Concentration inequalities for semidefinite least squares based on data

Concentration inequalities for semidefinite least squares based on data

Filippo Fabiani and Andrea Simonetto

***Abstract*—We study data-driven least squares (LS) problems with semidefinite (SD) constraints and derive finite-sample guarantees on the spectrum of their optimal solutions when these constraints are relaxed. In particular, we provide a high confidence bound allowing one to solve a simpler program in place of the full SDLS problem, while ensuring that the eigenvalues of the resulting solution are ε-close of those enforced by the SD constraints. The developed certificate, which consistently shrinks as the number of data increases, turns out to be easy-to-compute, distribution-free, and only requires independent and identically distributed samples. Moreover, when the SDLS is used to learn an unknown quadratic function, we establish bounds on the error between a gradient descent iterate minimizing the surrogate cost obtained with no SD constraints and the true minimizer.**

***Index Terms*—Data-driven modeling, Optimization, Machine learning.**

I. Introduction

LEAST squares problems are one of the workhorses of signal processing and machine learning, as well as a multitude of other domains. When the underlying problems are equipped with semidefinite (SD) constraints, we often use the term semidefinite least squares (SDLS), whose first occurrence traces back to the 60's [1]. The resulting constrained programs have a number of key applications in mechanics [2]–[4], finance [5], [6], stochastic control [7], and functional estimation [8]. Besides finding widespread application, one of the main challenges in SDLS is the presence of the SD constraint, which increases significantly the overall computational complexity.

In particular, throughout this letter we will be interested in *data-driven*, convex SDLS of the following general form:

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & \frac{1}{N}\|Ax-b\|^2+\rho\|x\|^2 \\ \text{s.t.} \quad & \Lambda(F(x))\in[m,L], \end{aligned} \tag{1}$$

where a dataset consisting of N samples $\{z^{(i)}\}_{i=1}^N$, with $z^{(i)}:=(x^{(i)},y^{(i)})$, $x^{(i)}\in\mathbb{R}^n$, and possibly noisy $y^{(i)}\in\mathbb{R}$, populate matrix $A\in\mathbb{R}^{N\times n}$ and vector $b\in\mathbb{R}^N$. While the convex set $\mathcal{X}\subseteq\mathbb{R}^n$ introduce inequality and equality constraints, $F:\mathbb{R}^n\to\mathbb{S}^\ell$ maps x into a symmetric matrix relation with spectrum $\Lambda(F(x))$ constrained within $[m,L]$. In particular, $F(\cdot)$ amounts to a (LMI), $F(x)=F_0+x_1F_1+\ldots+x_nF_n$, for given symmetric matrices $F_0,F_1,\ldots,F_n\in\mathbb{S}^\ell$. Moreover, while the first part in the cost performs mean squared error minimization, the second one denotes the regularization term with $\rho>0$, making the overall cost strongly convex with a unique optimal solution.

Note that one can readily transform (1) into the associated matrix version by the vectorization operator $\mathrm{vec}(\cdot)$ stacking the matrix columns into a vector, i.e., $x=\mathrm{vec}(X)$. We can then write $\|Ax-b\|^2=\|CX-B\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm, with $b=\mathrm{vec}(B)$ and, by exploiting the Kronecker product, $C=I\otimes A$. Therefore, popular instances that can be immediately recast in the form of (1) are at least two. The first one amounts to the *projection-onto-the-cone*:

$$\min_{X\succcurlyeq 0}\ \frac{1}{2}\|X-P\|_F^2,$$

for a given matrix $P\in\mathbb{S}^\ell$, which is typically solved via singular value decomposition and eigenvalue clipping [9]. It can be also extended to $\Lambda(X)\in[m,L]$ straightforwardly.

A less obvious instance is the *SD Procrustes problem*:

$$\min_{X\succcurlyeq 0}\ \frac{1}{2}\|TX-P\|_F^2.$$

When $T\in\mathbb{R}^{\ell\times\ell}$ is full rank, the factorization $X=E^\top E$ is exact and can be used to remove the SD constraint [2].

In general, however, the complexity of (1) is dominated by the SD constraint. In this letter, we ask the question of *when we can safely remove* the SD constraint and thereby transform the semidefinite program (SDP) (1) into an easier program frequently turning into a quadratic program (QP), according to the shape of $\mathcal{X}$. This can not be done in general if the SD constraints are added for the purpose of regularizing the problem. However, when A and b stack noisy samples as in (1), the "true" optimal solution $x^\star\in\mathcal{X}$ is the one obtained with non-noisy unbiased (possibly infinite) data samples, and $\Lambda(F(x^\star))\in[m,L]$, then one can safely relax the requirement on the span of $F(x)$, and analyze the distribution of the eigenvalues of $F(x_N^\star)$, where the (random) quantity $x_N^\star$ solves (1) with no SD constraints and N noisy samples.

The question of determining in probability when certain constraints are satisfied, which are so for the true process $x^\star$, goes under the umbrella of concentration inequalities, which offer bounds on how random variables deviate from a certain value (typically, the expectation). Concentration inequalities, also known as tail bounds or finite-sample certificates [10], have a long history in statistical learning—see, e.g., [11, Ch. 1]—and they have been recently used for data-driven control and decision-making [12]–[15] as well as ordinary least squares (LS) [16]. Here, we will be interested in matrix

This work was partially supported by the ANR-JCJC grant ANR-23-CE48-0011-01.

F. Fabiani is with the IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100, Lucca, Italy. A. Simonetto is with the Unité de Mathématiques Appliquées, ENSTA, Institut Polytechnique de Paris, 91120 Palaiseau, France (e-mail: filippo.fabiani@imtlucca.it, andrea.simonetto@ensta.fr).

concentration inequalities aligned to the McDiarmid's one as in [17, §7]. Although other type of concentration inequalities tailored for matrices exist, such as Bernstein [17, §6] [18] or PAC-Bayes [19], [20], none of them captures the deviation between the sampled value and the expected value of a matrix function when evaluated on independent random variables. This is a key aspect in our data-driven framework. Concentration of measures as in [21], instead, go beyond our interest.

Our contributions can then be summarized as follows:

1) We establish a distribution-free concentration bound on the span of $F(x_N^\star)$, where $x_N^\star$ minimizes (1) with no SD constraints, and it is therefore much easier to solve;
2) Based on the previous result, when the original SDLS is designed to learn a quadratic, yet unknown, function, we show how to bound the error between a gradient descent iterate minimizing the surrogate cost obtained with no SD constraints and the true minimizer $x^\star$.

Finally, our theoretical results are corroborated on a numerical example addressing a quadratic function fitting problem.

II. Examples of SDLS problems

To fully motivate the problem addressed in this letter, we discuss next several applications of the SDLS problem in (1).

Example 1 (Fitting a quadratic function [8])**.** *Given an unknown, quadratic, positive SD function, $f(x) \coloneqq x^\top Q x + c^\top x + r$, $Q \succcurlyeq 0$, $c \in \mathbb{R}^n$ and $r \in \mathbb{R}$, the task is to estimate the latter parameters through $(\hat{Q}, \hat{c}, \hat{r})$, thereby producing $\hat{f}(x) \coloneqq x^\top \hat{Q} x + \hat{c}^\top x + \hat{r}$, through measurements $y^{(i)} = f(x^{(i)}) + \eta^{(i)}$ collected at $x^{(i)} \in \mathbb{R}^n$ and affected by noise $\eta^{(i)}$, $i \in \{1, \dots, N\}$. We can formulate the problem as:*

$$\min_{\hat{Q}\succcurlyeq 0, \hat{c}, \hat{r}} \ \frac{1}{N} \sum_{i=1}^{N} (\hat{f}(x^{(i)}) - y^{(i)})^2,$$

which can be readily transformed into:

$$\min_{\xi \in \mathbb{R}^{n^2+n+1}} \ \frac{1}{N} \|A\xi - b\|^2 \ \text{ s.t. } \ F(\xi) \succcurlyeq 0, \tag{2}$$

for a suitable data matrix A, vector b, and LMI F. In general, we may also want to impose $\Lambda(F(\xi)) \in [m, L]$, for all ξ. □

Example 2 (Kernel ridge regression [22])**.** *Akin to the above, we now want to estimate some smooth, strongly convex function $f(x)$ through noisy measurements at $x^{(i)} \in \mathbb{R}^n$, $i \in \{1, \dots, N\}$. By assuming that f belongs to a reproducing kernel Hilbert space (RKHS) with kernel function $k(\cdot, \cdot)$ and bounded RKHS norm, exploiting the reproducing property makes our goal to finding coefficients $\alpha \in \mathbb{R}^N$ so that:*

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^N} \quad & \frac{1}{N} \|\mathrm{K}\alpha - y\|^2 + \rho \|\alpha\|^2 \\ \text{s.t.} \quad & \Lambda([\nabla^2 \kappa(x^s)] \cdot \alpha) \in [m, L], \ s \in \mathcal{S}, \end{aligned} \tag{3}$$

where $\mathrm{K} \in \mathbb{S}^N$ is the so-called Gram matrix, with entries $\mathrm{K}_{i,j} = k(x^{(i)}, x^{(j)})$, and $y = [f(x^{(1)}) + \eta^{(1)} \ \cdots \ f(x^{(N)}) + \eta^{(N)}]^\top$ obtained by means of noisy samples. By letting $\kappa(x) \coloneqq [k(x, x^{(1)}) \ \cdots \ k(x, x^{(N)})]^\top$, in the constraints we then indicate with the "dot" a weighted sum over the third tensor dimension, thereby forcing the eigenvalues of the Hessian associated to the estimate of $f(x)$ within $[m, L]$ at specific points $x^s \in \mathbb{R}^n$, $s \in \mathcal{S}$, to be selected. □

Example 3 (Elasticity and inertia estimation [2], [4])**.** *Given a matrix of data containing applied forces F and one collecting the resulting measured displacements X, one wishes to reconstruct the elasticity matrix K as follows:*

$$\min_{K \succcurlyeq 0} \ \frac{1}{2} \|KX - F\|_F^2.$$

The above can be generalized to estimate the matrix of inertia for rigid bodies, such as satellites, with angular matrix M and torque matrix T. For a suitable matrix R, we then have:

$$\min_{J \succ 0} \ \frac{1}{2} \|JM - T\|_F^2 \ \text{ s.t. } \ RJ \geq 0.$$ □

Example 4 (Covariance fitting [7])**.** *Given a stochastic linear system $\dot{x} = Ax + Bd$, with $A \in \mathbb{R}^{n \times n}$ being a Hurwitz matrix, and $B \in \mathbb{R}^{n \times m}$ mapping disturbance $d \in \mathbb{R}^m$ over the state, the LS covariance fitting problem can be formulates as:*

$$\begin{aligned} \min_{X \succcurlyeq 0, H} \quad & \frac{1}{2} \|X - \Sigma\|_F^2 \\ \text{s.t.} \quad & AX + XA^\top = -(BH + H^\top B^\top), \end{aligned}$$

where $\Sigma = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} {x^{(i)}}^\top \succcurlyeq 0$ is the sample covariance matrix obtained through N-state measurements $x^{(i)} \in \mathbb{R}^n$. □

III. Finite-sample guarantees for SDLS

Next, we establish the main result of our paper, i.e., we derive finite-sample guarantees on the spectrum of $F(x)$ in (1) when the constraints $\Lambda(F(x)) \in [m, L]$ is relaxed. To this end, we will consider a simplified version of (1) that does not include any SD-type of constraints $\Lambda(F(x))$, namely:

$$\min_{x \in \mathcal{X}} \ \frac{1}{N} \|Ax - b\|^2 + \rho \|x\|^2. \tag{4}$$

Let $x_N^\star$ denote the resulting optimal solution, where the subscript emphasizes its data-driven nature, i.e., $x_N^\star = x_N^\star(z^{(1)}, \dots, z^{(N)})$. Note that $x_N^\star$ is inherently random according to the *unknown* distribution $\mathbb{P}$ underlying data. Our goal is then to establish probabilistic guarantees on how the span of $F(x_N^\star)$ concentrates as N grows. To achieve our goal, we postulate the following working assumptions:

Standing Assumption 1. *The SDLS in* (1)*, and hence* (4)*, relies on N independent and identically distributed (i.i.d.) samples $\{z^{(i)}\}_{i=1}^N$ drawn according to some unknown probability distribution $\mathbb{P}$, and appropriately stacked into matrix A and vector b. The data is noisy but unbiased, and the underlying true process $x^\star$, for which we collect samples $\{z^{(i)}\}_{i=1}^N$ and that we want to estimate, verifies all the constraints in* (1)*.* □

Armed with these, we can therefore prove what follows:

Theorem 1. *Fix $\delta \in (0, 1)$ and $\rho > 0$. Then, there exists $\varepsilon = \varepsilon(\ell, \delta, N) > 0$ such that, with probability at least $1 - \delta$,*

$$\Lambda(F(x_N^\star)) \in [m - \varepsilon, \ L + \varepsilon]\,.$$ □

Proof. We use $F(\cdot)$ as an application mapping the N data via $x_N^\star$ into a symmetric matrix of dimension ℓ, i.e., $F(x_N^\star) = F(x_N^\star(z^{(1)}, \dots, z^{(N)}))$, and then apply [17, Cor. 7.5].

To this end, we first need to prove the bounded difference property for $F(x_N^\star) - F(x_{N^i}^\star) = F(x_N^\star(z^{(1)},\dots,z^{(N)})) - F(x_N^\star(z^{(1)},\dots,z^{(i-1)},z',z^{(i+1)},\dots,z^{(N)}))$, where z' is some sample drawn according to $\mathbb{P}$, i.i.d. with respect to (w.r.t.) the dataset $\{z^{(i)}\}_{i=1}^N$, which replaces the arbitrary i-th one. Specifically, we have to identify matrices $\Xi_i \in \mathbb{S}^\ell$ so that $(F(x_N^\star) - F(x_{N^i}^\star))^2 \preccurlyeq \Xi_i^2$, which allow us to define $\sigma^2 = \|\sum_{i=1}^N \Xi_i^2\|$. By analyzing $(F(x_N^\star) - F(x_{N^i}^\star))^2$ more in detail we obtain:

$$
\begin{aligned}
(F(x_N^\star) - F(x_{N^i}^\star))^2 &= \left(\left((x_N^\star - x_{N^i}^\star)^\top \otimes I_\ell\right) \begin{bmatrix} F_1 \\ \vdots \\ F_n \end{bmatrix} \right)^2 \\
&\overset{(a)}{=} \left((x_N^\star - x_{N^i}^\star)^\top \otimes I_\ell\right) \underbrace{\begin{bmatrix} F_1 \\ \vdots \\ F_n \end{bmatrix} \begin{bmatrix} F_1 & \dots & F_n \end{bmatrix}}_{=:H} \left((x_N^\star - x_{N^i}^\star) \otimes I_\ell\right) \\
&\overset{(b)}{\preccurlyeq} \lambda_{\max}(H) \left((x_N^\star - x_{N^i}^\star)^\top \otimes I_\ell\right) \left((x_N^\star - x_{N^i}^\star) \otimes I_\ell\right) \\
&\overset{(c)}{=} \lambda_{\max}(H) \|x_N^\star - x_{N^i}^\star\|^2 I_\ell,
\end{aligned}
$$

where (a) follows from the symmetry of $F(x)$, for all $x \in \mathbb{R}^n$, (b) since $H \in \mathbb{S}^{\ell n}_{\succcurlyeq 0}$ by construction, while (c) from standard properties of the Kronecker product. To upper bound the term $\|x_N^\star - x_{N^i}^\star\|$, we rely on admissibility-type of arguments from [23]. In particular, by denoting with $x_{N-1}^\star$ the optimal solution to (4) once removed an arbitrary sample from the N available $\{z^{(i)}\}_{i=1}^N$, direct application of [23, Lemma 21] to $\|x_N^\star - x_{N^i}^\star\| \leq \|x_N^\star - x_{N-1}^\star\| + \|x_{N^i}^\star - x_{N-1}^\star\|$ leads to $\|x_N^\star - x_{N^i}^\star\|^2 \leq 2B^2/\rho^2 N^2$. Here, B denotes some term upper bounding both x and the data-based quantities (A, b) that one can obtain, which happens to be finite in view of Standing Assumption 1. Specifically, it follows from i) the well-posedness of the learning problem, which rules out the possibility to collect possibly noisy samples that are unbounded, and ii) the fact that (4) contains a regularization term singling out a solution even if the underlying feasible set $\mathcal{X}$ may be unbounded. Thus, we obtain $\sigma = \sqrt{\|\sum_{i=1}^N \frac{2B^2\lambda_{\max}(H)}{\rho^2 N^2} I_\ell\|} = \frac{B}{\rho}\sqrt{\frac{2\lambda_{\max}(H)}{N}}$.

Then, [17, Cor. 7.5] establishes the following relation:

$$
\mathbb{P}^N \left\{\lambda_{\max}(F(x_N^\star) - \mathbb{E}_{\mathbb{P}}\left[F(x_N^\star)\right]) \leq \epsilon\right\} \geq 1 - \ell e^{-\epsilon^2/8\sigma^2}. \quad (5)
$$

Since we have assumed unbiasedness, from the above we directly obtain: $x^\star = \mathbb{E}_{\mathbb{P}}\left[x_N^\star\right] \implies F(x^\star) = F(\mathbb{E}_{\mathbb{P}}\left[x_N^\star\right]) = \mathbb{E}_{\mathbb{P}}\left[F(x_N^\star)\right]$ in view of the linearity of $F(\cdot)$. With this,

$$
\begin{aligned}
&\lambda_{\max}(F(x_N^\star) - \mathbb{E}_{\mathbb{P}}\left[F(x_N^\star)\right]) \leq \epsilon \\
&\quad \iff \lambda_{\max}(F(x_N^\star) - F(x^\star)) \leq \epsilon \\
&\quad \iff F(x_N^\star) - F(x^\star) \preccurlyeq \epsilon I_\ell \\
&\quad \iff F(x_N^\star) \preccurlyeq \epsilon I_\ell + F(x^\star) \\
&\quad \implies F(x_N^\star) \preccurlyeq (\epsilon + L) I_\ell \implies \lambda_{\max}(F(x_N^\star)) \leq \epsilon + L.
\end{aligned} \quad (6)
$$

Then, by deriving ϵ from the relation $\delta = \ell e^{-\epsilon^2/8\sigma^2}$, we end up with the following expression:

$$
\begin{aligned}
&\mathbb{P}^N \left\{\lambda_{\max}(F(x_N^\star)) \leq L + 2\sigma\sqrt{2\ln(\ell/\delta)}\right\} \geq 1 - \delta, \text{ i.e.,} \\
&\mathbb{P}^N \left\{\lambda_{\max}(F(x_N^\star)) \leq L + \tfrac{4B}{\rho\sqrt{N}}\sqrt{\lambda_{\max}(H)\ln(\ell/\delta)}\right\} \geq 1 - \delta,
\end{aligned}
$$

which holds true for arbitrarily high confidence $\delta \in (0, 1)$.

To derive the relation for the "opposite direction" involving $\lambda_{\min}(F(x_N^\star))$, from the proof of [17, Cor. 7.5] one is able to obtain also the following concentration bound:

$$
\mathbb{P}^N \left\{\lambda_{\max}(\mathbb{E}_{\mathbb{P}}\left[F(x_N^\star)\right] - F(x_N^\star)) \leq \epsilon\right\} \geq 1 - \ell e^{-\epsilon^2/8\sigma^2}.
$$

Then, we proceed as in (6), with the opposite signs to arrive at:

$$
\mathbb{P}^N \left\{\lambda_{\min}(F(x_N^\star)) \geq m - \epsilon\right\} \geq 1 - \ell e^{-\epsilon^2/8\sigma^2}.
$$

Again, by obtaining ϵ from $\delta = \ell e^{-\epsilon^2/8\sigma^2}$, and putting everything together we finally arrive at the following expression:

$$
\begin{aligned}
\mathbb{P}^N \Big\{\Lambda(F(x_N^\star)) \in \Big[&m - \tfrac{4B}{\rho\sqrt{N}}\sqrt{\lambda_{\max}(H)\ln(\ell/\delta)}, \\
&L + \tfrac{4B}{\rho\sqrt{N}}\sqrt{\lambda_{\max}(H)\ln(\ell/\delta)}\Big]\Big\} \geq 1 - \delta.
\end{aligned}
$$

Setting $\varepsilon = \frac{4B}{\rho\sqrt{N}}\sqrt{\lambda_{\max}(H)\ln(\ell/\delta)}$ concludes the proof. ■

As a crucial consequence of Theorem 1, instead of solving the SDLS in (1), one can then focus on the program in (4), which is easier to solve as it frequently reduces to a QP, according to the shape of $\mathcal{X}$, and still obtain finite-sample guarantees that the spectrum of $F(x_N^\star)$ will span a neighborhood of the desired one, i.e., $[m, L]$, which shrinks as N grows. The established bound is easy-to-compute, distribution-free, and only requires a dataset $\{z^{(i)}\}_{i=1}^N$ with i.i.d. samples.

IV. Optimization over learned functions

Focusing on Example 1, we now want to bound the distance between $x^\star = \operatorname{argmin}_{x\in\mathbb{R}^n} f(x)$, where $f(x) = x^\top Q x + c^\top x + r$ is the unknown function we wish to learn, and the generic iterate x_{k+1} of the gradient descent scheme applied to $\hat{f}(x) = x^\top \hat{Q}_N^\star x + \hat{c}_N^{\star\top} x + \hat{r}_N^\star$. The latter function is obtained, instead, by solving the variant of (2) without the constraints $\Lambda(F(\xi)) \in [m, L]$, yielding $\xi_N^\star = [\text{vec}(\hat{Q}_N^\star)^\top \ \hat{c}_N^{\star\top} \ \hat{r}_N^\star]^\top$.

We note that without imposing suitable constraints on the eigenvalues of the Hessian $\hat{Q} \succcurlyeq 0$ does not ensure the convergence for the gradient descent scheme when applied to minimize $\hat{f}(x)$. Then, let

$$
x_{k+1} = x_k - \gamma \nabla \hat{f}(x_k) = x_k - \gamma(\hat{Q}_N^\star x_k + \hat{c}_N^\star), \quad (7)
$$

denoting the generic k-th update of the gradient descent method, for given stepsize $\gamma > 0$ and initial condition $x_0 \in \mathbb{R}^n$. We can claim the following bound based on Theorem 1:

Theorem 2. *Fix $\delta \in (0, 1)$ and $\rho > 0$. Assume that $x^\star$ exists and is bounded. Then, there exists $\varepsilon = \varepsilon(\ell, \delta, N) > 0$ such that, if we select a N that guarantees $m - \varepsilon > 0$ with probability $1 - \delta$ and we select $\gamma < 2/(L + \varepsilon)$, then with the same probability, the iterations* (7) *converge as follows:*

$$
\limsup_{k\to\infty} \|x_{k+1} - x^\star\| = O(\varepsilon). \qquad \square
$$

Proof. By analyzing the iteration error, we readily obtain:

$$
\begin{aligned}
x_{k+1} - x^\star &= x_k - \gamma(\hat{Q}_N^\star x_k + \hat{c}_N^\star) - x^\star \\
&= x_k - \gamma(\hat{Q}_N^\star x_k + \hat{c}_N^\star) - x^\star + \gamma(Qx^\star + c) \\
&= (I - \gamma\hat{Q}_N^\star)(x_k - x^*) + \gamma(Q - \hat{Q}_N^\star)x^\star + \gamma(c - \hat{c}_N^\star).
\end{aligned}
$$

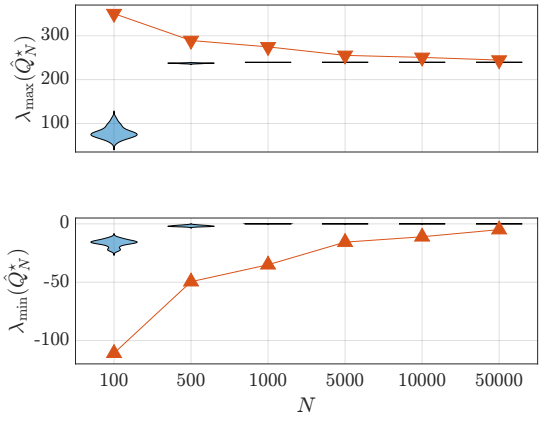

Fig. 1. Violin plots reporting the maximum (top figure) and minimum (bottom figure) eigenvalues of $\hat{Q}_N^\star$ obtained by solving (2), averaged over 20 trials with datasets of different size N. The red downward (respectively, upward)-pointing triangles denote the upper (resp., lower) bound in Theorem 1.

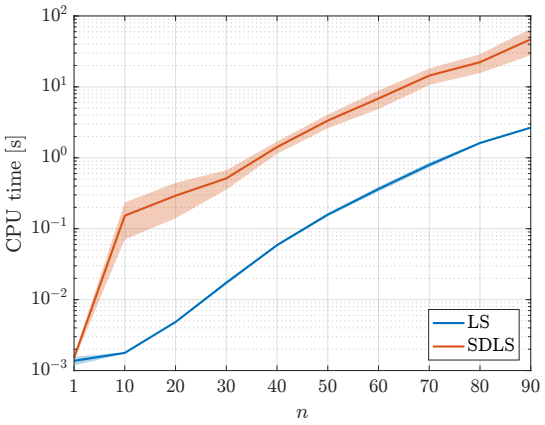

Fig. 2. Computational time for solving the LS in (2) and related SDLS variant, averaged over 20 different numerical instances.

By choosing γ and N as specified, we impose that the matrix $(1-\gamma\hat{Q}_N^\star)$ is smaller than one in norm, with probability at least $1-\delta$. Furthermore, the error terms $Q-Q_N^\star$ and $c-\hat{c}_N^\star$ can be bounded following the proof of Theorem 1. The former is immediate leading to $\|Q-\hat{Q}_N^\star\| \leq \varepsilon$; the latter is also easy by looking at the diagonal map $F' := \mathrm{diag}(c)$ and using the same steps to obtain $\|c-\hat{c}_N^\star\| \leq \varepsilon$. With this, and with probability at least $1-\delta$,

$$\|x_k - x^\star\| \leq \|I-\gamma\hat{Q}_N^\star\|^k \|x_0 - x^\star\| + O(\varepsilon),$$

and the theorem is proven. ■

The theorem ensures convergence of the sampled gradient method to an error ball of size proportional to $\varepsilon \propto \frac{1}{\sqrt{N}}\sqrt{\ln(\ell/\delta)}$. The more the points N, the smaller the error.

V. Numerical Experiments

We now corroborate our theoretical results on a numerical instance of Example 1. The simulations are run in MATLAB on a laptop with an Apple M2 chip featuring an 8-core CPU and 16 GB RAM, while Mosek [24] has been used as SDP solver, implemented in YALMIP environment [25].

Then, we randomly generate a quadratic function $f(x)$ that we wish to learn. By denoting with $\mathcal{U}([a,b])$ the uniform distribution over the interval $[a,b]$, we set $Q = U^\top U \succcurlyeq 0$, with each entry $U_{ij} \sim \mathcal{U}([0,1])$, $c \sim \mathcal{U}([0,1]^n)$, and $r \sim \mathcal{U}([0,1])$.

First, we test how the bound established in Theorem 1 changes by considering a dataset of different size N. In particular, the latter consists of pairs $(x^{(i)}, f(x^{(i)}) + \eta^{(i)})$ with each $x^{(i)} \in \mathcal{U}([-10,10]^n)$ and $\eta^{(i)} \sim \mathcal{N}(0,1)$, where $\mathcal{N}(\mu,\sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Matrix A and vector b in (2) then reads as follows:

$$A = \begin{bmatrix} x^{(1)\top} \otimes x^{(1)\top} & x^{(1)\top} & 1 \\ \vdots & & \\ x^{(N)\top} \otimes x^{(N)\top} & x^{(N)\top} & 1 \end{bmatrix}, \; b = \begin{bmatrix} f(x^{(1)}) + \eta^{(1)} \\ \vdots \\ f(x^{(N)}) + \eta^{(N)} \end{bmatrix}.$$

Notice that $\xi = [\mathrm{vec}(\hat{Q})^\top \; \hat{c}^\top \; \hat{r}]^\top \in \mathbb{R}^{n^2+n+1}$. Since we will solve (2) by adding an extra regularization term $\rho\|\xi\|^2$ in the cost and imposing symmetry only, i.e., simple equality constraints $\hat{Q} = \hat{Q}^\top$ and no SD ones, i.e., $\hat{Q} \succcurlyeq 0$ and $\hat{Q} \preccurlyeq \lambda_{\max}(Q)$, we are therefore primarily interested in guaranteeing bounds $\lambda_{\min}(\hat{Q}_N^\star) \geq -\frac{4B}{\rho\sqrt{N}}\sqrt{n\ln(n/\delta)}$ and $\lambda_{\max}(\hat{Q}_N^\star) \leq \lambda_{\max}(Q) + \frac{4B}{\rho\sqrt{N}}\sqrt{n\ln(n/\delta)}$ with probability at least $1-\delta$, since it turns out that, in the considered setting, $\ell = n$ and $\lambda_{\max}(H) = n$. To this end, Fig. 1 allows one to contrast the theoretical bound established in Theorem 1 and the maximum/minimum eigenvalues of $\hat{Q}_N^\star$ obtained by solving the QP in (2). In this case, the simulation is run by setting $\rho = 1$, $n = 30$, $\delta = 0.05$, and for each N averaged over 20 different dataset $\{(x^{(i)}, f(x^{(i)}) + \eta^{(i)})\}_{i=1}^N$. While one can observe a sort of monotonic behavior for the theoretical bounds, the computed $\lambda_{\max}(\hat{Q}_N^\star)$ and $\lambda_{\min}(\hat{Q}_N^\star)$ instead do not follow a specific trend, other than producing better (i.e., close to $240.14 = \lambda_{\max}(Q)$ and $0 = \lambda_{\min}(Q)$, respectively) estimates with very little variance for larger datasets.

Next, we set $N = 5000$ and compare the computational time required to solve the QP in (2), and the related SDLS variant including the SD constraint $0 \preccurlyeq \hat{Q} \preccurlyeq \lambda_{\max}(Q)$, for different values of n. Note that we actually solve (SD)LS from 3 up to 8191 decision variables. As illustrated in Fig. 2, solving the LS problem only requires at least an order of magnitude less than solving the associated SDLS.

VI. Conclusion

By focusing on data-driven LS problems with SD constraints, we have derived probabilistic guarantees holding with high confidence on the spectrum of their optimal solutions once these constraints are removed. Our certificate allows one to solve a simple program, which frequently turns into a QP, in place of the full SDLS problem, while ensuring that the span of the resulting solution is ε-close to that enforced by the SD constraints. Consistently, our bound shrinks as the number of data increases, it is distribution-free and only requires i.i.d. samples. As a consequence, when SDLS is designed to learn an unknown function, we have shown how to bound the error between a gradient descent iterate minimizing the surrogate cost obtained with no SD constraints and the true minimizer.

Along the line of the results in §IV, future work will explore the link between these bounds and data-driven optimization.

REFERENCES

[1] J. E. Brock, "Optimal matrices describing linear systems.," *AIAA Journal*, vol. 6, no. 7, pp. 1292–1296, 1968.

[2] K. G. Woodgate, "Efficient stiffness matrix estimation for elastic structures," *Computers & Structures*, vol. 69, no. 1, pp. 79–84, 1998.

[3] N. G. B. Krislock, *Numerical solution of semidefinite constrained least squares problems*. PhD thesis, University of British Columbia, 2003.

[4] Z. R. Manchester and M. A. Peck, "Recursive inertia estimation with semidefinite programming," in *AIAA Guidance, Navigation, and Control Conference*, p. 1902, 2017.

[5] J. Malick, "A dual approach to semidefinite least-squares problems," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 1, pp. 272–284, 2004.

[6] S. Boyd and L. Xiao, "Least-squares covariance matrix adjustment," *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 2, pp. 532–546, 2005.

[7] F. Lin and M. R. Jovanovic, "Least-squares approximation of structured covariances," *IEEE Transactions on Automatic Control*, vol. 54, no. 7, pp. 1643–1648, 2009.

[8] I. Notarnicola, A. Simonetto, F. Farina, and G. Notarstefano, "Distributed personalized gradient tracking with convex parametric models," *IEEE Transactions on Automatic Control*, vol. 68, no. 1, pp. 588–595, 2022.

[9] N. J. Higham, "Computing a nearest symmetric positive semidefinite matrix," *Linear Algebra and its Applications*, vol. 103, pp. 103–118, 1988.

[10] M. Krikheli and A. Leshem, "Finite sample performance of linear least squares estimation," *Journal of the Franklin Institute*, vol. 358, no. 15, pp. 7955–7991, 2021.

[11] F. Bach, *Learning theory from first principles*. MIT press, 2024.

[12] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International conference on machine learning*, pp. 1225–1234, PMLR, 2016.

[13] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite-sample perspective," *IEEE Control Systems Magazine*, vol. 43, no. 6, pp. 67–97, 2023.

[14] L. Ye, M. Chi, R. Liao, and V. Gupta, "Learning decentralized linear quadratic regulators with regret," *SIAM Journal on Control and Optimization*, vol. 62, no. 6, pp. 3341–3368, 2024.

[15] F. Fabiani and B. Franci, "Finite-sample guarantees for data-driven forward-backward operator methods," *IEEE Transactions on Automatic Control*, 2025. (Under review).

[16] O. Bousquet and A. Elisseeff, "Algorithmic stability and generalization performance," *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[17] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, pp. 389–434, 2012.

[18] R. I. Oliveira, "Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges," *arXiv preprint arXiv:0911.0600*, 2009.

[19] O. Catoni, "PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design," *arXiv preprint arXiv:1603.05229*, 2016.

[20] S. Minsker, "Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2871–2903, 2018.

[21] R. Couillet and Z. Liao, *Random matrix methods for machine learning*. Cambridge University Press, 2022.

[22] P.-C. Aubin-Frankowski and Z. Szabó, "Hard shape-constrained kernel machines," *Advances in Neural Information Processing Systems*, vol. 33, pp. 384–395, 2020.

[23] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[24] MOSEK ApS, *The MOSEK optimization toolbox for MATLAB. Version 11.0*, 2025.

[25] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proceedings of the CACSD Conference*, 2004.