Near-optimal Rank Adaptive Inference of High Dimensional Matrices

Frédéric Zheng
KTH
Stockholm, Sweden

Yassir Jedra
MIT
Cambridge, MA, USA

Alexandre Proutière
KTH
Stockholm, Sweden

Abstract

We address the problem of estimating a high-dimensional matrix from linear measurements, with a focus on designing optimal rank-adaptive algorithms. These algorithms infer the matrix by estimating its singular values and the corresponding singular vectors up to an effective rank, adaptively determined based on the data. We establish, for the first time, instance-specific lower bounds for the sample complexity of such algorithms. We uncover fundamental trade-offs in selecting the effective rank: balancing the precision of estimating a subset of singular values against the approximation cost incurred for the remaining ones. Our analysis identifies how the optimal effective rank depends on the matrix being estimated, the sample size, and the noise level. We propose an algorithm that combines a Least-Squares estimator with a universal singular value thresholding procedure. We provide finite-sample error bounds for this algorithm, that are tighter than those of existing rank-adaptive algorithms. Furthermore, our bounds nearly match the derived fundamental limits. Finally, we confirm experimentally that our algorithm outperforms existing rank-adaptive algorithms.

1 INTRODUCTION

We revisit the canonical problem of estimating a high-dimensional matrix from linear measurements. The learner has access to n samples, $(x_1, y_1), \dots, (x_n, y_n)$ where $y_i \in \mathbb{R}^{d_y}$ is a noisy realization of Ax_i with $x_i \in \mathbb{R}^{d_x}$. $A \in \mathbb{R}^{d_y \times d_x}$ is considered a priori unknown. The objective is to estimate the matrix A as accurately as possible, and more precisely to construct an estimator $\hat{A}_n$ with minimal Frobenius error $\|\hat{A}_n - A\|_{\mathrm{F}}$ with high probability. This problem has widespread applications across various domains, including healthcare, biology, computer vision, and control (see Wu et al. (2020) for a non-exhaustive list of examples). In this paper, we illustrate our results through two key applications: (1) multivariate regression where the covariates $(x_i)_{i\in\{1,\dots,n\}}$ form an i.i.d. sequence of random variables and (2) linear system identification where the covariates are the successive states of a linear time-invariant dynamical system governed by A, meaning that x_{i+1} is a noisy version of Ax_i.

The high-dimensional setting arises when the number of entries of A are comparable to or exceed the number of available observations. In such cases, imposing and leveraging additional structural properties — such as sparsity or low rank — is essential for statistically sound matrix estimation. Ideally, an estimation algorithm should automatically detect whether such a structure exists and construct an estimator accordingly. In this work, we focus on scenarios where the relevant structure is the matrix rank and aim to design optimal rank-adaptive algorithms. These algorithms estimate the matrix by inferring its singular values and corresponding singular vectors up to an effective rank, which is adaptively determined based on the data. The key questions we explore are: What are the fundamental performance limits of such algorithms? Can we devise an algorithm approaching these limits? How does the optimal effective rank depend on the matrix being estimated, the covariates, and the sample size? We answer these questions with, to the best of our knowledge, an unprecedented level of precision. Specifically, our contributions are as follows.

1. Instance-specific Sample Complexity Lower bounds. We establish the first instance-specific lower bounds for

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

the sample complexity[1] of rank-adaptive algorithms. These bounds reveal fundamental trade-offs in selecting the effective rank and precisely depend on both the matrix A and the covariates' distribution. For instance, in the case of multivariate regression, our results show that with n samples and with $d_x = d_y$, the Frobenius norm error scales at least as:

$$\min_k \left(\sigma^2 \frac{\log(\frac{1}{\delta}) + kd_x}{n\underline{\lambda}_k(\Sigma)} + \sum_{i>k} s_i^2(A) \right),$$

where σ is the noise level, $s_i(A)$ denotes the i-th largest singular value of A, Σ is the covariance matrix of covariates, and $\underline{\lambda}_k(\Sigma)$ is the average of its k smallest eigenvalues. Furthermore, if we were able to learn the value of k that minimizes this bound, we would directly obtain the optimal effective rank. The proof techniques towards our lower bounds are of independent interest. They intricately combine change-of-measure arguments, which provide explicit dependence on A and the covariates, with packing arguments, which capture the dependence on the dimensions d_x and d_y.

2. Thresholded Least-squares Estimation is nearly optimal. We propose an algorithm that combines a least squares estimator with a universal singular value thresholding procedure. We derive finite-sample error bounds for this estimator and show that it outperforms existing algorithms, while closely approaching the fundamental performance limits. Our analysis builds on an improved understanding of matrix denoising techniques based on singular value thresholding. Specifically, we demonstrate that applying universal singular value thresholding to an existing estimator generally enhances its performance guarantees. We illustrate this principle using the estimator obtained via nuclear norm regularization and the Least-Squares estimator.

3. Applications and Numerical Results. We exemplify our theoretical results to both multivariate regression and linear system identification. Finally, we present numerical experiments to complement and validate our theoretical findings.

Notation. We use $a \lesssim b$ (resp. $a \gtrsim b$) to mean that a is smaller (resp. larger) than b up to a universal multiplicative constant. We use $a \wedge b$ (resp. $a \vee b$) to denote $\min(a, b)$ (resp. $\max(a, b)$). Let $[d] = \{1, ..., d\}$. Given a matrix $M \in \mathbb{R}^{d_x \times d_y}$, $\|M\|_{\mathrm{F}}$ denotes its Frobenius norm, $\|M\|_2$ denotes its operator norm, and $\|M\|_1$ denotes its nuclear norm. Let $\underline{d} = d_x \wedge d_y$ and $\bar{d} = d_x \vee d_y$. The singular values (resp. eigenvalues) of M are denoted by $s_1(M), \ldots, s_{\underline{d}}(M)$ (resp. $\lambda_1(M), \ldots, \lambda_{\underline{d}}(M)$) in decreasing order, and its condition number is denoted by $\kappa(M) = s_1(M)/s_d(M)$. For $k \le \underline{d}$, we use $\Pi_k(M)$ to denote the best rank-k approximation of M. $M^\dagger$ is the Moore-Penrose pseudo-inverse of M. When M is symmetric positive semi-definite, we denote its maximum (resp. minimum) eigenvalue by $\lambda_{\max}(M)$ (resp. $\lambda_{\min}(M)$). We also define $\bar{\lambda}_k(M) := \sum_{i=1}^k \lambda_i(M)/k$ and $\underline{\lambda}_k(M) := \sum_{i=0}^{k-1} \lambda_{d-i}(M)/k$.

[1] By sample complexity here, we mean the minimum number of samples required for the existence of an (ε, δ)-PAC rank-adaptive algorithm (i.e., returning an estimator $\hat{A}_n$ such that $\|\hat{A}_n - A\|_F \le \varepsilon$ with probability at least $1-\delta$.)

2 PRELIMINARIES

Model and objective. We are given n observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, taking values in $\mathbb{R}^{d_x \times d_y}$. We consider a linear model whereby $y_i = Ax_i + \eta_i$ for all $i \in [n] = \{1, \ldots, n\}$. A is a matrix in $\mathbb{R}^{d_y \times d_x}$ that is a priori unknown. $(\eta_i)_{i \ge 1}$ is a sequence of i.i.d. zero-mean, σ^2-sub-Gaussian random variables taking values in $\mathbb{R}^{d_y}$ where $\sigma > 0$ denotes their variance proxy parameter. Our objective is to estimate the matrix A.

We define the empirical covariance matrix as $\hat{\Sigma} = (\sum_{i=1}^n x_i x_i^\top)/n$. We assume that $\mathbb{E}[\hat{\Sigma}] \succ 0$, ensuring identifiability of A by least squares. We are interested in two specific settings where this assumption naturally holds:

1. Multivariate regression. Here the covariates $x_1, \ldots, x_n$ are assumed to be i.i.d. random vectors with covariance matrix $\Sigma := \mathbb{E}[\hat{\Sigma}] = \mathbb{E}[x_1 x_1^\top]$.

2. Linear system identification. Here, $y_i = x_{i+1}$ for all $i \ge 1$, and thus the covariates $x_1, \ldots, x_n$ evolve according to the dynamical system: $x_{t+1} = Ax_t + \eta_t$ for all $t \ge 1$, with $x_1 = 0$. In this case, $d_x = d_y$. For all $t \ge 1$, the finite-time controllability Gramian of the system as $\Gamma_p(A) := \sum_{k=0}^p (A^k)(A^k)^\top$. We assume that the system is stable (i.e., $\rho(A) < 1$, where $\rho(A)$ denotes the spectral radius of A). Thus, we may define $\Gamma_\infty(A)$ as $\lim_{t \to \infty} \Gamma_t(A)$.

In what follows, we will often prefer to use matrix notations and write $Y = XA^\top + E$ where $Y^\top = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix} \in \mathbb{R}^{d_y \times n}$, $X^\top = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \in \mathbb{R}^{d_x \times n}$ and $E^\top = \begin{bmatrix} \eta_1 & \cdots & \eta_n \end{bmatrix} \in \mathbb{R}^{d_y \times n}$.

3 RELATED WORK

The inference of high-dimensional matrices using reduced-rank regression has a long history and has attracted significant attention over the past decades (see, e.g., Anderson (1951); Izenman (1975); Reinsel and Velu (1998); Anderson (1999); Negahban and Wainwright (2011); Bunea et al. (2010)). It also has numerous connections to iconic problems such as principal component analysis and matrix completion Candès

and Recht (2012); Koltchinskii et al. (2011); Chatterjee (2015). Below, we present a selection of papers that we consider most relevant to our analysis. We begin by discussing existing fundamental limits before moving on to algorithms with finite sample-size performance guarantees.

Fundamental limits. For multivariate regression and matrix completion, most existing lower bounds on the estimation error are minimax and concern matrices with known rank (or with a known upper bound on its rank). For instance, the lower bounds derived in Rohde and Tsybakov (2009) are minimax and assume that the design matrix X satisfies some properties such as the Restricted Isometry Property (RIP), and that the rank is fixed. Candes and Plan (2011) derived lower bounds satisfied when the design matrix X has the RIP property, and that remain valid for approximately low-rank matrices. However, these bounds are in expectation, and do not cover fully adaptive algorithms. The lower bounds presented in Bunea et al. (2010) are based on those from Rohde and Tsybakov (2009); Candes and Plan (2011) and hence suffer from the same shortcomings. Cai et al. (2015) considers the problem of optimal rank estimation for covariate matrices, and presents minimax rate detection limits, but these are only valid for exactly low-rank matrices. Closer to our analysis, Wu et al. (2020) proposes an instance-specific (or more precisely locally minimax) lower bound, but unfortunately these are only valid for the risk and not the matrix estimation error. Finally, it is worth noting recent efforts towards the derivation of error lower bounds for the problem of linear system identification Simchowitz et al. (2018); Jedra and Proutiere (2019, 2022); Djehiche and Mazhar (2021); Sun et al. (2023); Bakshi et al. (2023); Zhang et al. (2024). These bounds concern the estimation error in operator norm, they are not rank-adaptive, and most of them are minimax.

To the best of our knowledge, we are the first to derive precise instance-specific lower bounds that depend on the spectral properties (i.e., the singular values) of both the target matrix and the design matrix. This contrasts with traditional minimax lower bounds, which characterize performance limits for the worst-case matrix within a given class. As a result, minimax lower bounds fail to capture how algorithms can truly adapt to the specific matrix being estimated.

Algorithms and their finite-sample guarantees. Most existing algorithms for reduced-rank regression rely on optimization methods that incorporate a penalty in the objective function to encourage low-rank solutions. Common choices for this penalty include the nuclear norm or its weighted variant Yuan et al. (2007); Candes and Plan (2010); Negahban and Wainwright (2011); Chen et al. (2013), the rank itself Bunea et al. (2010), and the Schatten-p quasi-norm Rohde and Tsybakov (2009). When the rank of A is at most r, the most effective among these algorithms is the Rank-Constrained Selection (`RSC`) algorithm introduced by Bunea et al. (2010), and whose error bounds are analyzed in their Corollary 8: given n samples, with probability at least $1-\delta$,

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \lesssim \min_{k\in[r]} \left(k\sigma^2 \frac{\log(\frac{1}{\delta}) + \bar{d}}{n\lambda_{\min}(\hat{\Sigma})} + \kappa(\hat{\Sigma}) \sum_{i>k} s_i^2(A) \right). \tag{1}$$

The RSC algorithm is meant for minimizing the prediction error $\|X\hat{A}_n - XA\|_{\mathrm{F}}$. This explains the extra dependence on the condition number $\kappa(\hat{\Sigma})$ in the bound on the identification error $\|\hat{A}_n - A\|_{\mathrm{F}}$. Our algorithm, the Thresholded Least Squares Estimator (`T-LSE`), achieves provably better error bounds. It combines the `LSE` with a universal singular value thresholding procedure, inspired by the seminal works of Chatterjee (2015); Gavish and Donoho (2014). Finally, the design and analysis of algorithms with finite sample-size performance guarantees for identifying dynamical systems with low-rank structure, notably partially observed linear dynamical system, has also seen a surge of interest recently Fazel et al. (2013); Djehiche and Mazhar (2022); Sun et al. (2022); Oymak and Ozay (2019); Simchowitz et al. (2019); Sarkar et al. (2021); Bakshi et al. (2023). However, to the best of our knowledge, we are the first to present rank-adaptive algorithms with guarantees, exploiting the low-rank structure of the state matrix.

4 INSTANCE-SPECIFIC SAMPLE COMPLEXITY LOWER BOUNDS

Our goal in this section is to derive fundamental limits on the number of samples required to obtain an (ε, δ)-PAC estimate of the matrix A. To this end, we must focus on estimators or algorithms that genuinely adapt to the matrix A. For instance, an algorithm that always outputs A, regardless of the input data, would require no samples to be (ε, δ)-PAC when the true matrix is A, but would fail for any matrix other than A. Therefore, to derive instance-specific sample complexity lower bounds, we must consider algorithms that are (ε, δ)-PAC not just for A, but for all matrices in a *neighborhood* of A. The choice of this neighborhood involves a delicate trade-off. On the one hand, the neighborhood around A should be small enough to ensure that the class of algorithms considered is broad enough and that the derived lower bound reflects the difficulty of estimating the specific instance A. On the other hand, if the neighborhood is too narrow, the resulting lower bound may be overly restrictive and potentially unattainable.

Deriving lower bounds on the sample complexity of *rank-adaptive* algorithms presents significant challenges. We address this by decomposing the problem into two steps: (i) we first derive lower bounds (see §4.1) using a neighborhood of A obtained by adding perturbations of specific rank $r \geq \mathrm{rank}(A)$. As shown later, this lower bound will be tight for *rank-constrained* algorithms, i.e., those returning matrices of rank r. (ii) Building on this analysis, we then derive refined lower bounds (see §4.2) by considering a neighborhood of A obtained by adding perturbations of rank bounded by the effective rank that optimally balances estimation and approximation errors. In Section 6, we introduce a rank-adaptive algorithm whose sample complexity nearly matches these lower bounds (without any prior knowledge), hence proving that our lower bounds are tight.

Throughout this section, we assume that $(\eta_i)_{i\in[n]}$ are distributed according to $\mathcal{N}(0, \sigma^2 I_{d_y})$. Proofs of the results presented in this section are in Appendix A.

4.1 Lower bounds for rank-constrained algorithms

As a starting point, we derive fundamental limits for algorithms estimating a matrix A by a rank-r matrix. We begin by presenting the following packing bounds for Stiefel manifolds, which will play an important role in the definition of the neighborhood of A. For any $k \leq d$, the Stiefel manifold $\mathrm{St}_k^d(\mathbb{R})$ is defined as the set of semi-orthogonal matrices $\{Q \in \mathbb{R}^{d\times k} : \ Q^\top Q = I_k\}$.

Lemma 4.1. *For any $k \leq d/2$, there exists a finite $\sqrt{k}$-packing $\mathcal{P}_k^d$ of $\mathrm{St}_k^d(\mathbb{R})$ whose cardinality is larger than 2^{kd}. By $\sqrt{k}$-packing, we mean that for any $Q, R \in \mathcal{P}_k^d$, $Q \neq R$ implies $\|Q - R\|_{\mathrm{F}} \geq \sqrt{k}/C$ where $C \geq 1$ is a universal constant independent of k and d.*

The explicit value of C is given in Appendix A. We define the neighborhood of A as: $\mathcal{C}(A, r, \varepsilon) = \mathcal{C}_1(A, r, \varepsilon) \cup \mathcal{C}_2(A, r, \varepsilon) \cup \{A\}$ where

$$\begin{cases} \mathcal{C}_1(A, r, \varepsilon) & = \left\{A + \frac{2C\varepsilon}{\sqrt{r}} Q W_{-r}^\top : Q \in \mathcal{P}_r^{d_y}\right\} \\ \mathcal{C}_2(A, r, \varepsilon) & = \left\{A + \frac{2C\varepsilon}{\sqrt{r}} U_r R^\top : R \in \mathcal{P}_r^{d_x}\right\} \end{cases}. \quad (2)$$

Here, $U_r \in \mathrm{St}_r^{d_y}(\mathbb{R})$ is a semi-orthogonal matrix whose columns $u_1, ..., u_r$ are the left-singular vectors of A corresponding to $s_1(A), ..., s_r(A)$ while $W_{-r} \in \mathrm{St}_r^{d_x}(\mathbb{R})$ contains the eigenvectors corresponding to the r smallest eigenvalues of 1) Σ for multivariate regression, 2) $\Gamma_\infty(A)$ for dynamical systems. Our packing directly uses knowledge of the expected covariance matrix, avoiding the need for RIP assumption. We show below that this leads to a more precise dependency of the sample complexity with respect to the spectrum of Σ.

Definition 4.2 ((ε, δ, r)-stability)**.** *Let A such that $\mathrm{rank}(A) \leq r$. An algorithm is (ε, δ, r)-stable in A if it returns a rank-r matrix and the following assertion holds:*

$$\exists N \in \mathbb{N}, \ \forall n \geq N, \ \forall A' \in \mathcal{C}(A, r, \varepsilon), \\ \mathbb{P}_{A'}(\|\hat{A}_n - A'\|_{\mathrm{F}} \leq \varepsilon) \geq 1 - \delta. \quad (3)$$

If an algorithm is (ε, δ, r)-stable in A, we define its *sample complexity* as the minimal integer N such that Assertion (3) holds. We tried with the above definition to construct a neighborhood $\mathcal{C}(A, r, \varepsilon)$ as small as possible. In fact, it is finite, which implies that the class of (ε, r, δ)-stable algorithms in A is very broad. In particular, it includes all algorithms returning rank-r matrices and that are *consistent* in the sense that for any A, there exists a number of sample N_A such that $\mathbb{P}_A(\|\hat{A}_{N_A} - A\|_{\mathrm{F}} \leq \varepsilon) \geq 1 - \delta$. $\mathcal{C}(A, r, \varepsilon)$ is also not too small: we will show that there exist algorithms aware of r and whose sample complexity nearly matches the lower bound obtained by considering this neighborhood.

The following theorem provides sample complexity lower bounds in both the multivariate regression and the linear system identification settings.

Theorem 4.3. *Let $\varepsilon > 0$ and $\delta \in (0, 1)$. Suppose that $\mathbb{E}[\hat{\Sigma}] \succ 0$ and $\mathrm{rank}(A) \leq r \leq \underline{d}/2$. Then:*
(i) Multivariate regression. The sample complexity N of any (ε, δ, r)-stable algorithm in A satisfies:

$$N \gtrsim \frac{\sigma^2}{\varepsilon^2} \left(\frac{r d_x + \log(\frac{1}{\delta})}{\bar{\lambda}_r(\Sigma)} \vee \frac{r d_y + \log(\frac{1}{\delta})}{\underline{\lambda}_r(\Sigma)} \right). \quad (4)$$

(ii) Linear system identification. Further assume that $\varepsilon \leq \|\Gamma_\infty(A)\|_2^{-3}/12$. The sample complexity N of any (ε, δ, r)-stable algorithm in A satisfies:

$$N \gtrsim \sigma^2 \frac{r d_x + \log(\frac{1}{\delta})}{\varepsilon^2 \underline{\lambda}_r(\Gamma_\infty(A))}. \quad (5)$$

4.2 Lower bounds for rank-adaptive algorithms

We now derive a refined lower bound for rank-adaptive algorithms. To this aim, we define a neighborhood obtained by perturbing the matrix A using matrices whose ranks are smaller than the optimal effective rank. We begin by defining the latter. Using the results derived for rank-constrained algorithms, we know that for a given selected rank k and given n samples, the minimal Frobenius estimation error for the estimation

of $\Pi_k(A)$ should behave as, in the multivariate case,

$$\mathrm{ErrReg}(k,n,\Sigma) := \sigma^2\left(\frac{kd_x+\log(\frac{1}{\delta})}{n\bar{\lambda}_k(\Sigma)} \vee \frac{kd_y+\log(\frac{1}{\delta})}{n\underline{\lambda}_k(\Sigma)}\right) \tag{6}$$

and in the system identification case

$$\mathrm{ErrLti}(k,n,\Gamma_\infty(A)) := \sigma^2\frac{kd_x+\log(\frac{1}{\delta})}{n\underline{\lambda}_k(\Gamma_\infty(A))}. \tag{7}$$

Since we ignore the remaining singular values of A, this error square must be increased by $\sum_{i>k} s_i^2(A)$. We can hence guess that the optimal effective rank is $k^\star_{A,n}$ solving in the multivariate case (and similarly in the system identification case)

$$k^\star_{A,n} := \arg\min_{k\in[r]}\left(\mathrm{ErrReg}(k,n,\Sigma)+\sum_{i>k}s_i^2(A)\right). \tag{8}$$

We are now ready to define a notion of stability for rank-adaptive algorithms.

Definition 4.4 ((ε,δ)-stability)**.** *For any integer $m \geq 1$, define $\mathcal{D}(A,m,\varepsilon) = \mathcal{C}(A,k^\star_{A,m},\varepsilon)$. An algorithm is (ε,δ)-stable in A if the following assertion holds:*

$$\exists N\in\mathbb{N} \quad \text{s.t} \quad \begin{cases} (a)\ \|A-\Pi_{k^\star_{A,N}}(A)\|_F \leq \varepsilon \\ (b)\ \forall n\geq N,\ \forall A'\in\mathcal{D}(A,N,\varepsilon), \\ \quad \mathbb{P}_{A'}(\|\hat{A}_n - A'\|_{\mathrm{F}} \leq \varepsilon) \geq 1-\delta. \end{cases} \tag{9}$$

If an algorithm is (ε,δ)-stable in A, we define its *sample complexity* as the minimal integer N such that Assertion (9) holds. Observe that the definition of (ε,δ)-stability involves a neighborhood $\mathcal{D}(A,N,\varepsilon)$ of A that depends on the number of samples N. Such a dependence is essential because we wish to analyze algorithms that adaptively select the effective rank as a function of N. The class of algorithms that are (ε,δ)-stable in A is broad and includes all consistent algorithms.

Theorem 4.5. *Let $\varepsilon > 0$ and $\delta \in (0,1)$. Suppose that $\mathbb{E}[\hat{\Sigma}] \succ 0$ and $\mathrm{rank}(A) \leq r \leq \underline{d}/2$. Then:*
(i) Multivariate regression. The sample complexity N of any (ε,δ)-stable algorithm in A satisfies:

$$N \geq \min\left\{n : \gamma_A^\delta(n) \leq \frac{32}{\log(2)}\varepsilon^2\right\},$$

where we define for any $n \geq 1$,

$$\gamma_A^\delta(n) := \min_{k\in[r]}\left(\mathrm{ErrReg}(k,n,\Sigma)+\sum_{i>k}s_i^2(A)\right). \tag{10}$$

(ii) Linear system identification. Further assume that $\varepsilon \leq \|\Gamma_\infty(A)\|_2^{-3}/12$. The sample complexity N of any (ε,δ)-stable algorithm in A satisfies:

$$N \geq \min\{n : \beta_A^\delta(n) \leq \frac{640}{\log(2)}\varepsilon^2\},$$

where we define for any $n \geq 1$,

$$\beta_A^\delta(n) := \min_{k\in[r]}\left(\mathrm{ErrLti}(k,n,\Gamma_\infty(A))+\sum_{i>k}s_i^2(A)\right). \tag{11}$$

5 MATRIX DENOISING VIA SINGULAR VALUE THRESHOLDING

A key component of our algorithms is an adaptive singular value thresholding procedure that can be combined with any estimator of A. Specifically, let $\bar{A}$ be a given estimate of A, and denote $Z = \bar{A} - A$ its estimation error. Given a threshold parameter $\xi > 0$, the procedure runs an SVD decomposition to obtain $\bar{A} = \sum_{i=1}^{\underline{d}} s_i(\bar{A})u_iv_i^\top$, then produces

$$\bar{A}(\xi) = \sum_{i=1}^{\underline{d}} \mathbb{1}_{\{s_i(\bar{A})>\xi\}} s_i(\bar{A})u_iv_i^\top.$$

Singular value thresholding with tighter guarantees. In his seminal work Chatterjee (2015), Chatterjee proposes a universal way of selecting the threshold ξ and provides a performance analysis of the resulting algorithm. His analysis builds on a crucial result (Lemma 3.5 in Chatterjee (2015)) stating that for $\tau > 0$ and $\xi = (1+\tau)\|Z\|_2$, it must hold that $\|\bar{A}(\xi) - A\|_{\mathrm{F}}^2 \leq f(\tau)\|Z\|_2\|A\|_1$, where $f(\tau) = ((4+2\tau)\sqrt{2/\tau}+\sqrt{2+\tau})^2$. This upper bound appears conservative as it depends on the nuclear norm of A, and consequently on all the singular values of A. Furthermore, observe that when $\tau \to \infty$ then the upper bound also goes to infinity while it is clear that $\|\bar{A}(\xi) - A\|_{\mathrm{F}}^2 \to \|A\|_{\mathrm{F}}^2$. Ideally, we seek an error upper bound that is independent of the singular values of A exceeding the threshold ξ. In the following theorem, we derive an error upper bound satisfying this desired property.

Theorem 5.1. *For $\xi \geq 2\|Z\|_2$, we have:*

$$\|\bar{A}(\xi) - A\|_{\mathrm{F}}^2 \leq 18\min_{k\in[r]}\left(4k\xi^2+\sum_{i>k}s_i^2(A)\right).$$

The theorem has a significant implication for the design of rank-adaptive algorithms. Suppose that we can construct an initial estimate $\bar{A}$ of A and derive a concentration result for its error $Z = \bar{A} - A$. For instance, assume that we can upper bound $2\|Z\|_2$ by

ξ with high probability. In this case, a rank-adaptive algorithm that outputs $\bar{A}(\xi)$, i.e., with effective rank $k = \max\{i : s_i(\bar{A}) > \xi\}$, is expected to perform well. Theorem 5.1 is a consequence of the following decomposition lemma.

Lemma 5.2. *For $k \in [\underline{d}]$, we have:*

$$\|\Pi_k(\bar{A}) - A\|_{\mathrm{F}} \leq 2\sqrt{2}\|\Pi_k(Z)\|_{\mathrm{F}} + 3\|A - \Pi_k(A)\|_{\mathrm{F}}.$$

We can immediately deduce from this lemma that $\|\Pi_k(\bar{A}) - A\|_{\mathrm{F}}^2 \leq 18(k\|Z\|_2^2 + \sum_{i>k} s_i^2(A))$ for $k \leq \underline{d}$. The proofs of Theorem 5.1 and Lemma 5.2 are given in Appendix B.

Improved guarantees for estimators with nuclear norm penalization. To better appreciate Theorem 5.1, we provide in Appendix B, as an example, an improved analysis of an estimator with nuclear norm penalization, previously analyzed in Negahban and Wainwright (2011). For this estimator, an adaptive error upper bound was given in Theorem 12 in Bunea et al. (2010). We show in Corollary B.1 that our thresholding yields a smaller upper bound, by a multiplicative factor $\kappa(\Sigma)$.

6 THE THRESHOLDED LEAST SQUARES ESTIMATOR

In this section, we describe how Least Squares Estimator (`LSE`) can be combined with singular value thresholding to develop *rank-constrained* and *rank-adaptive* algorithms, for both the multivariate regression and linear system identification tasks. We provide performance guarantees for these algorithms. The proofs are deferred to Appendix C. Before proceeding, we define the `LSE` as $\bar{A} \in \arg\min_{A \in \mathbb{R}^{d_y \times d_x}} \sum_{i=1}^n \|y_i - Ax_i\|_2^2$. We can also express it in its closed form as follows $\bar{A} = (Y^\top X)(X^\top X)^\dagger$. We assume $n \geq d_x$, ensuring a nonzero probability that X is full rank and A is identifiable. This is standard in the literature (e.g Wainwright (2019)), because even for learning a rank-one matrix, observing at least d_x samples is necessary. In the multivariate regression case, we restrict our attention to Gaussian inputs for ease of exposition. However, we prove general versions of Theorems 6.2 and 6.3 for any $\hat{\Sigma}$ in Appendix C, and discuss when one can replace $\hat{\Sigma}$ by Σ with recent concentration results obtained by Barzilai and Shamir (2024).

6.1 Multivariate regression

Rank-constrained LSE (`R-LSE`). Here, we consider the case where the rank of A is known to be upper bounded by r, thus the algorithm must set the effective rank equal to r. We define the `R-LSE` as $\hat{A}_n := \Pi_r(\bar{A})$. To analyze the performance of `R-LSE` we first establish the concentration result below.

Lemma 6.1. *Let $\delta \in (0,1)$. For all $k \in [\underline{d}]$, with probability at least $1-\delta$, it holds that:*

$$\|\Pi_k(Z)\|_{\mathrm{F}}^2 \leq \frac{k\sigma^2\left(\sqrt{d_x} + \sqrt{d_y} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)^2}{n\underline{\lambda}_k^H(\hat{\Sigma})}, \quad (12)$$

where $\underline{\lambda}_k^H(\hat{\Sigma}) := \left(\frac{1}{k}\sum_{i=d_x-k+1}^{d_x} \frac{1}{\lambda_i(\hat{\Sigma})}\right)^{-1}$ is the harmonic mean of the k smallest eigenvalues of $\hat{\Sigma}$.

Combining this concentration result to Lemma 5.2, we immediately obtain:

Theorem 6.2. *Assume that $\mathrm{rank}(A) \leq r$, and $x_i \sim \mathcal{N}(0,\Sigma)$ with $\Sigma \succ 0$. For $n \gtrsim d_x + \log(\frac{1}{\delta})$, the `R-LSE` satisfies, with probability at least $1-\delta$:*

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \lesssim r\sigma^2 \frac{\bar{d} + \log(\frac{1}{\delta})}{n\underline{\lambda}_r^H(\Sigma)}.$$

The upper bound provided above for the `R-LSE` is tighter than that of the best-known rank constrained algorithm found in Corollary 6 of Bunea et al. (2010), by a factor $\underline{\lambda}_r^H(\Sigma)/\lambda_{\min}(\Sigma)$.

Since `R-LSE` is consistent and returns a rank-r matrix, it is (ε, δ, r)-stable for any $\epsilon > 0$ and $\delta \in (0,1)$. Its sample complexity verifies, see Appendix D,

$$N \lesssim r\sigma^2 \frac{\bar{d} + \log(\frac{1}{\delta})}{\varepsilon^2 \underline{\lambda}_r^H(\Sigma)}. \quad (13)$$

While the sample complexity upper bound provided for `R-LSE` exhibits a dependence on r, d_x, d_y, n and σ matching those identified in the lower bound result of Theorem 4.3, it still does not capture the right dependence on Σ. To fully match the lower bound, the term $\underline{\lambda}_r^H(\Sigma)$ should be replaced by either $\underline{\lambda}_r(\Sigma)$ if $\bar{\lambda}_r(\Sigma)(\log(1/\delta) + rd_y) \geq \underline{\lambda}_r(\Sigma)(\log(1/\delta) + rd_x)$ or $\bar{\lambda}_r(\Sigma)$ otherwise. However, for well-conditioned Σ, our bounds are tight. Furthermore, due to our application of Lemma 6.1, we obtain $r\log(\frac{1}{\delta})$ compared to $\log(\frac{1}{\delta})$ in our lower bound. We argue that this is negligible in the low-rank settings where $r = O(1)$.

Thresholded LSE (`T-LSE`). We now consider rank-adaptive algorithms. One can show that the error $Z = \bar{A} - A$ of the LSE satisfies the following concentration result (see Lemma 3 in Bunea et al. (2010)): with probability at least $1-\delta$,

$$2\|Z\|_2 \leq \xi_{\mathrm{MR}} := \frac{2\sigma\left(\sqrt{d_x} + \sqrt{d_y} + \sqrt{\log(\frac{1}{\delta})}\right)}{\sqrt{n\lambda_{\min}(\hat{\Sigma})}}. \quad (14)$$

Combining this concentration bound with Theorem 5.1, we obtain the following result.

Theorem 6.3. *Assume that $x_i \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. For $n \gtrsim d_x + \log(\frac{1}{\delta})$, the* `T-LSE` *$\hat{A}_n := \bar{A}(\xi_{\mathrm{MR}})$ satisfies, with probability at least $1-\delta$:*

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \lesssim \min_{k \in [r]} \left(k\sigma^2 \frac{\bar{d} + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>k} s_i^2(A) \right).$$

To the best of our knowledge, `T-LSE` enjoys state-of-the-art performance guarantees. It achieves tighter error upper bounds compared to both the rank-adaptive algorithm from Bunea et al. (2010), cf Equation (1), and the thresholded estimator with nuclear norm penalization introduced in Appendix B. More precisely, one can show (see Appendix D) that `T-LSE` is (ε, δ)-stable, for any (ε, δ), with sample complexity

$$N \leq \min \left\{ m : \phi(m) \leq c\varepsilon^2 \right\}, \tag{15}$$

where

$$\phi(m) = k^\star_{A,m}\sigma^2 \frac{\bar{d} + \log(\frac{1}{\delta})}{m\lambda_{\min}(\Sigma)} + \sum_{i > k^\star_{A,m}} s_i^2(A).$$

$c \in [0, 1]$ is a universal constant.

The error guarantees for `T-LSE` almost match the limits identified in Theorem 4.5. To achieve an exact match, it suffices to replace, in Theorem 6.3, $\lambda_{\min}(\Sigma)$ with either $\underline{\lambda}_k(\Sigma)$, if $\bar{\lambda}_k(\Sigma)(\log(1/\delta) + kd_y) \geq \underline{\lambda}_k(\Sigma)(\log(1/\delta) + kd_x)$, or $\bar{\lambda}_k(\Sigma)$ otherwise.

6.2 Linear System Identification

Analyzing the performance of the LSE in linear system identification requires leveraging recent concentration results on self-normalized processes (see, e.g., the proof of Theorem 3 in Jedra and Proutiere (2022)). This analysis allows us to upper bound the error Z, provided that the number of samples satisfies

$$n \geq \frac{\max(c_0\sigma^4, 1)\|\Gamma_\infty(A)\|^3}{\lambda_{\min}(\Gamma_\infty(A))} \left(\log\left(\frac{1}{\delta}\right) + d_x \right), \tag{16}$$

for some universal constant $c_0 > 0$.

Rank-constrained LSE. Assume first that the algorithm selects an effective rank equal to $r \geq \mathrm{rank}(A)$. The `R-LSE` is defined as $\hat{A}_n := \Pi_r(\bar{A})$. Assuming the noise is Gaussian with variance σ^2, we can establish the following concentration result for $\Pi_k(Z)$.

Lemma 6.4. *Assume n verifies inequality (16). Then, for all $k \in [\underline{d}]$, with probability at least $1-\delta$:*

1. *The error Z verifies*

$$\|\Pi_k(Z)\|_{\mathrm{F}}^2 \leq k\sigma^2 \frac{d_x + \log(\frac{1}{\delta})}{n\underline{\lambda}_k^H(\hat{\Sigma})}.$$

2. *$\hat{\Sigma}$ verifies*

$$\forall i \in [d_x], \quad \frac{\lambda_i(\Gamma_\infty(A))}{8} \leq \lambda_i(\hat{\Sigma}) \leq \frac{3\lambda_i(\Gamma_\infty(A))}{2}.$$

Combining these concentration results with Lemma 5.2, we obtain:

Theorem 6.5. *Let $\delta \in (0, 1)$. Assume that n verifies (16) and that $\mathrm{rank}(A) \leq r$. The* `R-LSE` *satisfies, with probability at least $1-\delta$:*

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \lesssim r\sigma^2 \frac{d_x + \log(\frac{1}{\delta})}{n\underline{\lambda}_r^H(\Gamma_\infty(A))}.$$

Thresholded LSE. Consider now rank-adaptive algorithms. When the noise is Gaussian with variance σ^2 and when n verifies inequality (16), we establish (refer to the proof of Lemma 6.4) that, with probability at least $1-\delta$:

$$2\|Z\|_2 \leq \xi_{\mathrm{SysID}} := 2\sigma \sqrt{\frac{d_x + \log(\frac{1}{\delta})}{n\lambda_{\min}(\hat{\Sigma})}}. \tag{17}$$

Combining the previous result and Theorem 5.1, we obtain the following result.

Theorem 6.6. *Assume n verifies inequality (16). The* `T-LSE` *satisfies, with probability at least $1-\delta$:*

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \lesssim \min_{k \in [r]} \left(\frac{k\sigma^2(d_x + \log(\frac{1}{\delta}))}{n\lambda_{\min}(\Gamma_\infty(A))} + \sum_{i>k} s_i^2(A) \right).$$

To the best of our knowledge, this is the first rank-adaptive performance bound for system identification. Furthermore, the same remarks as in the multivariate regression case can be made regarding the remaining existing gaps between our lower and upper bounds.

Table 1: Performance of T-LSE and RSC in presence of alignment between covariate and target matrices. The values correspond to the relative Frobenius error multiplied by 10^{-7} for $r = 10$ and 10^{-6} for $r = 45$.

	$r = 10$			$r = 45$		
	Error Avg	Error Std	Error Max	Error Avg	Error Std	Error Max
T-LSE	1.67	**1.03**	**4.75**	**5.49**	**0.69**	**8.20**
RSC	**1.37**	1.23	7.62	8.37	4.86	23.1

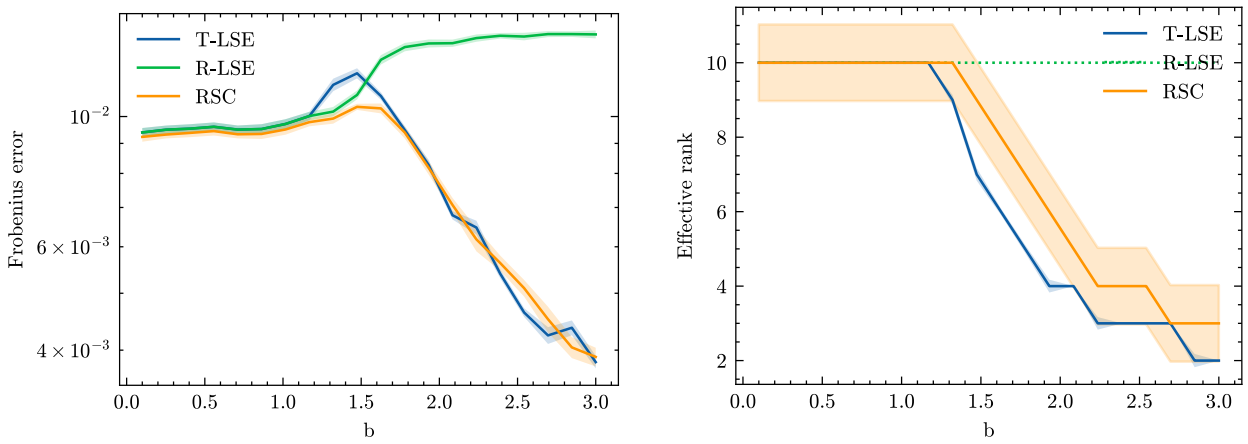

Figure 1: Frobenius error (left) and effective rank (right) vs noise level b, ($r = 10$).

7 NUMERICAL EXPERIMENTS

We test our algorithms experimentally in the multivariate regression case using similar settings as existing work Bunea et al. (2010). We estimate $A \in \mathbb{R}^{d\times d}$ of rank r with $d = 50$ using $n = 1000$ samples. Our experiments can also be conducted for higher values of (d, r, n), and these can be found in Appendix E. Each 'run' involves sampling the design matrix X and the Gaussian noise matrix E to compute $Y = XA+E$. We evaluate the performance of an estimator $\hat{A}_n$ using the relative Frobenius error $\|\hat{A}_n - A\|_{\mathrm{F}}^2 / \|A\|_{\mathrm{F}}^2$ averaged over $T = 30$ runs. We consider two settings: (1) Low-rank $r = 10, \sigma = 0.1$. (2) High-rank $r = 45, \sigma = 0.4$.

On the impact of singular subspace alignment. We begin by comparing `T-LSE` and the RSC algorithm Bunea et al. (2010) (initially designed to minimize $\|X\hat{A}_n - XA\|_{\mathrm{F}}$), and in particular show how the 'alignment' of the singular subspaces of the design matrix X and of A impact the performance of these two algorithms. To this aim, we start by sampling X: for $i = 1, ..., n, \quad x_i \sim \mathcal{N}(0, \Sigma)$ where Σ is a diagonal matrix such that $\Sigma_{j,j} = j^2$ for $j = 1, ..., d$. Let $X = U_X S_X V_X^\top$ be its SVD. Then we create a set of $d + 1$ matrices A with different levels of alignment, as follows: (1) Sample entries of A uniformly at random in $[0, 1]$, and let $A = USV^\top$ be its SVD. (2) Change the singular value of A as $s_j(A) = \frac{1}{j^b}$ for $j = 1, ..., r$ (we use $b = 1.5$ in the low-rank case and $b = 0.5$ in the high-rank case). Note that the assumption of decaying singular values is frequent, see Wu et al. (2020). (3) Change its right singular vectors to align them to those of X: $A = USPV_X^\top$, where P is one of the $(d + 1)$ circular shifting matrices (i.e., shifting circularly the positions of the rows $V_X^\top$). We have constructed $d + 1$ scenarios with different levels of alignment, and report the performance of `T-LSE` and RSC in Table 1, averaged across all scenarios as well as for the worst-case alignment. `T-LSE` demonstrates greater robustness and consistently outperforms RSC, particularly in challenging alignment conditions.

On the importance of adaptivity. Next we compare adaptive algorithms, `T-LSE` and RSC, to the non-adaptive algorithm `R-LSE` aware of the true rank of A, $r = 10$. If the number of samples n is fixed and if the noise level increases, we expect that optimal algorithms should target an effective rank strictly smaller than r. Here we increase the level of noise by increasing a parameter b. The singular values of A are now set to $s_j(A) = \frac{1}{j^b}$ for $j = 1, ..., r$, and we do not align its singular vectors with those of X. In Figure 1, for $r = 10$, we observe two interesting regimes. When $b \in [0.1, 1.5]$ is small, the signal is still strong enough for `R-LSE` to be competitive compared to the rank-adaptive algorithms. For $b \in [1.5, 3]$ both adaptive algorithms `T-LSE` and RSC target a lower effective rank and have better performance than `R-LSE`. Additional figures are provided in Appendix E for the high-rank regime.

8 CONCLUSION

We revisited the problem of estimating high-dimensional matrices via reduced-rank regression, focusing on rank-adaptive algorithms. These methods estimate the matrix by learning its singular values and corresponding singular vectors up to an effective rank, which is adaptively determined from the data. For this setting, we established instance-specific lower bounds on the sample complexity that any such algorithm must satisfy. These bounds explicitly depend on the spectral properties of the underlying matrix and the covariance structure of the covariates, offering insight into the effective rank that an optimal algorithm should infer.

We proposed the thresholded `LSE` , which combines the classical `LSE` with a universal singular value threshold-

ing procedure. We derived finite-sample error bounds for this estimator and showed that it achieves better perfomance guarantees than existing algorithms. A small gap remains between the performance of `LSE` and the theoretical limits. Closing this gap is an interesting direction for future research. Another promising avenue is the development of similarly adaptive estimation procedures under stronger norms—such as entry-wise norms—which are particularly relevant in applications like reinforcement learning, where such guarantees are often required.

Acknowledgments. This research is supported by Vetenskaprådet, Digital Futures, and the Wallenberg AI, Autonomous Systems and Software program.

References

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351.

Anderson, T. W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, 27(4):1141–1154.

Bakshi, A., Liu, A., Moitra, A., and Yau, M. (2023). A new approach to learning linear dynamical systems. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 335–348.

Barzilai, D. and Shamir, O. (2024). Simple relative deviation bounds for covariance and gram matrices. *arXiv preprint arXiv:2410.05754*.

Batir, N. (2008). Inequalities for the gamma function. *Archiv der Mathematik*, 91(6):554–563.

Bunea, F., She, Y., and Wegkamp, M. (2010). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39.

Cai, T., Ma, Z., and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3):781–815.

Candès, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candes, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920.

Djehiche, B. and Mazhar, O. (2021). Non asymptotic estimation lower bounds for LTI state space models with Cramer-Rao and van Trees. *arXiv preprint arXiv:2109.08582*.

Djehiche, B. and Mazhar, O. (2022). Efficient learning of hidden state LTI state space models of unknown order. *arXiv preprint arXiv:2202.01625*.

Fan, K. (1951). Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences*, 37(11):760–766.

Fazel, M., Pong, T. K., Sun, D., and Tseng, P. (2013). Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977.

Gavish, M. and Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053.

Gerchinovitz, S., Ménard, P., and Stoltz, G. (2020). Fano's inequality for random variables.

Henkel, O. (2005). Sphere-packing bounds in the Grassmann and Stiefel manifolds. *IEEE Transactions on Information Theory*, 51(10):3445–3456.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.

Jedra, Y. and Proutiere, A. (2019). Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, page 2676–2681. IEEE Press.

Jedra, Y. and Proutiere, A. (2022). Finite-time identification of linear systems: Fundamental limits and optimal algorithms. *IEEE Transactions on Automatic Control*, 68(5):2805–2820.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302 – 2329.

Ma, T., Verchand, K. A., and Samworth, R. J. (2024). High-probability minimax lower bounds. *arXiv preprint arXiv:2406.13447*.

Marshall, A. W., Olkin, I., and Arnold, B. C. (1979). Inequalities: theory of majorization and its applications.

Mataigne, S., Absil, P.-A., and Miolane, N. (2024). Bounds on the geodesic distances on the Stiefel manifold for a family of Riemannian metrics. *arXiv preprint arXiv:2408.07072*.

Mori, T. (2002). Comments on" A matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equation" by JM Saniuk and IB Rhodes. *IEEE transactions on automatic control*, 33(11):1088.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069 – 1097.

Oymak, S. and Ozay, N. (2019). Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate reduced-rank regression*. Springer.

Rohde, A. and Tsybakov, A. (2009). Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39:887–930.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. (2021). Finite time LTI system identification. *Journal of Machine Learning Research*, 22(26):1–61.

Simchowitz, M., Boczar, R., and Recht, B. (2019). Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR.

Sun, S., Li, J., and Mo, Y. (2023). Finite Time Performance Analysis of MIMO Systems Identification. *arXiv preprint arXiv:2310.11790*.

Sun, Y., Oymak, S., and Fazel, M. (2022). Finite sample identification of low-order lti systems via nuclear norm regularization. *IEEE Open Journal of Control Systems*, 1:237–254.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wu, Q., Wong, F. M., Li, Y., Liu, Z., and Kanade, V. (2020). Adaptive reduced rank regression. *Advances in Neural Information Processing Systems*, 33:4103–4114.

Xiang, S., Zhu, Y., Shen, X., and Ye, J. (2012). Optimal exact least squares rank minimization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 480–488.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.

Zhang, Y., Talebi, S., and Li, N. (2024). Learning low-dimensional latent dynamics from high-dimensional observations: Non-asymptotics and lower bounds. *arXiv preprint arXiv:2405.06089*.

Ziemann, I., Tsiamis, A., Lee, B., Jedra, Y., Matni, N., and Pappas, G. J. (2023). A tutorial on the non-asymptotic theory of system identification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 8921–8939. IEEE.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes. Sections 2 and 6 define the sampling model and least square estimator respectively.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes. Sections 4 and 6 presents an analysis of the sample complexity with lower/upper bounds. Section 5 provides with additional properties for thresholded estimators.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes. An anonymous Jupyter notebook with all required dependencies to run our experiments is provided.**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes. Section 4 presents our lower bounds with assumptions on expected covariance matrix and the rank. Section 6 presents concentration results with the required sample complexity.**
   (b) Complete proofs of all theoretical results. **Yes. All our theoretical results have corresponding proof in the supplementary material, see Appendices A, B, C, D.**
   (c) Clear explanations of any assumptions. **Yes. Our assumptions are standard for our objective i.e identifiability (positive definiteness of Σ), low-rankness ($2r \leq d$), empirical covariance concentration.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes. The Jupyter notebook in supplementary material is structured into several headings to clearly reproduce experiments.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes. We present choices of hyperparameters and their influence in Appendix E.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes. We define the Frobenius error and rank of our estimators either on Y-axis of a figure or as numerical value in a table.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes. See Appendix E.5.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes. We used a visualisation library whose repository is cited in Appendix E.5.**
   (b) The license information of the assets, if applicable. **Not Applicable.**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable.**
   (d) Information about consent from data providers/curators. **Not Applicable.**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

Contents

A Proofs of results presented in Section 4

In this section, we provide the proofs for our complexity sample lower bounds, namely Theorem 4.3 and Theorem 4.5. The proofs rely on intricate change-of-measure arguments that allow for multiple hypotheses. In §A.1, we start by presenting some of the notations and tools used in these arguments. In §A.2, we prove a result on the packing number of the Stiefel manifold given in Lemma 4.1. We then present the proofs of Theorem 4.3 and Theorem 4.5 in §A.3 and §A.4, respectively.

A.1 Tools for the change-of-measure argument

We start by introducing notation which we will be used throughout this section. We consider a collection of n observations $(x_1, y_1), \dots, (x_n, y_n)$. We will often view these observations as samples from a probability distribution with a density function f_A parameterized by $A \in \mathbb{R}^{d_y \times d_x}$. For some given $A \in \mathbb{R}^{d_y \times d_x}$, we use $\mathbb{E}_A$ (resp. $\mathbb{P}_A$) to denote the expectation (resp. the probability measure) under the distribution with the density f_A.

Computation of the expected log-likelihood ratio. Let $A, A' \in \mathbb{R}^{d_y \times d_x}$ such that $A \neq A'$. We define the log-likelihood ratio of a collection of n observations $(x_1, y_1), \dots, (x_n, y_n)$, under A and A' as follows:

$$L_n(A, A') = \log\left(\frac{f_A((x_1, y_1), \dots, (x_n, y_n))}{f_{A'}((x_1, y_1), \dots, (x_n, y_n))}\right).$$

When $(x_1, y_1), \dots, (x_n, y_n)$ are generated under a linear model as presented in §2 with Gaussian noise then the expected log-likelihood can be computed explicitly:

Lemma A.1. *Assume that $(x_1, y_1), \dots, (x_n, y_n)$ are generated under a linear model as presented in §2, i.e., $y_i = Ax_i + \eta_i$, where $(\eta_i)_{i \geq 1}$ is a sequence of i.i.d. of random variables distributed as $\mathcal{N}(0, \sigma^2 I_{d_y})$. Then, the expected log-likelihood ratio under A and A' is as follows:*

$$\mathbb{E}_A[L_n(A, A')] = \frac{1}{2\sigma^2}\mathrm{Tr}\left((A - A')^\top (A - A')\mathbb{E}_A[X^\top X]\right).$$

The computations leading to the above result follow the same steps as in Jedra and Proutiere (2019), we provide a proof for completeness.

Proof of Lemma A.1. Consider that we have n observations $(x_1, y_1), \dots, (x_n, y_n)$. Before computing the expectation of the log-likelihood ratio, we use conditional independence between the observations to write the ratio of densities as a product of ratios. In the multivariate regression setting, we have by independence

$$\frac{f_A((x_i, y_i)_{i=1}^n)}{f_{A'}((x_i, y_i)_{i=1}^n)} = \prod_{i=1}^n \frac{f_A(x_i, y_i)}{f_{A'}(x_i, y_i)} = \prod_{i=1}^n \frac{f_A(y_i|x_i)f(x_i)}{f_{A'}(y_i|x_i)f(x_i)} = \prod_{i=1}^n \frac{f_A(y_i|x_i)}{f_{A'}(y_i|x_i)}.$$

In the system identification case, the observations $((x_i, y_i)_{i=1}^n)$ simplify to $(x_i)_{i=1}^{n+1}$. In a similar fashion, by conditional independence, we have

$$\frac{f_A(x_1, ..., x_{n+1})}{f_{A'}(x_1, ..., x_{n+1})} = \prod_{i=1}^n \frac{f_A(x_{i+1}|x_i)}{f_{A'}(x_{i+1}|x_i)} = \prod_{i=1}^n \frac{f_A(y_i|x_i)}{f_{A'}(y_i|x_i)}$$

with $y_i = x_{i+1}$.

Therefore,

$$\mathbb{E}_A[L_n(A,A')] = \mathbb{E}_A \sum_{i=1}^n \left[\log \frac{f_A(y_i|x_i)}{f_{A'}(y_i|x_i)}\right] = \sum_{i=1}^n \mathbb{E}_A \left[\log \frac{f_A(y_i|x_i)}{f_{A'}(y_i|x_i)}\right]$$
$$= \sum_{i=1}^n \mathbb{E}_A \left[\mathbb{E}_A \left[\log \frac{f_A(y_i|x_i)}{f_{A'}(y_i|x_i)}|x_i\right]\right].$$

Noting that $f_A(y_i|x_i) = \mathcal{N}(Ax_i, \sigma^2 I_{d_y})$ (and similarly for A') we obtain

$$\begin{aligned}
\mathbb{E}_A[L_n(A,A')] &= \sum_{i=1}^n \mathbb{E}_A \left[\mathrm{KL}\left(\mathcal{N}(Ax_i, \sigma^2 I_{d_y}), \mathcal{N}(A'x_i, \sigma^2 I_{d_y})\right)\right] \\
&= \sum_{i=1}^n \mathbb{E}_A \left[\frac{1}{2\sigma^2} \mathrm{Tr}(x_i^\top (A-A')^\top (A-A') x_i)\right] \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^n \mathbb{E}_A \left[\mathrm{Tr}((A-A')^\top (A-A') x_i x_i^\top)\right] \\
&= \frac{1}{2\sigma^2} \mathbb{E}_A \left[\mathrm{Tr}((A-A')^\top (A-A') X^\top X)\right] \\
&= \frac{1}{2\sigma^2} \mathrm{Tr}\left((A-A')^\top (A-A') \mathbb{E}_A[X^\top X]\right).
\end{aligned}$$

□

Data-processing inequality for multiple hypotheses. We present a variant of the data processing inequality which allows for using multiple hypothesis in deriving lower bounds. This result is borrowed from Jedra and Proutiere (2022) (see their Proposition 2).

Lemma A.2. *Let $kl(p,q)$ the KL-divergence between two Bernoulli with parameters p and q. Let n, m be two positive integers and consider $\mathcal{E}_1, ..., \mathcal{E}_m$ disjoint events belonging to the filtration $\mathcal{F}_n$. Then, for all $A, A_1, ..., A_m$ such that for $i = 1, ..., m$, $A_i \neq A$, we have the following*

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{A_i}(L_n(A_i, A)) \geq kl\left(\frac{1}{m} \sum_{i=1}^m \mathbb{P}_{A_i}(\mathcal{E}_i), \frac{\mathbb{P}_A(\cup_{i=1}^m \mathcal{E}_i)}{m}\right).$$

Furthermore, if $\mathbb{P}_{A_i}(\mathcal{E}_i) \geq 1-\delta$ for all $i \in [m]$, and $\cup_{i=1}^m \mathcal{E}_i \subseteq \mathcal{E}^c$ for some event $\mathcal{E}$ such that $\mathbb{P}_A(\mathcal{E}) \geq 1-\delta$ then

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{A_i}(L_n(A_i, A)) \geq kl\left(1-\delta, \frac{\delta}{m}\right) \geq \frac{1}{2} \log\left(\frac{m}{4\delta}\right).$$

Remark A.3. The first statement of Lemma A.2 resembles Fano's inequality, which states that for all events $\mathcal{E}_i$ and probability measures $\mathbb{P}_i, \mathbb{Q}_i$ with $i \in [m]$, one has

$$\frac{1}{m} \sum_{i=1}^m \mathbb{P}_i(\mathcal{E}_i) \leq \frac{\frac{1}{m}\sum_{i=1}^m \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i) + \log(2)}{\log m}. \tag{18}$$

More precisely, both results are derived from the elementary data-processing inequality which involves analyzing $\mathrm{kl}\left(\frac{1}{m}\sum_{i=1}^m \mathbb{P}_i(\mathcal{E}_i), \frac{1}{m}\sum_{i=1}^m \mathbb{Q}_i(\mathcal{E}_i)\right)$. While this quantity can be additionally simplified leading to Fano's inequality, see Proposition 4 of Gerchinovitz et al. (2020) for technical details, we instead leverage it with our stability definitions. This allows us to derive a tight dependency in $\log(\frac{1}{\delta})$ by carefully choosing the events $\mathcal{E}_i$, as shown by the second statement of Lemma A.2.

On the other hand, Fano's inequality cannot recover alone this $\log(\frac{1}{\delta})$ dependency. As a remedy, authors of Ma et al. (2024) propose a separate LeCam's two point method (see their Corollary 6). However as its name indicates

and as far as we are aware, it can only be applied to two hypothesis, and therefore cannot leverage the packing argument from which we derive the correct dependency of $\log(\frac{1}{\delta})$ with respect to all model parameters in our lower bounds.

A.2 Proof of Lemma 4.1

Proof. Let $k \leq \frac{d}{2}$. Let Γ be the usual Gamma function. We first introduce the geodesic distance on the Stiefel manifold, following the formalism of Henkel (2005). For any skew-symmetric $A \in \mathbb{R}^{k\times k}$ (i.e., such that $A^\top = -A$), and $B \in \mathbb{R}^{(d-k)\times k}$, we define

$$X = \begin{pmatrix} A & -B^\top \\ B & 0 \end{pmatrix} \in \mathbb{R}^{d\times d}.$$

The geodesic distance between the matrix $Q = \exp(X)\begin{pmatrix} I_k \\ 0 \end{pmatrix} \in \mathrm{St}_k^d(\mathbb{R})$ and $I_{d,k} = \begin{pmatrix} I_k \\ 0 \end{pmatrix} \in \mathrm{St}_k^d(\mathbb{R})$ (where $I_k \in \mathbb{R}^{k\times k}$ is the identity matrix) is defined by $d_g(Q, I_{d,k}) = \frac{1}{2}\|X\|_{\mathrm{F}}^2 = \frac{1}{2}\|A\|_{\mathrm{F}}^2 + \|B\|_F^2$. The geodesic distance $d_g(Q_1, Q_2)$ between arbitrary $Q_1, Q_2 \in \mathrm{St}_k^d(\mathbb{R})$ follows from the isometric transformation $Q_1' = Q_2^T Q_1,\ Q_2' = I_{d,k}$: it corresponds to $d_g(Q_1', I_{d,k})$.

For any $m \geq 1$, Theorem 4.1 in Henkel (2005) states the existence of a packing $\mathcal{P}_{k,d}$ of $\mathrm{St}_k^d(\mathbb{R})$ of size m, with minimal geodesic distance between its elements

$$\forall (Q_1, Q_2) \in \mathcal{P}_{k,d} \quad d_g(Q_1, Q_2) \geq d_0 := ab.$$

where

$$a = \left(\frac{1}{2}\right)^{\frac{\log_2 m}{2dk-k^2}}, \qquad b = \left((2\sqrt{\pi})^k \frac{\Gamma\left(\frac{k(2d-k)}{2} + 1\right)}{\prod_{i=d-k+1}^{d} \Gamma(i)}\right)^{\frac{1}{2dk-k^2}}.$$

We will prove that with an appropriate choice of m, we obtain a packing with the properties required in Lemma 4.1. Let us select m such that $\log_2(m) = 2dk - k^2 = k(2d-k)$. With this choice, we can see that $a = \frac{1}{2}$. The next lemma, proved below, gives a lower bound on b.

Lemma A.4. *There exists a universal constant $c \approx 0.17$ such that $b \geq e^{\frac{-1-\log(2)}{3} - \frac{1}{2} - c}\sqrt{\frac{k}{2}}$.*

From this lemma, we deduce that the distance between two elements of the packing satisfies:

$$d_0 = ab \geq \frac{e^{\frac{-1-\log(2)}{3} - \frac{1}{2} - c}}{2\sqrt{2}}\sqrt{k}.$$

Hence we have proved that there exists a packing of size $2^{k(2d-k)} \geq 2^{kd}$ with minimal geodesic distance lower bounded by $\sqrt{k}$ up to a universal constant. To complete the proof of Lemma 4.1, we have to relate the geodesic distance to that induced by the Frobenius norm. To this aim, we use Corollary 7.2 of Mataigne et al. (2024) stating that

$$\forall (Q_1, Q_2) \in \mathrm{St}_k^d(\mathbb{R}), \quad d_g(Q_1, Q_2) \leq \frac{\pi}{2}\|Q_1 - Q_2\|_{\mathrm{F}}.$$

Hence the packing in the geodesic distance also induces a packing in the Frobenius distance with minimal distance $\frac{\sqrt{k}}{C}$ where $C := \pi\sqrt{2}e^{\frac{1+\log(2)}{3} + \frac{1}{2} + c}$.

□

Proof of Lemma A.4. We use Theorem 1.6 in Batir (2008): For any $x \geq 1$

$$\left(\frac{x}{e}\right)^x \sqrt{2\pi x} \leq \Gamma(x+1) \leq \left(\frac{x}{e}\right)^x \sqrt{2\pi(x+c)}$$

where $c \approx 0.17$.

Let $x = \frac{\log_2 m}{2}$, then we can use this result to bound b as follows

$$\begin{aligned} b &\geq \left((2\sqrt{\pi})^k \frac{(\frac{x}{e})^x \sqrt{2\pi x}}{\prod_{i=d-k}^{d-1} (\frac{i}{e})^i \sqrt{2\pi i}} \right)^{\frac{1}{2dk-k^2}} \left(\prod_{i=d-k}^{d-1} \sqrt{1+\frac{c}{i}} \right)^{\frac{-1}{2dk-k^2}} \\ &= \sqrt{\frac{x}{e}} \left(\frac{\sqrt{2}^{k+1}\sqrt{\pi}}{\prod_{i=d-k}^{d-1} (\frac{i}{e})^i \sqrt{i}} \right)^{\frac{1}{2dk-k^2}} \left(\prod_{i=d-k}^{d-1} \sqrt{1+\frac{c}{i}} \right)^{\frac{-1}{2dk-k^2}} \\ &\geq \sqrt{\frac{x}{e}} \underbrace{\left(\frac{1}{\prod_{i=d-k}^{d-1} (\frac{i}{e})^i \sqrt{i}} \right)^{\frac{1}{2dk-k^2}}}_{:=b_0} \underbrace{\left(\prod_{i=d-k}^{d-1} \sqrt{1+\frac{c}{i}} \right)^{\frac{-1}{2dk-k^2}}}_{:=b_1}, \end{aligned}$$

where the second line holds since $(\frac{x}{e})^{\frac{x}{2dk-k^2}} = \sqrt{\frac{x}{e}}$, and $\sqrt{x}^{\frac{1}{2dk-k^2}} = x^{\frac{1}{x}} \geq 1$. Next we compute the logarithms of b_0 and b_1.

Computing $\log(b_1)$. We have:

$$\begin{aligned} \log\left(\prod_{i=d-k}^{d-1} \sqrt{1+\frac{c}{i}} \right) &= \frac{1}{2} \sum_{i=d-k}^{d-1} \log\left(1+\frac{c}{i}\right) \leq \frac{1}{2} k \log\left(1+\frac{c}{d-k}\right) \\ &\leq \frac{kc}{2(d-k)} \leq \frac{c}{2} \qquad \text{since } k \leq \frac{d}{2}. \end{aligned}$$

We conclude that: $b_1 = \left(\prod_{i=d-k}^{d-1} \sqrt{1+\frac{c}{i}} \right)^{\frac{-1}{2dk-k^2}} \geq e^{-\frac{c}{2(2dk-k^2)}} \geq e^{-c}$.

Computing $\log(b_0)$. We decompose the product $\prod_{i=d-k}^{d-1} (\frac{i}{e})^i \sqrt{i}$ into two products $\prod_{i=d-k}^{d-1} \sqrt{i}$ and $\prod_{i=d-k}^{d-1} (\frac{i}{e})^i$.

- We upper bound the logarithm of the first product as follows:
  $\log\left(\prod_{i=d-k}^{d-1} \sqrt{i} \right) = \frac{1}{2} \sum_{i=d-k}^{d-1} \log(i) \leq \frac{1}{2} k \log(d)$. We obtain

$$\left(\prod_{i=d-k}^{d-1} \sqrt{i} \right)^{\frac{-1}{2dk-k^2}} \geq e^{-\frac{k\log(d)}{2k(2d-k)}} \geq e^{-\frac{\log(d)}{2(2d-k)}} \geq e^{-\frac{1}{3}}.$$

- For the second product, its logarithm can also be upper bounded as follows: $\log\left(\prod_{i=d-k}^{d-1} \left(\frac{i}{e}\right)^i \right) = \sum_{i=d-k}^{d-1} i\log(i) - \sum_{i=d-k}^{d-1} i \leq \sum_{i=d-k}^{d-1} i\log(i)$. Since $t\log(t)$ is an increasing convex function for $t \geq 1$, we get

$$\begin{aligned} \sum_{i=d-k}^{d-1} i\log(i) &\leq \int_{i=d-k}^{d} t\log(t)dt \leq k\frac{(d-k)\log(d-k) + d\log(d)}{2} \\ &= k\frac{(d-k)\log(d-k) + d(\log(d-k) + \log(1+\frac{k}{d-k}))}{2} \\ &\leq k\frac{(2d-k)\log(d-k) + d\log 2}{2}. \end{aligned}$$

  We deduce that $\left(\prod_{i=d-k}^{d-1} \left(\frac{i}{e}\right)^i \right)^{\frac{-1}{2dk-k^2}} \geq e^{-\frac{\log(d-k)}{2} - \frac{d\log(2)}{2(2d-k)}} \geq \frac{1}{\sqrt{d-k}} e^{\frac{-\log(2)}{3}}$.

Plugging the above inequalities together, we obtain $b_0 \geq e^{\frac{-1-\log(2)}{3}}\sqrt{\frac{1}{d-k}}$.
We conclude that

$$b = \sqrt{\frac{x}{e}}b_0 b_1 \geq e^{\frac{-1-\log(2)}{3}-\frac{1}{2}-c}\sqrt{\frac{x}{d-k}} = e^{\frac{-1-\log(2)}{3}-\frac{1}{2}-c}\sqrt{\frac{k(2d-k)}{2(d-k)}}$$
$$\geq e^{\frac{-1-\log(2)}{3}-\frac{1}{2}-c}\sqrt{\frac{k}{2}}.$$

□

A.3 Proof of Theorem 4.3

Proof. Let $\varepsilon > 0$ and $\delta \in (0,1)$. We suppose that $\mathbb{E}(\hat{\Sigma}) \succ 0$. Let A be such that $\text{rank}(A) \leq r \leq \frac{1}{2}\underline{d}$. Let $\hat{A}_n$ be an (ε, δ, r)-stable algorithm in A and let N be its sample complexity. The lower bound on N stated in Theorem 4.3 is the maximum over two lower bounds, one that depends on d_x and one that depends on d_y. We prove both lower bounds below.

1) Lower bound that depends on d_y. We start by applying Lemmas A.1 and A.2 to A using the confusing models from $\mathcal{C}_1(A, r, \varepsilon)$. More precisely, introduce for all $i \in [2^{rd_y}]$,

$$A_i = A + \frac{2C\varepsilon}{\sqrt{r}}Q_i W_{-r}^\top, \quad \text{where} \quad Q_i \in \mathcal{P}_r^{d_y},$$

and $C \geq 1$ is a universal constant previously defined in Lemma 4.1.

Application of Lemma A.1: We have

$$\forall i \in [2^{rd_y}], \quad \mathbb{E}_{A_i}[L_N(A_i, A)] = \frac{1}{2\sigma^2}\text{Tr}\left((A-A_i)^\top(A-A_i)\mathbb{E}_{A_i}[X^\top X]\right),$$

where $X \in \mathbb{R}^{N\times d_x}$ is the (random) matrix of covariates.

- In the multivariate regression case, we clearly have: $\mathbb{E}_{A_i}[X^\top X] = \mathbb{E}[X^\top X] = N\Sigma$, and therefore,

$$\mathbb{E}_{A_i}[L_N(A_i, A)] = \frac{N}{2\sigma^2}\text{Tr}((A-A_i)^\top(A-A_i)\Sigma) = \frac{4NC^2\varepsilon^2}{2\sigma^2 r}\text{Tr}(Q_i W_{-r}^\top \Sigma W_{-r} Q_i^\top)$$
$$= \frac{2NC^2\varepsilon^2}{\sigma^2 r}\text{Tr}(W_{-r}^\top \Sigma W_{-r}) = \frac{2NC^2\varepsilon^2\underline{\lambda}_r(\Sigma)}{\sigma^2}.$$

- In the system identification case, the trajectory is generated by A itself. Lemma 8 in Jedra and Proutiere (2022) shows that

$$\mathbb{E}_{A_i}\left[\sum_{n=0}^{N-2} x_n x_n^\top\right] = \sum_{n=0}^{N-2}\Gamma_n(A_i) \preceq N\Gamma_\infty(A) + N\left(\Gamma_\infty(A_i) - \Gamma_\infty(A)\right).$$

Furthermore, since $\|A_i - A\|_2 = \frac{2C\varepsilon}{\sqrt{r}} \leq \frac{\|\Gamma_\infty(A)\|_2^{-3}}{4}$ then Lemma 1 in Jedra and Proutiere (2022) implies that

$$\|\Gamma_\infty(A_i) - \Gamma_\infty(A)\|_2 \leq 16\|A_i - A\|_2\|\Gamma_\infty(A)\|_2^3.$$

Hence

$$\begin{aligned}\mathrm{Tr}(W_{-r}^\top \mathbb{E}_{A_i}[X^\top X]W_{-r}) &= \sum_{n=0}^{N-2} \mathrm{Tr}(W_{-r}^\top \Gamma_n(A_i) W_{-r}) \\ &\leq N\mathrm{Tr}(W_{-r}^\top \Gamma_\infty(A) W_{-r}) + N\mathrm{Tr}\Big(W_{-r}^\top \big(\Gamma_\infty(A_i) - \Gamma_\infty(A)\big) W_{-r}\Big) \\ &\leq N\mathrm{Tr}(W_{-r}^\top \Gamma_\infty(A) W_{-r}) + N\|\Gamma_\infty(A_i) - \Gamma_\infty(A)\|_2 \mathrm{Tr}(W_{-r} W_{-r}^\top),\end{aligned}$$

where we applied the trace inequality found in Theorem 1 of Mori (2002), $\mathrm{Tr}(AB) \leq s_1(A)\mathrm{Tr}(B)$.

Hence,

$$\begin{aligned}\mathrm{Tr}(W_{-r}^\top \mathbb{E}_{A_i}[X^\top X]W_{-r}) &\leq Nr\underline{\lambda}_r(\Gamma_\infty(A))) + 16Nr\|A_i - A\|_2 \|\Gamma_\infty(A)\|_2^3 \\ &\leq Nr\underline{\lambda}_r(\Gamma_\infty(A)) + 16Nr \\ &\leq 17Nr\underline{\lambda}_r(\Gamma_\infty(A)),\end{aligned}$$

where the last inequality comes from $\underline{\lambda}_r(\Gamma_\infty(A)) \geq 1$ since $\Gamma_\infty(A) \succeq I$.

Finally, we obtain the following upper bound in the system identification case

$$\mathbb{E}_{A_i}[L_N(A_i, A)] \leq \frac{2C^2\varepsilon^2}{\sigma^2 r} 17Nr\underline{\lambda}_r(\Gamma_\infty(A)) \leq \frac{40NC^2\varepsilon^2\underline{\lambda}_r(\Gamma_\infty(A))}{\sigma^2}.$$

Application of Lemma A.2: Consider

$$\forall i \in [2^{rd_y}] \quad \mathcal{E}_i = \{\|A_i - \hat{A}_n\|_{\mathrm{F}} \leq \varepsilon\} \quad \text{and} \quad \mathcal{E} = \{\|A - \hat{A}_n\|_{\mathrm{F}} \leq \varepsilon\}.$$

Given N samples, by stability of $\hat{A}_n$, one has for $i \in [2^{rd_y}]$, $\mathbb{P}_{A_i}(\mathcal{E}_i) \geq 1 - \delta$ and $\mathbb{P}_A(\mathcal{E}) \geq 1 - \delta$.

Furthermore, these confusing models verify

$$\forall i \in [2^{rd_y}], \quad \|A_i - A\|_{\mathrm{F}} = \frac{2C\varepsilon}{\sqrt{r}}\|Q_i W_{-r}^\top\|_{\mathrm{F}} = 2C\varepsilon \geq 2\varepsilon,$$

since both Q_i, W_{-r} are semi-orthogonal and the Frobenius norm is unitarily invariant. Similarly, we have

$$\forall i \neq j \in [2^{rd_y}], \quad \|A_i - A_j\|_{\mathrm{F}} = \frac{2C\varepsilon}{\sqrt{r}}\|(Q_i - Q_j)W_{-r}^\top\|_{\mathrm{F}} = \frac{2C\varepsilon}{\sqrt{r}}\|Q_i - Q_j\|_{\mathrm{F}} \geq 2\varepsilon,$$

where the last inequality holds by Lemma 4.1. Since the A_i are all distant from each other by at least 2ε then the $\mathcal{E}_i$ are all pairwise disjoint. Furthermore, since the A_i are also distant from A by at least 2ε then $\cup_{i=1}^{2^{rd_y}} \mathcal{E}_i \subset \mathcal{E}^c$. We can now apply Lemma A.2 and obtain

- For i.i.d multivariate regression:

$$\begin{aligned}&\frac{1}{2}\log\left(\frac{2^{rd_y}}{4\delta}\right) \leq \frac{1}{2}\log\left(\frac{|\mathcal{P}_r^{d_y}|}{4\delta}\right) \leq \sum_{i=1}^{2^{rd_y}} \frac{\mathbb{E}_{A_i}[L_N(A_i, A)]}{2^{rd_y}} \\ \implies \quad &rd_y \log(2) + \log(\frac{1}{4\delta}) \leq \frac{4NC^2\varepsilon^2\underline{\lambda}_r(\Sigma)}{\sigma^2} \\ \implies \quad &\log(2)(rd_y + \log(\frac{1}{\delta}) - 1) \leq \frac{4NC^2\varepsilon^2\underline{\lambda}_r(\Sigma)}{\sigma^2} \\ \implies \quad &\frac{\log 2}{2}(rd_y + \log(\frac{1}{\delta})) \leq \frac{4NC^2\varepsilon^2\underline{\lambda}_r(\Sigma)}{\sigma^2}.\end{aligned}$$

This implies the following lower bound

$$N \geq \frac{\sigma^2 \log(2)}{8} \frac{\log(\frac{1}{\delta}) + rd_y}{\varepsilon^2 \underline{\lambda}_r(\Sigma)} \gtrsim \sigma^2 \frac{\log(\frac{1}{\delta}) + rd_y}{\varepsilon^2 \underline{\lambda}_r(\Sigma)}.$$

- For dynamical systems:

$$\frac{1}{2}\log\left(\frac{2^{rd_y}}{4\delta}\right) \leq \frac{1}{2}\log\left(\frac{|\mathcal{P}_r^{d_y}|}{4\delta}\right) \leq \sum_{i=1}^{2^{rd_y}} \frac{\mathbb{E}_{A_i}[L_N(A_i, A)]}{2^{rd_y}}$$
$$\implies \quad rd_y \log(2) + \log(\frac{1}{4\delta}) \leq \frac{80NC^2\varepsilon^2 \underline{\lambda}_r(\Gamma_\infty(A))}{\sigma^2}.$$

This implies the following lower bound

$$N \geq \frac{\sigma^2 \log(2)}{160} \frac{\log(\frac{1}{\delta}) + rd_y}{\varepsilon^2 \underline{\lambda}_r(\Gamma_\infty(A))} \gtrsim \sigma^2 \frac{\log(\frac{1}{\delta}) + rd_y}{\varepsilon^2 \underline{\lambda}_r(\Gamma_\infty(A))}.$$

2) Lower bound that depends on d_x. This case is relevant only for the multivariate regression since in this case, d_x might differ from d_y.

Application of Lemma A.1: We use the confusing models from $\mathcal{C}_2(A, r, \varepsilon)$. Introduce for all $i \in [2^{rd_y}]$,

$$A_i = A + \frac{2C\varepsilon}{\sqrt{r}} U_r R_i^\top, \qquad \text{where } R_i \in \mathcal{P}_r^{d_x}.$$

Then

$$\mathbb{E}_{A_i}[L_N(A_i, A)] = \frac{N}{2\sigma^2}\mathrm{Tr}((A - A_i)^\top (A - A_i)\Sigma) = \frac{4NC^2\varepsilon^2}{2\sigma^2 r}\mathrm{Tr}(U_r R_i^\top \Sigma R_i U_r^\top)$$
$$= \frac{2NC^2\varepsilon^2}{\sigma^2 r}\mathrm{Tr}(R_i^\top \Sigma R_i) \leq \frac{2NC^2\varepsilon^2 \bar{\lambda}_r(\Sigma)}{\sigma^2} \quad \text{by Lemma A.5.}$$

Application of Lemma A.2: Using the same arguments as to derive the lower bound that depends on d_y, we obtain

$$\frac{1}{2}\log\left(\frac{2^{rd_x}}{4\delta}\right) \leq \frac{1}{2}\log\left(\frac{|\mathcal{P}_r^{d_x}|}{4\delta}\right) \leq \sum_{i=1}^{2^{rd_x}} \frac{\mathbb{E}_{A_i}[L_N(A_i, A)]}{2^{rd_x}}$$
$$\implies \quad rd_x \log(2) + \log(\frac{1}{4\delta}) \leq \frac{4NC^2\varepsilon^2 \bar{\lambda}_r(\Sigma)}{\sigma^2}.$$

This implies the following lower bound

$$N \gtrsim \sigma^2 \frac{\log(\frac{1}{\delta}) + rd_x}{\varepsilon^2 \bar{\lambda}_r(\Sigma)}.$$

Combining both lower bounds on the sample complexity, we obtain the stated result. □

A.4 Proof of Theorem 4.5

Proof. Let $\varepsilon > 0$ and $\delta \in (0, 1)$. We suppose that $\mathbb{E}(\hat{\Sigma}) \succ 0$. Let A such that $\mathrm{rank}(A) \leq r \leq \frac{1}{2}\underline{d}$. Let $\hat{A}_n$ be an (ε, δ)-stable algorithm in A and let N be its sample complexity. By stability assumption, N verifies $\|A - \Pi_{k^\star_{A,N}}(A)\|_{\mathrm{F}} \leq \varepsilon$. Hence, letting $\varepsilon' = 2\varepsilon - \|A - \Pi_{k^\star_{A,N}}(A)\|_{\mathrm{F}} \geq \varepsilon$, we also have by stability

$$\forall A' \in \mathcal{D}(A, N, \varepsilon), \quad \mathbb{P}_{A'}(\|\hat{A}_n - A'\|_{\mathrm{F}} \leq \varepsilon') \geq 1 - \delta.$$

The remainder of the proof is identical to that of Theorem 4.3. We only provide the proof for the multivariate regression case (simply replace Σ by $\Gamma_\infty(A)$ for dynamical systems, and adjust the universal constant).

1) Lower bounds that depends on d_y: Consider the confusing models from $\mathcal{C}_1(A, k^\star_{A,N}, \varepsilon)$:

$$\forall i \in [2^{k^\star_{A,N} d_y}], \quad A_i = A + \frac{2C\varepsilon'}{\sqrt{k^\star_{A,N}}} Q_i W^\top_{-k^\star_{A,N}}, \qquad \text{where } Q_i \in \mathcal{P}^{d_y}_{k^\star_{A,N}}.$$

We apply Lemmas A.1 and A.2 as in the proof of Theorem 4.3 (replace r by $k^\star_{A,N}$, and ε by ε'), and derive following lower bound:

$$\varepsilon'^2 \geq \frac{\sigma^2 \log(2)}{8} \frac{\log(\frac{1}{\delta}) + k^\star_{A,N} d_y}{N \underline{\lambda}_{k^\star_{A,N}}(\Sigma)}.$$

Since $\varepsilon \geq \|A - \Pi_{k^\star_{A,N}}(A)\|_{\mathrm{F}}$, we have:

$$\begin{aligned}\varepsilon'^2 &= 4\varepsilon^2 + \|A - \Pi_{k^\star_{A,N}}(A)\|^2_{\mathrm{F}} - 4\varepsilon \|A - \Pi_{k^\star_{A,N}}(A)\|_{\mathrm{F}} \\ &\leq 4\varepsilon^2 - 3\|A - \Pi_{k^\star_{A,N}}(A)\|^2_{\mathrm{F}}.\end{aligned}$$

Combining the two previous inequalities yields:

$$\begin{aligned}4\varepsilon^2 &\geq \frac{\sigma^2 \log(2)}{8} \frac{\log(\frac{1}{\delta}) + k^\star_{A,N} d_y}{N \underline{\lambda}_{k^\star_{A,N}}(\Sigma)} + 3\|A - \Pi_{k^\star_{A,N}}(A)\|^2_{\mathrm{F}} \\ \implies \frac{32}{\log(2)} \varepsilon^2 &\geq \gamma^\delta_A(N),\end{aligned}$$

This finally implies that:

$$N \geq \min\{n : \gamma^\delta_A(n) \leq \frac{32}{\log(2)} \varepsilon^2\}.$$

2) Lower bound that depends on d_x: The proof follows along the same lines as those in the proof of Theorem 4.3.

□

A.5 Extremal partial trace

Lemma A.5. *For any $k \leq d$ and two matrices $Q \in \mathrm{St}^d_k(\mathbb{R}), \Sigma \in \mathbb{R}^{d\times d}$ symmetric positive definite, we have*

$$\sum_{i=d-k+1}^{d} \lambda_i(\Sigma) \leq \mathrm{Tr}(Q^\top \Sigma Q) \leq \sum_{i=1}^{k} \lambda_i(\Sigma). \tag{19}$$

Proof. The proof of this classical result, sometimes referred to as *extremal partial trace*, can be found in 20.A.2 of Marshall et al. (1979). □

B Proofs of results presented in Section 5

In this section, we first consider an estimator $\bar{A}$ of A and derive a tight upper bound on $\|\Pi_k(\bar{A}) - A\|_{\mathrm{F}}$ for arbitrary values of k (Lemma 5.2), highlighting the trade-off between estimation error and approximation error. We then address the problem of selecting the optimal value of k to minimize this upper bound. While solving this optimization problem exactly would require knowledge of A, which is unknown by definition, we propose a universal thresholding rule that bypasses this limitation (Theorem 5.1). Finally in §B.3, we explain how to apply the thresholding procedure to improve existing estimators. We illustrate the procedure for the estimator obtained via regression with nuclear norm regularization.

B.1 Proof of Lemma 5.2

Proof. Let $k \in [\underline{d}]$, we have by the Eckart-Young's theorem:

$$\|\Pi_k(A+Z) - A - Z\|_{\mathrm{F}}^2 \leq \|\Pi_k(A) - A - Z\|_{\mathrm{F}}^2.$$

By decomposing the squares, we obtain:

$$\|\Pi_k(A+Z) - A\|^2 \leq \|A - \Pi_k(A)\|_{\mathrm{F}}^2 + 2\langle \Pi_k(A+Z) - \Pi_k(A), Z\rangle.$$

For the l.h.s, we have, by reverse triangle inequality,

$$\begin{aligned}\|\Pi_k(A+Z) - A\|^2 &\geq \left|\|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}} - \|A - \Pi_k(A)\|_{\mathrm{F}}\right|^2 \\ &= \|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}}^2 + \|A - \Pi_k(A)\|_{\mathrm{F}}^2 - 2\|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}}\|A - \Pi_k(A)\|_{\mathrm{F}}.\end{aligned}$$

Hence, we get

$$\frac{1}{2}\|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}}^2 \leq \langle \Pi_k(A+Z) - \Pi_k(A), Z\rangle + \|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}}\|A - \Pi_k(A)\|_{\mathrm{F}}.$$

Furthermore, since $\Pi_k(A+Z) - \Pi_k(A)$ is of rank at most $2k$, we have by Lemma 2 of Xiang et al. (2012)

$$\langle \Pi_k(A+Z) - \Pi_k(A), Z\rangle \leq \|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}}\|\Pi_{2k}(Z)\|_{\mathrm{F}}.$$

This implies:

$$\frac{1}{2}\|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}} \leq \|\Pi_{2k}(Z)\|_{\mathrm{F}} + \|A - \Pi_k(A)\|_{\mathrm{F}}.$$

We conclude by noting that:

$$\begin{aligned}\|\Pi_k(\bar{A}) - A\|_{\mathrm{F}} &\leq \|\Pi_k(A+Z) - \Pi_k(A)\|_{\mathrm{F}} + \|A - \Pi_k(A)\|_{\mathrm{F}} \\ &\leq 2\|\Pi_{2k}(Z)\|_{\mathrm{F}} + 3\|A - \Pi_k(A)\|_{\mathrm{F}}.\end{aligned}$$

□

B.2 Proof of Theorem 5.1

Proof. We first prove an auxiliary useful inequality.

Let $\alpha \geq 0$ and $K = \max\{i : s_i(\bar{A}) \geq (2+\alpha)\|Z\|_2\}$. Further define $k^\star = \max\{i : s_i(A) \geq (3+\alpha)\|Z\|_2\}$. By Weyl's inequality, we have:

- $s_{K+1}(A) \leq s_{K+1}(\bar{A}) + \|Z\|_2 \leq (3+\alpha)\|Z\|_2$, which implies that $k^\star \leq K$.
- $\forall i \leq K,\ s_i(A) \geq s_i(\bar{A}) - \|Z\|_2 \geq (1+\alpha)\|Z\|_2$.

By Lemma 5.2:

$$\begin{aligned}\frac{\|\Pi_K(\bar{A}) - A\|_{\mathrm{F}}^2}{18} &\leq K\|Z\|_2^2 + \sum_{i>K} s_i^2(A) = k^\star\|Z\|_2^2 + \sum_{i>k^\star} s_i^2(A) + \psi \\ &\leq k^\star\|Z\|_2^2 + \sum_{i>k^\star} s_i^2(A) \\ &\leq (3+\alpha)^2 k^\star\|Z\|_2^2 + \sum_{i>k^\star} s_i^2(A) \\ &= \min_{k\in[r]} \left\{ (3+\alpha)^2 k\|Z\|_2^2 + \sum_{i=k+1}^{r} s_i^2(A) \right\}, \end{aligned} \tag{20}$$

where $\psi = \sum_{i=k^\star+1}^{K} \left(\|Z\|_2^2 - s_i^2(A)\right) < 0$. Let $\xi \geq 2\|Z\|_2$ and $\alpha = \frac{\xi}{\|Z\|_2} - 2$. Then we have: $\xi = (2+\alpha)\|Z\|_2$. The Theorem is obtained by combining (20) and the observation that $(3+\alpha)\|Z\|_2 = (1 + \frac{\xi}{\|Z\|_2})\|Z\|_2 \leq 2\xi$. □

B.3 Thresholded nuclear norm estimator

We demonstrate how Theorem 5.1 can be applied to devise a rank-adaptive algorithm, starting from an estimator $\bar{A}$ obtained via regression with nuclear norm regularization. We restrict our attention to the case of multivariate regression with Gaussian noise. The estimator $\bar{A}$ is defined in Negahban and Wainwright (2011) by:

$$\bar{A} = \arg\min_B \{\|Y - XB\|_{\mathrm{F}}^2 + \mu\|B\|_1\}.$$

The performance of this estimator is well understood (see Corollary 10.14 in Wainwright (2019)): if $\bar{A}$ is constructed using n samples and with $\mu = 10\sigma\sqrt{\lambda_{\max}(\hat{\Sigma})}(\sqrt{\frac{d_x+d_y}{n}} + \sqrt{\frac{\log(\frac{1}{2\delta})}{2n}})$, then the error $Z = \bar{A} - A$ satisfies, with probability at least $1-\delta$:

$$2\|Z\|_2 \leq \xi := \frac{20\sigma\sqrt{\lambda_{\max}(\hat{\Sigma})}(\sqrt{\frac{d_x+d_y}{n}} + \sqrt{\frac{\log(\frac{1}{2\delta})}{2n}})}{\lambda_{\min}(\hat{\Sigma})}.$$

Combining this concentration result to the result of Theorem 5.1 yields:

Corollary B.1. *The thresholded nuclear norm regularized estimator $\hat{A}_n := \bar{A}(\xi)$ satisfies, with probability at least $1-\delta$: $\|\hat{A}_n - A\|_{\mathrm{F}}^2 \lesssim \min_{k\in[r]} \left[k\sigma^2\kappa(\hat{\Sigma})\frac{d_x+d_y+\log(\frac{1}{\delta})}{n\lambda_{\min}(\hat{\Sigma})} + \sum_{i>k} s_i^2(A)\right]$.*

Proof. We plug in the value of $\xi = \frac{20\sigma\sqrt{\lambda_{\max}(\hat{\Sigma})}(\sqrt{\frac{d_x+d_y}{n}}+\sqrt{\frac{\log(\frac{1}{2\delta})}{2n}})}{\lambda_{\min}(\hat{\Sigma})}$ inside the upper bound given by Theorem 5.1. □

In Bunea et al. (2010), the authors derive, in their Theorem 12, an error upper bound for the plain nuclear norm regularized estimator (without thresholding). Their upper bound is similar to the one presented in the above corollary, but is $\kappa(\hat{\Sigma})$ times larger. Hence, incorporating the singular value thresholding procedure results in an algorithm with stronger performance guarantees.

C Proofs of results presented in Section 6

In this section, we provide the proofs of all results presented in Section 6. These results concern the performance of our algorithms `R-LSE` and `T-LSE`. We first observe that all stated inequalities in Section 6 are trivially true when $\hat{\Sigma}$ is singular since then each upper bound becomes infinity. Therefore, we will assume from now on that $\hat{\Sigma}$ is non-singular. In that case, the LSE is expressed as

$$\hat{A}_n = Y^\top X(X^\top X)^{-1} = A + E^\top X(X^\top X)^{-1} := A + Z,$$

allowing us to apply denoising results from Section 5. Since the norms considered are invariant by transposition, we will consider $Z := (X^\top X)^{-1}X^\top E$ for notational convenience.

C.1 Multivariate regression

C.1.1 Proof of Lemma 6.1 and Equation (14)

Proof. Let $\delta \in (0,1)$ and $k \in [\underline{d}]$. We have:

$$Z := (X^\top X)^{-1}X^\top E = (X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top E.$$

Furthermore,

- Lemma 3 of Bunea et al. (2010) states: with probability at least $1-\delta$,

$$\|X(X^\top X)^{-1}X^\top E\|_2 \le \sigma\left(\sqrt{d_x} + \sqrt{d_y} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right).$$

- For all matrices P, Q, we have

$$\|\Pi_k(PQ)\|_{\mathrm{F}}^2 = \sum_{i=1}^{k} s_i^2(PQ) \le \sum_{i=1}^{k} s_i^2(P)s_1^2(Q) = \|\Pi_k(P)\|_{\mathrm{F}}^2\|Q\|_2^2.$$

Hence, we can write

$$\begin{aligned}
\|\Pi_k(Z)\|_{\mathrm{F}}^2 &\le \|\Pi_k\left((X^\top X)^{-1}X^\top\right)\|_{\mathrm{F}}^2\|X(X^\top X)^{-1}X^\top E\|_2^2 \\
&\le \left(\sum_{i=d_x-k+1}^{d_x} \frac{1}{n\lambda_i(\hat{\Sigma})}\right)\|X(X^\top X)^{-1}X^\top E\|_2^2 \\
&\le \frac{\sum_{i=d_x-k+1}^{d_x}\frac{1}{\lambda_i(\hat{\Sigma})}}{k}\frac{k\sigma^2(\sqrt{d_x}+\sqrt{d_y}+\sqrt{\log(\frac{1}{\delta})})^2}{n},
\end{aligned}$$

which concludes the proof of Lemma 6.1.

The proof of Equation (14) is similar and uses the sub-multiplicativity of the operator norm to obtain

$$\begin{aligned}
\|Z\|_2 &\le \|(X^\top X)^{-1}X^\top\|_2\|X(X^\top X)^{-1}X^\top E\|_2 \\
&\le \frac{1}{\sqrt{n\lambda_{\min}(\hat{\Sigma})}}\|X(X^\top X)^{-1}X^\top E\|_2 \\
&\le \frac{\sigma\left(\sqrt{d_x}+\sqrt{d_y}+\sqrt{\log(\frac{1}{\delta})}\right)}{\sqrt{n\lambda_{\min}(\hat{\Sigma})}}.
\end{aligned}$$

□

C.1.2 General error bounds for any $\hat{\Sigma}$

Before stating the proofs of our main Theorems 6.2 and 6.3 presented in our main body, we start by stating general concentration-free result, valid for any $\hat{\Sigma}$.

Theorem C.1. *Assume that* $\mathrm{rank}(A) \leq r$, *the* `R-LSE` *satisfies, with probability at least* $1-\delta$*:*

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \leq 6\sqrt{2} r\sigma^2 \frac{\bar{d} + \log(\frac{1}{\delta})}{n\underline{\lambda}_r^H(\hat{\Sigma})}.$$

Proof. We first apply Lemma 6.1 with $k = r$ which gives with probability at least $1-\delta$

$$\|\Pi_r(Z)\|_{\mathrm{F}}^2 \leq \frac{r\sigma^2(\sqrt{d_x} + \sqrt{d_y} + \sqrt{\log(\frac{1}{\delta})})^2}{n\underline{\lambda}_r^H(\hat{\Sigma})}.$$

We plug this inequality in Lemma 5.2, also applied to $k = r$, which gives

$$\begin{aligned}\|\hat{A}_n - A\|_{\mathrm{F}} &= \|\Pi_r(\bar{A}) - A\|_{\mathrm{F}} \leq 2\sqrt{2}\|\Pi_r(Z)\|_{\mathrm{F}} \\ &\leq 2\sqrt{2}\frac{r\sigma^2(\sqrt{d_x} + \sqrt{d_y} + \sqrt{\log(\frac{1}{\delta})})^2}{n\underline{\lambda}_r^H(\hat{\Sigma})} \leq C_{\texttt{R-LSE}}\frac{r\sigma^2(d_x + d_y + \log(\frac{1}{\delta}))}{n\underline{\lambda}_r^H(\hat{\Sigma})},\end{aligned}$$

where $C_{\texttt{R-LSE}} = 6\sqrt{2}$.

□

In the adaptive case, one has the following

Theorem C.2. *The* `T-LSE` $\hat{A}_n := \bar{A}(\xi_{\mathrm{MR}})$ *satisfies, with probability at least* $1-\delta$*:*

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \leq 864 \min_{k\in[r]} \left(k\sigma^2 \frac{\bar{d} + \log(\frac{1}{\delta})}{n\lambda_{\min}(\hat{\Sigma})} + \sum_{i>k} s_i^2(A) \right).$$

Proof. We proved above that $\xi_{\mathrm{MR}} := \frac{2\sigma\left(\sqrt{d_x}+\sqrt{d_y}+\sqrt{\log(\frac{1}{\delta})}\right)}{\sqrt{n\lambda_{\min}(\hat{\Sigma})}} \geq 2\|Z\|_2$ with probability at least $1-\delta$. We can plug this value inside the upper bound given by Theorem 5.1 to obtain

$$\begin{aligned}\|\hat{A}_n - A\|_{\mathrm{F}}^2 &= \|\bar{A}(\xi_{\mathrm{MR}}) - A\|_{\mathrm{F}} \\ &\leq 18 \min_{k\in[r]} \left(4k\xi_{\mathrm{MR}}^2 + \sum_{i>k} s_i^2(A) \right) \\ &\leq 18 \min_{k\in[r]} \left(4k\frac{4\sigma^2\left(\sqrt{d_x} + \sqrt{d_y} + \sqrt{\log(\frac{1}{\delta})}\right)^2}{n\lambda_{\min}(\hat{\Sigma})} + \sum_{i>k} s_i^2(A) \right) \\ &\leq C_{\texttt{T-LSE}} \min_{k\in[r]} \left(\frac{k\sigma^2\left(d_x + d_y + \log(\frac{1}{\delta})\right)}{n\lambda_{\min}(\hat{\Sigma})} + \sum_{i>k} s_i^2(A) \right),\end{aligned}$$

where $C_{\texttt{T-LSE}} = 18 \times 16 \times 3 = 864$. □

To obtain matching lower and upper bounds, we need to further apply concentration results uniformly on the spectrum of $\hat{\Sigma}$. Such results have recently been studied in Barzilai and Shamir (2024) and applied to Gaussian inputs. We collect some of their findings in Appendix D and leverage them to prove Theorems 6.2 and 6.3.

C.1.3 Performance of R-LSE: proof of Theorem 6.2

Proof. Let $n \geq 288(d_x + \log(\frac{2}{\delta}))$. We first apply Theorem C.1 which gives with probability at least $1 - \delta/2$:

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \leq 6\sqrt{2} r \sigma^2 \frac{\bar{d} + \log(\frac{2}{\delta})}{n \underline{\lambda}_r^H(\hat{\Sigma})}.$$

By Lemma D.1, we can replace $\hat{\Sigma}$ by Σ up to a universal constant with probability at least $1 - \delta/2$.

Combining both results concludes the proof of Theorem 6.2.

□

C.1.4 Performance of T-LSE: proof of Theorem 6.3

Proof. Let $n \geq 288(d_x + \log(\frac{2}{\delta}))$. We first apply Theorem C.2 which gives with probability at least $1 - \delta/2$:

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \leq 864 \min_{k \in [r]} \left(k\sigma^2 \frac{\bar{d} + \log(\frac{2}{\delta})}{n \lambda_{\min}(\hat{\Sigma})} + \sum_{i>k} s_i^2(A) \right).$$

By Lemma D.1, we can replace $\hat{\Sigma}$ by Σ up to a universal constant with probability at least $1 - \delta/2$.

Combining both results concludes the proof of Theorem 6.3.

□

C.2 Linear system identification

C.2.1 Proof of Lemma 6.4 and Equation (17)

Proof. Let n verifying inequality (16).

We first prove (i). Let $\delta \in (0, 1)$ and $k \in [\underline{d}]$.

We use a slightly different decomposition of the LSE error compared to the multivariate regression case. This decomposition will allow us to apply recent concentration results.

$$Z = (X^\top X)^{-1} X^\top E = (X^\top X)^{-\frac{1}{2}} (X^\top X)^{-\frac{1}{2}} X^\top E.$$

In the proof of Theorem 3 of Jedra and Proutiere (2022), it is shown that

$$\mathbb{P}\left[\|(X^\top X)^{-\frac{1}{2}} X^\top E\|_2 \geq \sigma \sqrt{d_x + \log\left(\frac{1}{\delta}\right)} \right] \leq \delta, \tag{21}$$

when the following condition is verified

$$\lambda_{\min}\left(\sum_{i=0}^{n-2} \Gamma_i(A) \right) \geq c_0 \sigma^4 \left(\sum_{i \geq 0} \|A^i\|_2 \right)^2 \left(d_x + \log\left(\frac{1}{\delta}\right) \right).$$

for some universal constant $c_0 > 0$.

Applying items (i) and (iii) of Lemma D.3, we can replace $\lambda_{\min}\left(\sum_{i=0}^{n-2} \Gamma_i(A)\right)$ by $n\lambda_{\min}(\Gamma_\infty(A))$ and $\left(\sum_{i \geq 0} \|A^i\|_2\right)^2$ by $\|\Gamma_\infty(A)\|_2^3$ up to some positive universal constants which we incorporate into c_0. Therefore, verifying inequality (16) is sufficient to obtain Equation 21.

In that case, using similar arguments as in the proof of Lemma 6.1, we have with probability at least $1-\delta$,

$$\begin{aligned}\|\Pi_k(Z)\|_{\mathrm{F}}^2 &\leq \|\Pi_k\left((X^\top X)^{-\frac{1}{2}}\right)\|_{\mathrm{F}}^2 \|(X^\top X)^{-\frac{1}{2}}X^\top E\|_2^2 \\ &\leq \frac{\sum_{i=d_x-k+1}^{d_x}\frac{1}{\lambda_i(\hat{\Sigma})}}{k}\frac{k\sigma^2(d_x+\log(\frac{1}{\delta}))^2}{n},\end{aligned}$$

We now prove (ii). Since $\Gamma_\infty(A) \succ 1$ then $\frac{\|\Gamma_\infty(A)\|_2^3}{\lambda_{\min}(\Gamma_\infty(A))} \geq \|\Gamma_\infty(A)\|_2$. Therefore verifying inequality (16) ensures that $n \geq 2\log(2)\|\Gamma_\infty(A)\|_2$. By (iii) of Theorem D.2, we obtain the desired result. This concludes the proof of Lemma 6.4.

The proof of Equation (17) uses the sub-multiplicativity of the operator norm to obtain

$$\|Z\|_2 \leq \|(X^\top X)^{-\frac{1}{2}}\|_2\|(X^\top X)^{-\frac{1}{2}}X^\top E\|_2 \leq \sigma\sqrt{\frac{d_x+\log(\frac{1}{\delta})}{n\lambda_{\min}(\hat{\Sigma})}}.$$

□

C.2.2 Performance of R-LSE: proof of Theorem 6.5

Proof. We first apply both (i) and (ii) of Lemma 6.4 with $k=r$, which gives with probability at least $1-\delta$,

$$\|\Pi_r(Z)\|_{\mathrm{F}}^2 \lesssim \frac{r\sigma^2(\sqrt{d_x}+\sqrt{d_y}+\sqrt{\log(\frac{1}{\delta})})^2}{n\underline{\lambda}_r^H(\Gamma_\infty(A))}.$$

We plug this inequality in Lemma 5.2, also applied to $k=r$, which gives $\|\Pi_r(\bar{A})-A\|_{\mathrm{F}} \leq 2\sqrt{2}\|\Pi_r(Z)\|_{\mathrm{F}}$ to obtain the desired result. □

C.2.3 Performance of T-LSE: proof of Theorem 6.6

Proof. We proved above that $\xi_{\mathrm{SysID}} := 2\sigma\sqrt{\frac{d_x+\log(\frac{1}{\delta})}{n\lambda_{\min}(\hat{\Sigma})}} \geq 2\|Z\|_2$ with probability at least $1-\delta$. We can plug this value inside the upper bound given by Theorem 5.1 to obtain

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \lesssim \min_{k\in[r]}\left(\frac{k\sigma^2(d_x+\log(\frac{1}{\delta}))}{n\lambda_{\min}(\hat{\Sigma})} + \sum_{i>k}s_i^2(A)\right).$$

We then use (ii) of Lemma 6.4 to replace $\hat{\Sigma}$ by $\Gamma_\infty(A)$ up to a positive universal constant and obtain the desired result. □

D Stability of R-LSE and T-LSE

In this section, we present results showing that R-LSE and T-LSE are stable in the sense of the definitions presented in §4. The results follow from Theorems 6.2 and Theorem 6.3 but require the intermediate step of establishing concentration bounds on the spectrum empirical covariance matrix $\hat{\Sigma}$ in both the multivariate regression and system identification settings. In Lemma D.1 and Theorem D.2, we present concentration on the spectrum of such matrices. Equipped with these concentration bounds we can revisit our guarantees for R-LSE and T-LSE to provide upper bounds that only exhibit a dependence on the true covariance matrix Σ in the case multivariate regression and on $\Gamma_\infty(A)$ in the case of linear system identification.

D.1 Concentration of the empirical covariance matrix

To obtain sample complexity upper bounds for R-LSE and T-LSE , we need to derive concentration bounds on the entire spectrum of $\hat{\Sigma}$. The derivation of such concentrations depends a lot on the collection of random variables $x_1, \dots, x_n$.

Multivariate regression. When $x_1, \dots, x_n$ are i.i.d. gaussian random variables with zero mean and covariance Σ, we have the following result.

Lemma D.1 (Theorem 8 of Barzilai and Shamir (2024))**.** *Let $X \in \mathbb{R}^{n \times d_x}$, with $d_x \le n$, be a matrix whose rows $x_i \sim \mathcal{N}(0, \Sigma)$ are i.i.d. Then for any $\delta > 0$, it holds with probability at least $1 - \delta$,*

$$\forall i \in [d_x], \quad |\lambda_i(\hat{\Sigma}) - \lambda_i(\Sigma)| \le 2\varepsilon + \varepsilon^2, \quad \text{where} \quad \varepsilon = \sqrt{\frac{d_x}{n}} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}}.$$

In particular, for any $\delta > 0$, provided $n \ge 288(d_x + \log(\frac{1}{\delta}))$, it holds with probability at least $1 - \delta$,

$$\forall i \in [d], \quad \frac{1}{2}\lambda_i(\Sigma) \le \lambda_i(\hat{\Sigma}).$$

Linear system identification. In this setting, the random variables $x_1, \dots, x_n$ are dependent, which makes the analysis of $\hat{\Sigma}$ more challenging, especially if we need a concentration result on the entire spectrum. We present the following result.

Theorem D.2. *Let $\delta > 0$. Assume that for all $i \ge 0$, $x_{i+1} = Ax_i + \eta_i$, where $x_0 = 0$, and $(\eta_i)_{i \ge 0}$ are i.i.d. zero-mean, σ^2-subgaussian random variables. Recall that $\hat{\Sigma} = \frac{1}{n} X^\top X$ and $\Sigma = \frac{1}{n}\sum_{i=0}^{n-1} \Gamma_i(A)$. Then, with probability $1 - \delta$, it holds*

$$\forall i \in [d_x], \qquad \frac{1}{2}\lambda_i(\Sigma) \le \lambda_i(\hat{\Sigma}) \le \frac{3}{2}\lambda_i(\Sigma)$$

provided that $n \ge \frac{c_1 \sigma^4 \|\Gamma_\infty(A)\|_2^3}{\lambda_{\min}(\Sigma)} \left(\log\left(\frac{1}{\delta}\right) + c_2 d_x\right)$ where c_1, c_2 are positive universal constants.

In particular, with probability $1 - \delta$, it holds that

$$\forall i \in [d_x], \qquad \frac{1}{8}\lambda_i(\Gamma_\infty(A)) \le \lambda_i(\hat{\Sigma}) \le \frac{3}{2}\lambda_i(\Gamma_\infty(A))$$

provided that $n \ge \frac{c_3 \sigma^4 \|\Gamma_\infty(A)\|_2^3}{\lambda_{\min}(\Gamma_\infty(A))} \left(\log\left(\frac{1}{\delta}\right) + c_4 d_x\right)$ where c_3, c_4 are positive universal constants.

Before proving this theorem, we need to review some key intermediate lemmas. First, we provide the following result which lists certain properties about the controllability Gramians.

Lemma D.3 (Lemma 8 in Jedra and Proutiere (2022)). *Assume that A is stable, meaning its spectral radius satisfies $\rho(A) < 1$. Then, the following properties hold*

(i) $\sum_{t=0}^{\infty} \|A^t\|_2 \leq (1+\sqrt{2})\|\Gamma_\infty(A)\|_2^{3/2}$

(ii) for all $i \geq 0$, $\left(1 - \exp(-\frac{i+1}{\|\Gamma_\infty(A)\|_2 - 1})\right)\Gamma_\infty(A) \preceq \Gamma_i(A) \preceq \Gamma_\infty(A)$

(iii) for $n \geq 2\log(2)\|\Gamma_\infty(A)\|_2$, *we have* $\frac{1}{4}\Gamma_\infty(A) \preceq \frac{1}{n}\sum_{i=0}^{n-1}\Gamma_i(A) \preceq \Gamma_\infty(A)$.

Proof of Lemma D.3. The results corresponding of *(i)* and *(ii)* are borrowed from Lemma 8 in Jedra and Proutiere (2022). We will therefore omit their proofs here. To prove *(iii)*, note that $\Gamma_\infty(A) \succeq \frac{1}{n}\sum_{i=0}^{n-1}\Gamma_i(A)$ is immediate from *(ii)*. Let us then prove the lower bound which also follows from *(ii)*. Note that

$$\begin{aligned}
\Sigma = \frac{1}{n}\sum_{i=0}^{n-1}\Gamma_i(A) &\succ \frac{1}{n}\sum_{i=\lfloor n/2\rfloor}^{n-1}\Gamma_i(A) \\
&\succeq \frac{1}{n}\left(\sum_{i=\lfloor n/2\rfloor}^{n-1}(1 - e^{-\frac{i+1}{\|\Gamma_\infty(A)\|_2 - 1}})\Gamma_\infty(A)\right) \\
&\succeq \frac{1}{n}\left(1 - e^{-\frac{\lfloor n/2\rfloor+1}{\|\Gamma_\infty(A)\|_2-1}}\right)(n - \lfloor n/2\rfloor)\Gamma_\infty(A) \\
&\succeq \frac{1}{2}\left(1 - e^{-\frac{\lfloor n/2\rfloor+1}{\|\Gamma_\infty(A)\|_2-1}}\right)\Gamma_\infty(A).
\end{aligned}$$

We also have

$$n \geq 2\log(2)\|\Gamma_\infty(A)\|_2 \implies \left(1 - e^{-\frac{\lfloor n/2\rfloor+1}{\|\Gamma_\infty(A)\|_2-1}}\right) \geq \frac{1}{2}.$$

Thus, whenever the above condition holds it follows that

$$\Sigma \succeq \frac{1}{4}\Gamma_\infty(A).$$

□

Next, we present a concentration result which is central in the proof of Theorem D.2.

Lemma D.4 (Lemma 4 in Jedra and Proutiere (2022)). *Let $\varepsilon, \delta > 0$. Assume that for all $i \geq 0$, $x_{i+1} = Ax_i + \eta_i$, where $x_0 = 0$, and $(\eta_i)_{i\geq 0}$ are i.i.d. zero-mean, σ^2-subgaussian random variables. We recall that $\Sigma = \frac{1}{n}\sum_{i=0}^{n-1}\Gamma_i(A)$. Then, with probability at least $1-\delta$, it holds*

$$\left\|\frac{1}{n}\left(X\Sigma^{-\frac{1}{2}}\right)^\top\left(X\Sigma^{-\frac{1}{2}}\right) - I_{d_x}\right\|_2 \leq \max(\varepsilon, \varepsilon^2),$$

provided $n \geq \frac{576(3+2\sqrt{2})\sigma^4\|\Gamma_\infty(A)\|_2^3}{\varepsilon^2\lambda_{\min}(\Sigma)}\left(\log\left(\frac{1}{\delta}\right) + \log(18)d_x\right)$.

We refer the reader to Jedra and Proutiere (2022) for a proof (see their Lemma 4) or Ziemann et al. (2023) for a revised proof with exact constants (see their Lemma B.1). We further used Lemma D.3 to upper bound $\|\sum_{i=0}^{n-1}\Gamma_i(A)\|_2$.

Next, we present a relative perturbation bound on the spectrum of the empirical covariance matrix.

Lemma D.5 (Theorem 5 in Barzilai and Shamir (2024)). *Let $X \in \mathbb{R}^{n\times d_x}$, with $n \geq d_x$. Let $\Sigma \in \mathbb{R}^{d_x\times d_x}$ be a symmetric positive definite matrix. Define $\hat{\Sigma} = \frac{1}{n}X^\top X$. For all $i \in [d_x]$, we have*

$$|\lambda_i(\hat{\Sigma}) - \lambda_i(\Sigma)| \leq \lambda_i(\Sigma) \left\| \frac{1}{n}\left(X\Sigma^{-\frac{1}{2}}\right)^\top \left(X\Sigma^{-\frac{1}{2}}\right) - I_{d_x} \right\|_2$$

Proof of Theorem D.2. The concentration result follows immediately by applying first Lemma D.4 with $\varepsilon = 1/2$, then using the relative perturbation bounds given by Lemma D.5. The second statement of the Theorem is an immediate consequence of the first and simply follows by further using the inequality *(iii)* of Lemma D.3. □

D.2 On the (ε, δ, r)-stability of R-LSE

Corollary D.6. *Let $\varepsilon > 0$, $\delta \in (0,1)$ and A such that $r \geq \operatorname{rank}(A)$. Then:*

(i) Multivariate regression. Let $x_i \sim \mathcal{N}(0,\Sigma)$ with $\Sigma \succ 0$. For any number of samples $n \gtrsim d_x + \log(\frac{1}{\delta})$, R-LSE satisfies

$$\mathbb{P}_A\left[\|\hat{A} - A\|_{\mathrm{F}} \lesssim r\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\underline{\lambda}_r^H(\Sigma)}\right] \geq 1 - \delta,$$

and is (ε, δ, r)-stable with sample complexity

$$N \lesssim r\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{\varepsilon^2 \underline{\lambda}_r^H(\Sigma)}.$$

(ii) Linear system identification. For any number of samples n verifying inequality (16), R-LSE satisfies

$$\mathbb{P}_A\left[\|\hat{A} - A\|_{\mathrm{F}} \lesssim r\sigma^2 \frac{d_x + \log(\frac{1}{\delta})}{n\underline{\lambda}_r^H(\Gamma_\infty(A))}\right] \geq 1 - \delta,$$

and is (ε, δ, r)-stable with sample complexity

$$N \lesssim r\sigma^2 \frac{d_x + \log(\frac{1}{\delta})}{\varepsilon^2 \underline{\lambda}_r^H(\Gamma_\infty(A))}.$$

Proof. Let $\varepsilon > 0$ and $\delta \in (0,1)$ and A such that $r \geq \operatorname{rank}(A)$. We focus on the multivariate regression case (for system identification, just replace Σ by $\Gamma_\infty(A)$). Let $n \geq 288(d_x + \log(\frac{1}{\delta}))$.

Let $C_{\texttt{R-LSE}} > 1$ the universal constant defined in the Proof of Theorem 6.2 such that given n samples, one has with probability at least $1 - \delta$,

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \leq C_{\texttt{R-LSE}} \frac{r\sigma^2(d_x + d_y + \log(\frac{1}{\delta}))}{n\underline{\lambda}_r^H(\Sigma)}.$$

Consider now

$$n = 2C_{\texttt{R-LSE}} \frac{r\sigma^2(d_x + d_y + \log(\frac{1}{\delta}))}{\varepsilon^2 \underline{\lambda}_r^H(\Sigma)},$$

such that by Theorem 6.2 , we have $\|\hat{A}_n - A\|_{\mathrm{F}}^2 \leq \varepsilon^2/2 \leq \varepsilon^2$.

Let now $A' \in \mathcal{C}(A, r, \varepsilon)$. Since A' has rank at most $2r$ then given n samples we also have by Theorem 6.2, with probability at least $1 - 2\delta$,

$$\|\hat{A}_n - A'\|_{\mathrm{F}}^2 \leq C_{\texttt{R-LSE}} \frac{2r\sigma^2(d_x + d_y + \log(\frac{1}{\delta}))}{n\underline{\lambda}_r^H(\Sigma)} \leq \varepsilon^2.$$

Hence, we have just shown that R-LSE is $(\varepsilon, 2\delta, r)$-stable with n samples (we can easily replace 2δ by δ simply by adding universal constants in the definition of n). Since the sample complexity N is defined as the minimum integer such that stability holds then we get that

$$N \leq n \lesssim r\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{\varepsilon^2 \underline{\lambda}_r^H(\Sigma)}.$$

□

D.3 On the (ε, δ)-stability of T-LSE

Corollary D.7. *Let $\varepsilon > 0$, $\delta \in (0,1)$ and A such that $r \geq \mathrm{rank}(A)$. Then:*

(i) Multivariate regression. Let $x_i \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. For any number of samples $n \gtrsim d_x + \log(\frac{1}{\delta})$, T-LSE satisfies

$$\mathbb{P}_A\left[\|\hat{A}_n - A\|_{\mathrm{F}} \lesssim \min_{k\in[r]} \left(k\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>k} s_i^2(A) \right)\right] \geq 1 - \delta,$$

and is (ε, δ)-stable with sample complexity

$$N \leq \min\left\{ m : k^\star_{A,m}\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{m\lambda_{\min}(\Sigma)} + \sum_{i>k^\star_{A,m}} s_i^2(A) \leq c_5\varepsilon^2 \right\},$$

where $0 < c_5 < 1$ is a universal constant.

(ii) Linear system identification. For any number of samples n verifying inequality (16), T-LSE satisfies

$$\mathbb{P}_A\left[\|\hat{A}_n - A\|_{\mathrm{F}} \lesssim \min_{k\in[r]} \left(k\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Gamma_\infty(A))} + \sum_{i>k} s_i^2(A) \right)\right] \geq 1 - \delta,$$

and is (ε, δ)-stable with sample complexity

$$N \leq \min\left\{ m : k^\star_{A,m}\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{m\lambda_{\min}(\Gamma_\infty(A))} + \sum_{i>k^\star_{A,m}} s_i^2(A) \leq c_6\varepsilon^2 \right\},$$

where $0 < c_6 < 1$ is a universal constant.

Proof. Let $\varepsilon > 0$ and $\delta \in (0,1)$ and A such that $r \geq \mathrm{rank}(A)$. We focus on the multivariate regression case (for system identification, just replace Σ by $\Gamma_\infty(A)$). Let $n \geq 288(d_x + \log(\frac{1}{\delta}))$.

Let $C_{\texttt{T-LSE}} > 1$ the universal constant defined in the Proof of Theorem 6.3 such that given n samples, one has with probability at least $1 - \delta$,

$$\|\hat{A}_n - A\|_{\mathrm{F}}^2 \leq C_{\texttt{T-LSE}} \min_{k\in[r]} \left(k\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>k} s_i^2(A) \right).$$

Consider

$$n = \min\left\{ m : k^\star_{A,m}\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{m\lambda_{\min}(\Sigma)} + \sum_{i>k^\star_{A,m}} s_i^2(A) \leq \frac{\varepsilon^2}{2C_{\texttt{T-LSE}}} \right\}.$$

By Theorem 6.3, we have

$$\begin{aligned}\|\hat{A}_n - A\|_{\mathrm{F}}^2 &\leq C_{\texttt{T-LSE}} \min_{k\in[r]} \left(k\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>k} s_i^2(A) \right) \\ &\leq C_{\texttt{T-LSE}} \left(k^\star_{A,n}\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>k^\star_{A,n}} s_i^2(A) \right) \\ &\leq \frac{\varepsilon^2}{2} \leq \varepsilon^2.\end{aligned}$$

The previous computation shows also that n must verify

$$\|A - \Pi_{k^\star_{A,n}}(A)\|_{\mathrm{F}} \leq \varepsilon$$

Let now $A' \in \mathcal{D}(A, n, \varepsilon)$ and denote for simplicity $K = k^\star_{A,n}$.

A' can also be rewritten as following

$$A' = \Pi_K(A) + \frac{2C\varepsilon}{\sqrt{K}} Q W_{-K}^\top + A - \Pi_K(A),$$

for some $Q \in \mathcal{P}_K^{d_y}$ (replace $QW_{-K}^\top$ by $U_K R^\top$ for some $R \in \mathcal{P}_K^{d_x}$ in the other possible case).

We have the following result on singular values, see Theorem 2 of Fan (1951): For any matrices P, Q and integers (i, j),

$$s_{i+j-1}(P+Q) \leq s_i(P) + s_j(Q).$$

Therefore, for any $i = 1, ..., r-K$ (A' has rank at most $r+K$),

$$\begin{aligned}s_{2K+i}(A') &\leq s_{2K+1}(\Pi_K(A) + \frac{2C\varepsilon}{\sqrt{K}} Q W_{-K}^\top) + s_i(A - \Pi_K(A)) \\ &= s_i(A - \Pi_K(A)) \\ &= s_{K+i}(A).\end{aligned}$$

Given n samples, we also have by Theorem 6.3, with probability at least $1 - 2\delta$,

$$\begin{aligned}\|\hat{A}_n - A'\|_{\mathrm{F}}^2 &\leq C_{\texttt{T-LSE}} \min_{k\in[r]} \left[2k\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>k} s_i^2(A') \right] \\ &\leq C_{\texttt{T-LSE}} \left(2K\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>2K} s_i^2(A') \right) \\ &\leq 2C_{\texttt{T-LSE}} \left(K\sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{n\lambda_{\min}(\Sigma)} + \sum_{i>K} s_i^2(A) \right) \\ &\leq \varepsilon^2.\end{aligned} \tag{22}$$

Hence, we have just shown that `T-LSE` is $(\varepsilon, 2\delta)$-stable with n samples (again, we can easily replace 2δ by δ by adding universal constants in the definition of n). Since the sample complexity N is defined as the minimum integer such that stability holds then we get that

$$N \leq n = \min \left\{ m : k^\star_{A,m} \sigma^2 \frac{d_x + d_y + \log(\frac{1}{\delta})}{m \lambda_{\min}(\Sigma)} + \sum_{i > k^\star_{A,m}} s_i^2(A) \leq \frac{\varepsilon^2}{2 C_{\texttt{T-LSE}}} \right\}.$$

□

Remark D.8. Finally, note that these results also hold for other rank-constrained and rank-adaptive algorithms, for instance those of Bunea et al. (2010) or the nuclear norm estimator in Appendix B, since they enjoy similar but looser upper bounds (up to model-dependent constants).

E Numerical experiments

In this section, we present further numerical experiments to illustrate the theoretical results obtained in Sections 5 and 6. We first compare our denoising lemma (derived in Section 5) to Lemma 3.5 of Chatterjee (2015), using a simple synthetic example. Next, we study the performance of our estimators `R-LSE` and `T-LSE`, and investigate the tightness of the error upper bound of `T-LSE` derived in Theorem 6.3. Finally, we describe the computational resources used for our experiments.

E.1 Matrix denoising lemmas

We first restate Lemma 3.5 in Chatterjee (2015): for $Z = \bar{A} - A$,

$$\|\bar{A}\left((1+\tau)\|Z\|_2\right) - A\|_{\mathrm{F}}^2 \leq f(\tau)\|Z\|_2\|A\|_1$$

where $f(\tau) = ((4+2\tau)\sqrt{\frac{2}{\tau}} + \sqrt{2+\tau})^2$.

In comparison, our Lemma 5.2 yields

$$\|\bar{A}\left((1+\tau)\|Z\|_2\right) - A\|_{\mathrm{F}}^2 \leq 18(k(\tau)\|Z\|_2^2 + \sum_{i>k(\tau)} s_i^2(A))$$

where $k(\tau) = \max\{i : s_i(\bar{A}) \geq (1+\tau)\|Z\|_2\}$. We are interested in comparing both upper bounds with respect to the parameter τ. A is a 50×50 matrix of rank $r = 10$ with entries sampled uniformly at random in $[-1, 1]$. We then generate $\bar{A} = A + Z$ where the entries of Z are i.i.d sampled from $\mathcal{N}(0, 1)$. Given Z and A, we evaluate the two upper bounds for several values of τ, and average them over $T = 100$ runs.

Figure 2 depicts the behavior of both upper bounds as a function of τ. We can observe that our upper bound is indeed smaller, and stays bounded between $r\|Z\|_2^2$ (corresponding to $k(\tau) = r$) and $\|A\|_{\mathrm{F}}^2$ (corresponding to $k(\tau) = 0$).

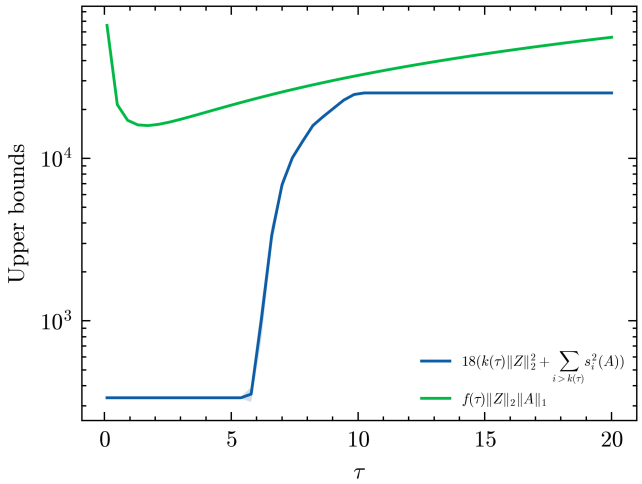

Figure 2: Comparison of Lemma 5.2 with the upper bound from Chatterjee (2015) as a function of the threshold τ.

E.2 On the importance of adaptivity

We consider the same experiments as those already presented in Section 7, which we re-describe briefly. We estimate $A \in \mathbb{R}^{d\times d}$ ($d = 50$), of rank r, from a linear model $Y = XA + E$. 1) Entries of A are first sampled uniformly at random in $[0, 1]$. 2) We compute its SVD $A = USV^\top$. 3) We keep its singular vectors but change its singular values as follows: $s_j(A) = \frac{1}{j^b}$ where b is a parameter acting as a proxy for signal-to-noise ratio. In Section 7, we considered the case where the rank of A was $r = 10$. Here we investigate the high-rank regime

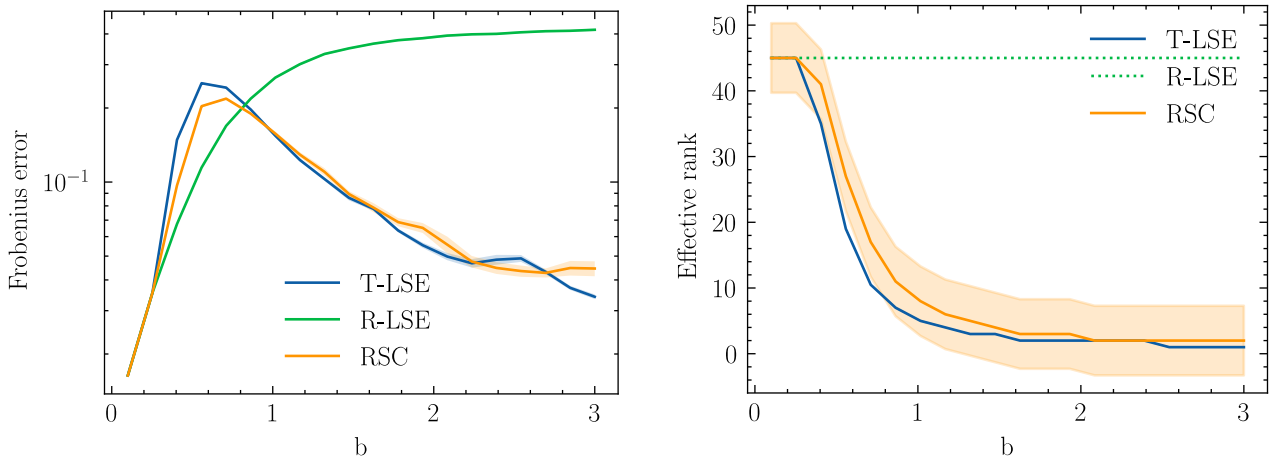

Figure 3: Multivariate regression: Frobenius error (left) and effective rank (right) vs noise level b, ($r = 45$).

with $r = 45, \sigma = 0.4$. Figure 3 shows the performance of `R-LSE` and its adaptive counterparts, `T-LSE` and RSC, as a function of b, using $n = 1000$ samples. Similar to the low-rank regime, we observe that as the noise level increases, both `T-LSE` and RSC adapt their effective rank to concentrate on the singular subspaces that can be reliably estimated from the noisy data.

E.3 System identification

We further give an example of rank-adaptive estimation for the case of linear system identification. We first construct a stable matrix $A \in \mathbb{R}^{d\times d}$ with $d = 50$ (stable means such that $\rho(A) < 1$). To do so, we consider a symmetric positive definite A constructed as follows. 1) We first sample its entries uniformly at random in $[0, 1]$. 2) We compute its SVD $A = USV^\top$. 3) We change A to $USU^\top$. 4) We change the entries of S to $(\frac{1}{(j+1)^b})_{j=1}^r$. For this choice of A, the system is stable. We assume that $x_{t+1} = Ax_t + \mathcal{N}(0, \sigma^2 I_d)$ and $x_0 = 0$ with $\sigma = 0.1$. Similarly to the Gaussian input case, Figure 4 shows that adaptiveness becomes preferable than its rank-constrained counterpart as b increases and the signal to noise ratio decreases.

E.4 Tightness of our upper bound

In this section, we investigate the tightness of the error upper bound for `T-LSE` established in Theorem 6.3. Specifically, we aim to determine whether the bound accurately captures the dependence of the true error $\|\hat{A}_n - A\|_{\mathrm{F}}^2$ (and not the relative error anymore) on the underlying problem parameters.

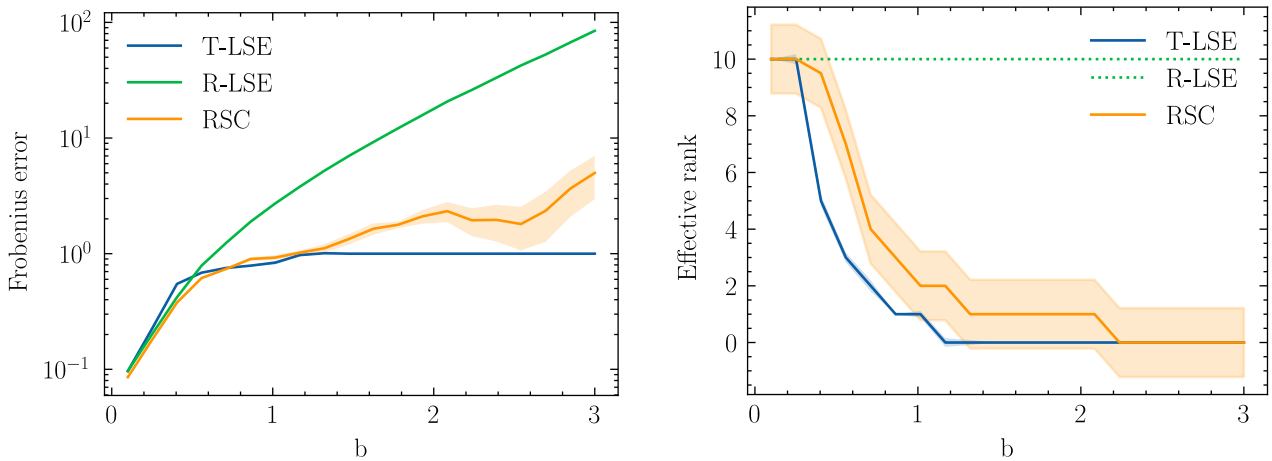

Figure 4: System identification: Frobenius error (left) and effective rank (right) vs noise level b, ($r = 10$).

For sake of simplicity, we consider the multivariate regression setting. A is a (d, d)-square matrix, with rank r. Its entries are sampled uniformly at random in $[0, 1]$. The noise matrix E entries are i.i.d sampled from $\mathcal{N}(0, 1)$. The covariates x_i are i.i.d sampled from a multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. The outputs y_i are the rows of $Y = XA + E$. We study the performances of `T-LSE` as a function of 1) number of samples n, 2) dimension d. We compare them to the upper bound given by Theorem 6.3.

For each experiment, i.e., for each choice of (n, d), we sample A, X once and then sample T times E to obtain Y and $\hat{A}_n$. We compute the identification error $\|\hat{A}_n - A\|_{\mathrm{F}}^2$ in each trial and output the average.

1. As a function of n ($d = 50$, $r = 10$). The true error of `T-LSE` and our upper bound are close. Both exhibit a hyperbolic decrease towards zero as can be seen in Figure 5.

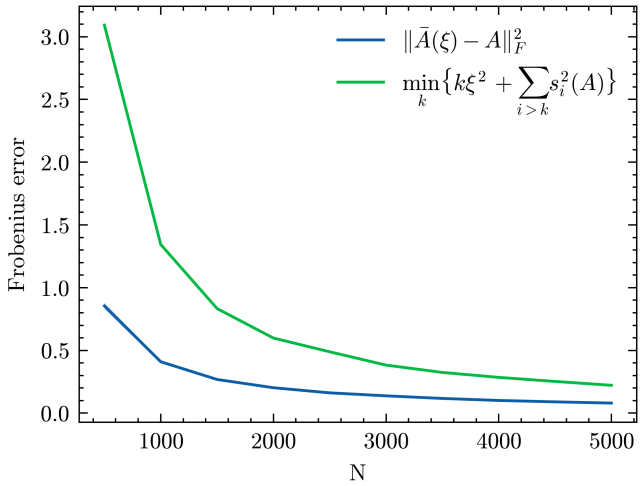

Figure 5: Performance of `T-LSE` and its corresponding error upper bound as a function of n.

2. As a function of d ($n = 5000$, $r = 10$). We observe in Figure 6 that the `T-LSE` error is linear in d as predicted by the upper bound (and lower bound). For the upper bound, the observed deviation from perfect linearity arises from the degradation of the accuracy of the empirical eigenvalue $\underline{\lambda}_r^H(\hat{\Sigma})$ as n is kept fixed.

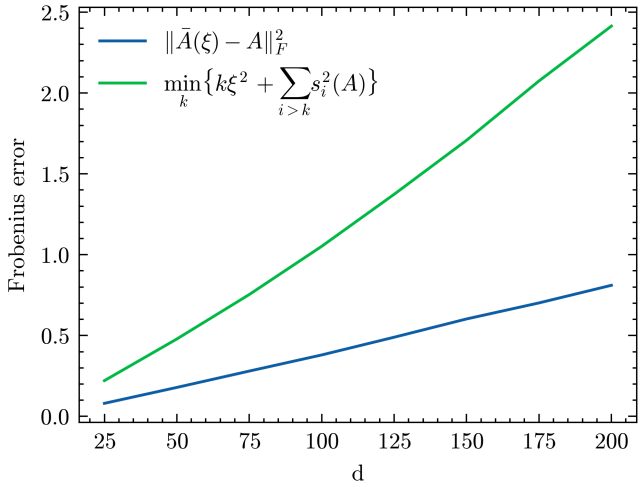

Figure 6: Performance of `T-LSE` and its corresponding error upper bound as a function of d.

Overall, our error upper bound reflects the dependence of `T-LSE`'s performance on the system parameters, but it is not always tight. Deriving sharper bounds remains an open direction for future work. In particular, an interesting step would be to replace, in our Theorems 6.3 and 6.6, $\lambda_{\min}(\hat{\Sigma})$ by $\underline{\lambda}_k^H(\hat{\Sigma})$.

E.5 Computing resources

All experiments were performed on a 12th Gen Intel(R) Core(TM) i7-1280P 1.80 GHz, with 64 Go of available RAM memory.

Coding was done on Python and standard numerical analysis libraries (e.g NumPy and Matplotlib) were used. The publicly available SciencePlots, https://github.com/garrettj403/SciencePlots library, in its Version 1.0.9, was used for plots.