performance guarantees in terms of cumulative regret and learning accuracy for the presented algorithm. In order to state the theorems, we first need to present some assumptions on the system.

Given $\Theta \in \mathbb{R}^{p\times q}$ and $L \in \mathbb{R}^{r\times p}$, define $\widetilde{L} = [I, -L^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{q\times p}$ and let $\Lambda \in \mathbb{R}^{p\times p}$ be a solution to the following Lyapunov equation

$$\Lambda - \Theta\widetilde{L}\Lambda\widetilde{L}^{\mathsf{T}}\Theta^{\mathsf{T}} = I. \tag{10}$$

If the closed loop system $(A^0 - B^0L)$ is stable then the solution to the above equation exists and the state vector $x(t)$ has a Normal stationary distribution with covariance Λ.

We proceed by introducing an *identifiable regulator*.

Definition 3.1. *For a k-sparse matrix $\Theta^0 = [A^0, B^0] \in \mathbb{R}^{p\times q}$ and $L \in \mathbb{R}^{r\times p}$, define $\widetilde{L} = [I, -L^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{q\times p}$ and let $H = \widetilde{L}\Lambda\widetilde{L}^{\mathsf{T}}$ where Λ is the solution of Eq. (10) with $\Theta = \Theta^0$. Define L to be $(\rho, C_{\min}, \alpha)$ identifiable (with respect to Θ^0) if it satisfies the following conditions for all $S \subseteq [q]$, $|S| \le k$.*

$$(1)\ \|A^0 - B^0L\|_2 \le \rho < 1, \quad (2)\ \lambda_{\min}(H_{SS}) \ge C_{\min}, \quad (3)\ \|H_{S^cS}H_{SS}^{-1}\|_\infty \le 1 - \alpha.$$

The first condition simply states that if the system is controlled using the regulator L then the closed loop autonomous system is asymptotically stable. The second and third conditions are similar to what is referred to in the sparse signal recovery literature as the *mutual incoherence* or *irrepresentable* conditions. Various examples and results exist for the matrix families that satisfy these conditions [18]. Let S be the set of indices of the nonzero entries in a specific row of Θ^0. The second condition states that the corresponding entries in the extended state variable $y = [x^{\mathsf{T}}, u^{\mathsf{T}}]$ are sufficiently distinguishable from each other. In other words, if the trajectories corresponding to this group of state variables are observed, non of them can be *well approximated* as a linear combination of the others. The third condition can be thought of as a quantification of the first vs. higher order dependencies. Consider entry j in the extended state variable. Then, the dynamic of y_j is directly influenced by entries y_S. However they are also influenced indirectly by other entries of y. The third condition roughly states that the indirect influences are sufficiently weaker than the direct influences. There exists a vast literature on the applicability of these conditions and scenarios in which they are known to hold. These conditions are *almost* necessary for the successful recovery by ℓ_1 relaxation. For a discussion on these and other similar conditions imposed for sparse signal recovery we refer the reader to [19] and [20] and the references therein.

Define $\Theta_{\min} = \min_{i\in[p],j\in[q],\Theta^0_{ij}\neq 0} |\Theta^0_{ij}|$. Our first result states that the system can be learned efficiently from its trajectory observations when it is controlled by an identifiable regulator.

Theorem 3.2. *Consider the LQ system of Eq. (1) and assume $\Theta^0 = [A^0, B^0]$ is k-sparse. Let $u(t) = -Lx(t)$ where L is a $(\rho, C_{\min}, \alpha)$ identifiable regulator with respect to Θ^0 and define $\ell = \max(1, \max_{j\in[r]} \|L_j\|_2)$. Let n denote the number of samples of the trajectory that is observed. For any $0 < \epsilon < \min(\Theta_{\min}, \frac{\ell}{2}, \frac{3}{1-\rho})$, there exists λ such that, if*

$$n \ge \frac{4\cdot 10^3\, k^2\ell^2}{\alpha^2(1-\rho)C_{\min}^2}\left(\frac{1}{\epsilon^2} + \frac{k}{(1-\rho)^2}\right)\log(\frac{4kq}{\delta})\,, \tag{11}$$

then the ℓ_1-regularized least squares solution $\widehat{\Theta}$ of Eq. (5) satisfies $\mathsf{d}(\widehat{\Theta}, \Theta^0) \le \epsilon$ with probability larger than $1-\delta$. In particular, this is achieved by taking $\lambda = 6\ell\sqrt{\log(4q/\delta)/(n\alpha^2(1-\rho))}$.

Our second result states that equipped with an efficient learning algorithm, the LQ system of Eq. (1) can be controlled with regret $\tilde{O}(p\sqrt{T}\log^{\frac{3}{2}}(1/\delta))$ under suitable assumptions.

Define an ϵ-neighborhood of Θ^0 as $\mathcal{N}_\epsilon(\Theta^0) = \{\Theta \in \mathbb{R}^{p\times q} \mid \mathsf{d}(\Theta^0, \Theta) \le \epsilon\}$. Our assumption asserts the identifiably of $L(\Theta)$ for Θ close to Θ^0.

Assumption: There exist $\epsilon, C > 0$ such that for all $\Theta \in \mathcal{N}_\epsilon(\Theta^0)$, $L(\Theta)$ is identifiable w.r.t. Θ^0 and

$$\sigma_L(\Theta^0, \epsilon) = \sup_{\Theta\in\mathcal{N}_\epsilon(\Theta^0)} \|L(\Theta)\|_2 \le C, \quad \sigma_K(\Theta^0, \epsilon) = \sup_{\Theta\in\mathcal{N}_\epsilon(\Theta^0)} \|K(\Theta)\|_2 \le C.$$

Also define

$$\ell(\Theta^0, \epsilon) = \sup_{\Theta\in\mathcal{N}_\epsilon(\Theta^0)} \max(1, \max_{j\in[r]} \|L_j(\Theta)\|_2)\,.$$

Note that $\ell(\Theta^0, \epsilon) \le \max(C, 1)$, since $\max_{j\in[r]} \|L_j(\Theta)\|_2 \le \|L(\Theta)\|_2$.

Theorem 3.3. *Consider the LQ system of Eq.* (1). *For some constants* $\epsilon, C_{\min}$ *and* $0 < \alpha, \rho < 1$, *assume that an initial* $(\rho, C_{\min}, \alpha)$ *identifiable regulator* $L^{(0)}$ *is given. Further, assume that for any* $\Theta \in \mathcal{N}_\epsilon(\Theta^0)$, $L(\Theta)$ *is* $(\rho, C_{\min}, \alpha)$ *identifiable. Then, with probability at least* $1-\delta$ *the cumulative regret of* ALGORITHM *(cf. the table) is bounded as*

$$R(T) \leq \tilde{O}(p\sqrt{T}\log^{\frac{3}{2}}(1/\delta))\,, \tag{12}$$

where $\tilde{O}$ *is hiding the logarithmic factors.*

4 Analysis

4.1 Proof of Theorem 3.2

To prove theorem 3.2 we first state a set of sufficient conditions for the solution of the ℓ_1-regularized least squares to be within some distance, as defined by $\mathsf{d}(\cdot,\cdot)$, of the true parameter. Subsequently, we prove that these conditions hold with high probability.

Define $X = [x(0), x(1), \ldots, x(n-1)] \in \mathbb{R}^{p\times n}$ and let $W = [w(1), \ldots, w(n)] \in \mathbb{R}^{p\times n}$ be the matrix containing the Gaussian noise realization. Further let the W_u denote the u^{th} row of W.

Define the normalized gradient and Hessian of the likelihood function (6) as

$$\widehat{G} = -\nabla\mathcal{L}(\Theta_u^0) = \frac{1}{n}\widetilde{L}XW_u^{\mathsf{T}}\,, \qquad \widehat{H} = \nabla^2\mathcal{L}(\Theta_u^0) = \frac{1}{n}\widetilde{L}XX^{\mathsf{T}}\widetilde{L}^{\mathsf{T}}\,. \tag{13}$$

The following proposition, a proof of which can be found in [20], provides a set of sufficient conditions for the *accuracy* of the ℓ_1-regularized least squares solution.

Proposition 4.1. *Let* S *be the support of* Θ_u^0 *with* $|S| < k$, *and* H *be defined per Definition 3.1. Assume there exist* $0 < \alpha < 1$ *and* $C_{\min} > 0$ *such that*

$$\lambda_{\min}(H_{S,S}) \geq C_{\min}\,, \qquad \|H_{S^c,S}H_{S,S}^{-1}\|_\infty \leq 1-\alpha\,. \tag{14}$$

For any $0 < \epsilon < \Theta_{\min}$ *if the following conditions hold*

$$\|\widehat{G}\|_\infty \leq \frac{\lambda\alpha}{3}\,, \qquad \|\widehat{G}_S\|_\infty \leq \frac{\epsilon C_{\min}}{4k} - \lambda, \tag{15}$$

$$\|\widehat{H}_{S^cS} - H_{S^cS}\|_\infty \leq \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}\,, \qquad \|\widehat{H}_{SS} - H_{SS}\|_\infty \leq \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}\,, \tag{16}$$

the ℓ_1*-regularized least square solution* (5) *satisfies* $\mathsf{d}(\widehat{\Theta}_u, \Theta_u^0) \leq \epsilon$.

In the sequel, we prove that the conditions in Proposition 4.1 hold with high probability given that the assumptions of Theorem 3.2 are satisfied. A few lemmas are in order proofs of which are deferred to the Appendix.

The first lemma states that $\widehat{G}$ concentrates in infinity norm around its mean of zero.

Lemma 4.2. *Assume* $\rho = \|A^0 - B^0L\|_2 < 1$ *and let* $\ell = \max(1, \max_{i\in[r]}\|L_i\|_2)$. *Then, for any* $S \subseteq [q]$ *and* $0 < \epsilon < \frac{\ell}{2}$

$$\mathbb{P}\big\{\|\widehat{G}_S\|_\infty > \epsilon\big\} \leq 2|S|\,\exp\left(-\frac{n(1-\rho)\epsilon^2}{4\ell^2}\right)\,. \tag{17}$$

To prove the conditions in Eq. (16) we first bound in the following lemma the absolute deviations of the elements of $\widehat{H}$ from their mean H, i.e., $|\widehat{H}_{ij} - H_{ij}|$.

Lemma 4.3. *Let* $i, j \in [q]$, $\rho = \|A^0 - B^0L\|_2 < 1$, *and* $0 < \epsilon < \frac{3}{1-\rho} < n$. *Then,*

$$\mathbb{P}(|\widehat{H}_{ij} - H_{ij}| > \epsilon) \leq 2\exp\left(-\frac{n(1-\rho)^3\epsilon^2}{24\ell^2}\right)\,. \tag{18}$$

The following corollary of Lemma 4.3 bounds $\|\widehat{H}_{JS} - H_{JS}\|_\infty$ for $J, S \subseteq [q]$.

Corollary 4.4. *Let $J, S \subseteq [q]$, $\rho = \|A^0 - B^0 L\|_2 < 1$, $\epsilon < \frac{3|S|}{1-\rho}$, and $n > \frac{3}{1-\rho}$. Then,*

$$\mathbb{P}(\|\widehat{H}_{JS} - H_{JS}\|_\infty > \epsilon) \leq 2|J||S| \exp\left(-\frac{n(1-\rho)^3\epsilon^2}{24|S|^2\ell^2}\right). \tag{19}$$

The proof of Corollary 4.4 is by applying union bound as

$$\mathbb{P}(\|\widehat{H}_{JS} - H_{JS}\|_\infty > \epsilon) \leq |J||S| \max_{i\in J, j\in S} \mathbb{P}(|\widehat{H}_{ij} - H_{ij}| > \epsilon/|S|). \tag{20}$$

Proof of Theorem 3.2. We show that the conditions given by Proposition 4.1 hold. The conditions in Eq. (14) are true by the assumption of identifiability of L with respect to Θ^0. In order to make the first constraint on $\widehat{G}$ imply the second constraint on $\widehat{G}$, we assume that $\lambda\alpha/3 \leq \epsilon C_{\min}/(4k) - \lambda$, which is ensured to hold if $\lambda \leq \epsilon C_{\min}/(6k)$. By Lemma 4.2, $\mathbb{P}(\|\widehat{G}\|_\infty > \lambda\alpha/3) \leq \delta/2$ if

$$\lambda^2 = \frac{36\ell^2}{n(1-\rho)\alpha^2} \log(\frac{4q}{\delta}). \tag{21}$$

Requiring $\lambda \leq \epsilon C_{\min}/(6k)$, we obtain

$$n \geq \frac{36^2\, k^2\ell^2}{\epsilon^2\alpha^2 C_{\min}^2(1-\rho)} \log(\frac{4q}{\delta}). \tag{22}$$

The conditions on $\widehat{H}$ can also be aggregated as $\|\widehat{H}_{[q],S} - H_{[q],S}\|_\infty \leq \alpha C_{\min}/(12\sqrt{k})$. By Corollary 4.4, $\mathbb{P}(\|\widehat{H}_{[q]S} - H_{[q]S}\|_\infty > \alpha C_{\min}/(12\sqrt{k})) \leq \delta/2$ if

$$n \geq \frac{3456\, k^3\ell^2}{\alpha^2(1-\rho)^3 C_{\min}^2} \log(\frac{4kq}{\delta}). \tag{23}$$

Merging the conditions in Eq. (22) and (23) we conclude that the conditions in Proposition 4.1 hold with probability at least $1-\delta$ if

$$n \geq \frac{4\cdot 10^3\, k^2\ell^2}{\alpha^2(1-\rho)C_{\min}^2}\left(\frac{1}{\epsilon^2} + \frac{k}{(1-\rho)^2}\right)\log(\frac{4kq}{\delta}). \tag{24}$$

Which finishes the proof of Theorem 3.2. □

4.2 Proof of Theorem 3.3

The high-level idea of the proof is similar to the proof of main Theorem in [1]. First, we give a decomposition for the gap between the cost obtained by the algorithm and the optimal cost. We then upper bound each term of the decomposition separately.

4.2.1 Cost Decomposition

Writing the Bellman optimality equations [5, 4] for average cost dynamic programming, we get

$$J(\widetilde{\Theta}_t) + x(t)^\mathsf{T} K(\widetilde{\Theta}_t) x(t) = \min_u \left\{ x(t)^\mathsf{T} Q x(t) + u^\mathsf{T} R u + \mathbb{E}\big[z(t+1)^\mathsf{T} K(\widetilde{\Theta}_t) z(t+1) | \mathcal{F}_t\big] \right\},$$

where $\widetilde{\Theta}_t = [\widetilde{A}, \widetilde{B}]$ is the estimate used at time t, $z(t+1) = \widetilde{A}_t x(t) + \widetilde{B}_t u + w(t+1)$, and $\mathcal{F}_t$ is the σ-field generated by the variables $\{(z_\tau, x_\tau)\}_{\tau=0}^t$. Notice that the left-hand side is the average cost occurred with initial state $x(t)$ [5, 4]. Therefore,

$$\begin{aligned}
J(\widetilde{\Theta}_t) + x(t)^\mathsf{T} K(\widetilde{\Theta}_t) x(t) &= x(t)^\mathsf{T} Q x(t) + u(t)^\mathsf{T} R u(t) \\
&\quad + \mathbb{E}\big[(\widetilde{A}_t x(t) + \widetilde{B}_t u(t) + w(t+1))^\mathsf{T} K(\widetilde{\Theta}_t)(\widetilde{A}_t x(t) + \widetilde{B}_t u(t) + w(t+1)) | \mathcal{F}_t\big] \\
&= x(t)^\mathsf{T} Q x(t) + u(t)^\mathsf{T} R u(t) + \mathbb{E}\big[(\widetilde{A}_t x(t) + \widetilde{B}_t u(t))^\mathsf{T} K(\widetilde{\Theta}_t)(\widetilde{A}_t x(t) + \widetilde{B}_t u(t)) | \mathcal{F}_t\big] \\
&\quad + \mathbb{E}\big[w(t+1)^\mathsf{T} K(\widetilde{\Theta}_t) w(t+1) | \mathcal{F}_t\big] \\
&= x(t)^\mathsf{T} Q x(t) + u(t)^\mathsf{T} R u(t) + \mathbb{E}\big[x(t+1)^\mathsf{T} K(\widetilde{\Theta}_t) x(t+1) | \mathcal{F}_t\big] \\
&\quad + \Big((\widetilde{A}_t x(t) + \widetilde{B}_t u(t))^\mathsf{T} K(\widetilde{\Theta}_t)(\widetilde{A}_t x(t) + \widetilde{B}_t u(t)) \\
&\quad - (A^0 x(t) + B^0 u(t))^\mathsf{T} K(\widetilde{\Theta}_t)(A^0 x(t) + B^0 u(t))\Big).
\end{aligned}$$

Consequently

$$\sum_{t=0}^{T}\big(x(t)^\mathsf{T}Qx(t)+u(t)^\mathsf{T}Ru(t)\big)=\sum_{t=0}^{T}J(\widetilde{\Theta}_t)+C_1+C_2+C_3, \tag{25}$$

where

$$C_1=\sum_{t=0}^{T}\Big(x(t)^\mathsf{T}K(\widetilde{\Theta}_t)x(t)-\mathbb{E}\big[x(t+1)^\mathsf{T}K(\widetilde{\Theta}_{t+1})x(t+1)\big|\mathcal{F}_t\big]\Big), \tag{26}$$

$$C_2=-\sum_{t=0}^{T}\mathbb{E}\big[x(t+1)^\mathsf{T}(K(\widetilde{\Theta}_t)-K(\widetilde{\Theta}_{t+1}))x(t+1)\big|\mathcal{F}_t\big], \tag{27}$$

$$\begin{aligned}C_3=-\sum_{t=0}^{T}\Big((\widetilde{A}_tx(t)+\widetilde{B}_tu(t))^\mathsf{T}K(\widetilde{\Theta}_t)(\widetilde{A}_tx(t)+\widetilde{B}_tu(t))\\ -(A^0x(t)+B^0u(t))^\mathsf{T}K(\widetilde{\Theta}_t)(A^0x(t)+B^0u(t))\Big).\end{aligned} \tag{28}$$

4.2.2 Good events

We proceed by defining the following two events in the probability space under which we can bound the terms C_1, C_2, C_3. We then provide a lower bound on the probability of these events.

$$\mathcal{E}_1=\{\Theta^0\in\Omega^{(i)}, \text{ for } i\geq 1\},\quad \mathcal{E}_2=\{\|w(t)\|\leq 2\sqrt{p\log(T/\delta)}, \text{ for } 1\leq t\leq T+1\}.$$

4.2.3 Technical lemmas

The following lemmas establish upper bounds on C_1, C_2, C_3.

Lemma 4.5. *Under the event $\mathcal{E}_1\cap\mathcal{E}_2$, the following holds with probability at least $1-\delta$.*

$$C_1\leq\frac{\sqrt{128}\,C}{(1-\rho)^2}\sqrt{T}\,p\log(\frac{T}{\delta})\sqrt{\log(\frac{1}{\delta})}\,. \tag{29}$$

Lemma 4.6. *Under the event $\mathcal{E}_1\cap\mathcal{E}_2$, the following holds.*

$$C_2\leq\frac{8C}{(1-\rho)^2}p\log(\frac{T}{\delta})\log T\,. \tag{30}$$

Lemma 4.7. *Under the event $\mathcal{E}_1\cap\mathcal{E}_2$, the following holds with probability at least $1-\delta$.*

$$|C_3|\leq 800\Big(\frac{C}{1-\rho}\Big)^{\frac{5}{2}}k\sqrt{\Big(1+\frac{k\epsilon^2}{(1-\rho)^2}\Big)}\cdot\frac{1+C}{C_{\min}}\cdot\log(\frac{pT}{\delta})\sqrt{\log(\frac{4kq}{\delta})}\cdot p\log T\sqrt{T}\,. \tag{31}$$

Lemma 4.8. *The following holds true.*

$$\mathbb{P}(\mathcal{E}_1)\geq 1-\delta,\quad \mathbb{P}(\mathcal{E}_2)\geq 1-\delta. \tag{32}$$

Therefore, $\mathbb{P}(\mathcal{E}_1\cap\mathcal{E}_2)\geq 1-2\delta$.

We are now in position to prove Theorem 3.3.

Proof (Theorem 3.3). Using cost decomposition (Eq. (25)), under the event $\mathcal{E}_1\cap\mathcal{E}_2$, we have

$$\begin{aligned}\sum_{t=0}^{T}(x(t)^\mathsf{T}Qx(t)+u(t)^\mathsf{T}Ru(t))&=\sum_{t=0}^{T}J(\widetilde{\Theta}_t)+C_1+C_2+C_3\\ &\leq TJ(\Theta^0)+C_1+C_2+C_3,\end{aligned}$$

where the last inequality stems from the choice of $\widetilde{\Theta}_t$ by the algorithm (cf. Eq (9)) and the fact that $\Theta^0\in\Omega_t$, for all t under the event $\mathcal{E}_1$. Hence, $R(T)\leq C_1+C_2+C_3$. Now using the bounds on C_1, C_2, C_3, we get the desired result. □

Acknowledgments

The authors thank the anonymous reviewers for their insightful comments. A.J. is supported by a Caroline and Fabian Pease Stanford Graduate Fellowship.

References

[1] Y. Abbasi-Yadkori and C. Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. *Proceeding of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

[2] Y. Bar-Shalom and E. Tse. Dual effect, certainty equivalence, and separation in stochastic control. *Automatic Control, IEEE Transactions on*, 19(5):494–500, 1974.

[3] J. Bento, M. Ibrahimi, and A. Montanari. Learning networks of stochastic differential equations. *Advances in Neural Information Processing Systems 23*, pages 172–180, 2010.

[4] D. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, 1987.

[5] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 3rd edition, 2007.

[6] S. Bittanti and M. Campi. Adaptive control of linear time invariant systems: the bet on the best principle. *Communications in Information and Systems*, 6(4):299–320, 2006.

[7] E. Bradlow, B. Bronnenberg, G. Russell, N. Arora, D. Bell, S. Duvvuri, F. Hofstede, C. Sismeiro, R. Thomadsen, and S. Yang. Spatial models in marketing. *Marketing Letters*, 16(3):267–278, 2005.

[8] M. Campi. Achieving optimality in adaptive control: the bet on the best approach. In *Decision and Control, 1997., Proceedings of the 36th IEEE Conference on*, volume 5, pages 4671–4676. IEEE, 1997.

[9] V. Dani, T. Hayes, and S. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

[10] G. Feichtinger, R. Hartl, and S. Sethi. Dynamic optimal control models in advertising: recent developments. *Management Science*, pages 195–226, 1994.

[11] L. Guo and H. Chen. The åstrom-wittenmark self-tuning regulator revisited and els-based adaptive trackers. *Automatic Control, IEEE Transactions on*, 36(7):802–812, 1991.

[12] P. Kumar and A. Becker. A new family of optimal adaptive controllers for markov chains. *Automatic Control, IEEE Transactions on*, 27(1):137–146, 1982.

[13] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[14] T. Lai and C. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

[15] C. Marinelli and S. Savin. Optimal distributed dynamic advertising. *Journal of Optimization Theory and Applications*, 137(3):569–591, 2008.

[16] T. Seidman, S. Sethi, and N. Derzko. Dynamics and optimization of a distributed sales-advertising model. *Journal of Optimization Theory and Applications*, 52(3):443–462, 1987.

[17] S. Sethi. Dynamic optimal control models in advertising: a survey. *SIAM review*, pages 685–725, 1977.

[18] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3):1030–1051, 2006.

[19] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

[20] P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.