



First Order Methods For Globally Optimal Distributed Controllers Beyond Quadratic Invariance

Working Paper

Author(s):

Furieri, Luca ; Kamgarpour, Maryam 

Publication date:

2019-08-21

Permanent link:

<https://doi.org/10.3929/ethz-b-000359773>

Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

First Order Methods For Distributed Control: Unique Stationarity Beyond Quadratic Invariance

Luca Furieri and Maryam Kamgarpour

Abstract—We study the distributed Linear Quadratic Gaussian (LQG) control problem in finite-horizon, where the controller depends linearly on the history of the outputs and it is required to possess a given sparsity pattern. It is well-known that a convex formulation of this problem in terms of disturbance-feedback control policies is only possible when a condition known as Quadratic Invariance (QI) holds. In this paper, we first show that given QI sparsity constraints the LQG problem is solved to global optimality with a first-order steepest descent method in the domain of output-feedback controllers. Convergence is obtained despite non-convexity of the problem with respect to output-feedback controllers. Second, we characterize a class of Uniquely Stationary (US) problems, for which first-order methods are guaranteed to converge to a global optimum. We show that the class of US problems is strictly larger than that of strongly QI problems, and that it is not included in that of QI problems. Finally, we develop a tractable test for the US property. Our test provides an a priori global optimality certificate for a class of problems which are not QI.

I. INTRODUCTION

The safe and efficient operation of emerging networked dynamical systems, such as the smart grid and autonomous vehicles, relies on the decision making of multiple interacting agents. Controlling these systems optimally is challenged by an inherent lack of information about the systems' internal variables, possibly due to privacy concerns, geographic distance or the high cost of implementing a reliable communication network. The classical works of [1], [2] highlighted that, given information constraints, even simple instances of the Linear Quadratic Gaussian (LQG) control problem can result in highly intractable optimization tasks.

A vast amount of literature has focused on approaching the distributed LQG problem and its variants with convex programming in the Youla parameter [3]. This enables utilizing efficient off-the-shelf software for numerical computation. A main challenge inherent to this approach is that the distributed control problem admits an exact convex reformulation if and only if the information constraints and the system dynamics interact in a Quadratically Invariant (QI) manner [4], [5]. This limitation severely reduces the class of problems for which a certificate of global optimality is available. For this reason, a variety of approximation methods and alternative controller implementations have been devised to partially deal with the non-QI cases, based both on convex

programming and nonlinear optimization. We refer the reader to [6]–[11] for a collection of recent results.

The past few years have witnessed a rapid growth of interest in developing learning-based, model-free techniques for optimal control problems. Specifically, some scenarios envision a system that is completely unknown, for which an optimal behavior is obtained by observing the system response to different controllers and iteratively improving the control policy. Since operating within the Youla domain is impractical due to the system and thus the exact mapping to the Youla parameters being unknown, these scenarios motivate computing controllers in a direct way, for instance by devising first-order, gradient-descent based methods. Convergence of these methods to a global optimum was recently proven for the LQR problem in the non-distributed case [12]–[15]. When carrying on these methods to the distributed controller case, however, one can in general only guarantee convergence to a stationary point, which may not be a global optimum [15], [16]. For the infinite-horizon and static-controller cases, this is mainly due to the set of stabilizing distributed controllers being disconnected in general [17]. To the best of the authors' knowledge, classes of distributed control problems solvable to global optimality with first-order methods are yet to be characterized, and a connection with the QI notion is yet to be established. Furthermore, a condition that is more general than QI for global optimality certificates has not been found yet. Indeed, the QI notion is closely linked to using convex programming, whereas we show that more general global optimality conditions can be found with first-order methods.

Motivated as per above, in this paper we study first-order methods to solve the distributed LQG problem with dynamic controllers in finite-horizon. Our contributions are as follows. First, we show that given QI sparsity constraints, one can use first-order methods to find a globally optimal output-feedback controller directly, as an alternative to solving a convex program in the disturbance-feedback (Youla) domain. This will enable devising learning-based policy gradient methods for distributed control in future works. Second, we characterize a new class of Uniquely Stationary (US) control problems, which can be solved to global optimality using first-order methods. We show that every strongly QI problem is US and that there are instances of US problems which are not QI. This implies that first-order methods enable global optimality certificates strictly beyond QI. We conclude the paper by discussing tests for the US property.

The paper is structured as follows. Section II introduces the necessary notation and background. Section III contains

This research was gratefully funded by the European Union ERC Starting Grant CONENE.

Luca Furieri and Maryam Kamgarpour are with the Automatic Control Laboratory, Department of Information Technology and Electrical Engineering, ETH Zürich, Switzerland. E-mails: {furieri, lkamgar}@control.ee.ethz.ch

our first result about global optimality given strong QI and a numerical example. Section IV characterizes the class of US problems and establishes the result about certificates of global optimality beyond QI. We conclude the paper in Section V.

II. BACKGROUND AND PROBLEM STATEMENT

We start this section by providing the necessary notation. We then proceed with stating the distributed LQG problem and reviewing results about disturbance-feedback control policies and quadratic invariance.

A. Notation

We use \mathbb{R} to denote the set of real numbers. The (i, j) -th element in a matrix $Y \in \mathbb{R}^{m \times n}$ is referred to as $Y_{i,j}$. We use I_n to denote the identity matrix of size $n \times n$, $0_{m \times n}$ to denote the zero matrix of size $m \times n$. Whenever the subscripts are omitted, the dimensions are implied by the context. We write $M = \text{blkdg}(M_1, \dots, M_n)$ to denote a block-diagonal matrix where the blocks are the matrices M_1, \dots, M_n . For a symmetric matrix $M = M^\top$ we write $M \succ 0$ (resp. $M \succeq 0$) if and only if it is positive definite (resp. positive semidefinite), that is its eigenvalues are strictly positive (resp. non-negative). For two matrices M, P of any dimensions $M \otimes P$ denotes the Kronecker product and for two matrices of equal dimensions $M \odot P$ denotes the Hadamard product¹. For any matrix $K \in \mathbb{R}^{m \times n}$, $\text{vec}(K) \in \mathbb{R}^{mn}$ is a vector obtained by stacking the columns of K into a single column. Given a binary matrix $X \in \{0, 1\}^{m \times n}$, we define the associated *sparsity subspace* as

$$\text{Sparse}(X) := \{Y \mid Y_{i,j} = 0 \text{ for all } i, j \text{ such that } X_{i,j} = 0\}.$$

Similarly, given $Y \in \mathbb{R}^{m \times n}$, we define $X = \text{Struct}(Y)$ as the binary matrix given by

$$X_{i,j} := \begin{cases} 0 & \text{if } Y_{i,j} = 0 \\ 1 & \text{otherwise.} \end{cases}$$

Let $X, \hat{X} \in \{0, 1\}^{m \times n}$ and $Z \in \{0, 1\}^{n \times p}$ be binary matrices. We adopt the following conventions: $X + \hat{X} := \text{Struct}(X + \hat{X})$, $XZ := \text{Struct}(XZ)$, $X \leq \hat{X}$ if and only if $X_{i,j} \leq \hat{X}_{i,j} \forall i, j$. The Euclidean norm of a vector $v \in \mathbb{R}^n$ is denoted by $\|v\|_2^2 = v^\top v$ and the Frobenius norm of a matrix $M \in \mathbb{R}^{m \times n}$ is denoted by $\|M\|_F^2 = \text{Trace}(M^\top M)$. Given a matrix $K \in \mathbb{R}^{m \times n}$ and a continuously differentiable function $J : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ we define $\nabla J(K)$ as the $m \times n$ matrix such that

$$\nabla J(K)_{i,j} = \frac{\partial J(K)}{\partial K_{i,j}}.$$

For a vector $v \in \mathbb{R}^n$ and a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ we denote the gradient by $\nabla f(v) \in \mathbb{R}^n$ and the Hessian by $\nabla^2 f(v) \in \mathbb{R}^{n \times n}$. Given a subspace $\mathcal{K} \subseteq \mathbb{R}^{m \times n}$ we denote its orthogonal complement as \mathcal{K}^\perp . The symbol $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution with expected value $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n} \succ 0$, and $x \sim \mathcal{N}(\mu, \Sigma)$

¹ $(M \odot P)_{i,j} = M_{i,j} P_{i,j}$

means that the random vector $x \in \mathbb{R}^n$ follows the distribution $\mathcal{N}(\mu, \Sigma)$.

B. Problem Setup

We consider time-varying linear systems in discrete-time

$$\begin{aligned} x_{t+1} &= A_t x_t + B_t u_t + w_t, \\ y_t &= C_t x_t + v_t, \end{aligned} \quad (1)$$

where $x_t \in \mathbb{R}^n$ is the system state at time t affected by additive noise $w_t \sim \mathcal{N}(0, \Sigma_t^w)$ with $x_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, $y_t \in \mathbb{R}^p$ is the output at time t affected by additive noise $v_t \sim \mathcal{N}(0, \Sigma_t^v)$ and $u_t \in \mathbb{R}^m$ is the control input at time t . We consider the evolution of (1) in finite-horizon for $t = 0, \dots, N$, where $N \in \mathbb{N}$. By defining the matrices $\mathbf{A} = \text{blkdg}(A_0, \dots, A_N)$,

$$\mathbf{B} = \begin{bmatrix} \text{blkdg}(B_0 \dots B_{N-1}) \\ 0_{n \times mN} \end{bmatrix}, \mathbf{C} = \begin{bmatrix} \text{blkdg}(C_0 \dots C_{N-1})^\top \\ 0_{n \times pN} \end{bmatrix}^\top,$$

and the vectors $\mathbf{x} = (x_0^\top, \dots, x_N^\top)^\top \in \mathbb{R}^{n(N+1)}$, $\mathbf{y} = (y_0^\top, \dots, y_{N-1}^\top)^\top \in \mathbb{R}^{pN}$, $\mathbf{u} = (u_0^\top, \dots, u_{N-1}^\top)^\top \in \mathbb{R}^{mN}$, $\mathbf{w} = (w_0^\top, w_0^\top, \dots, w_{N-1}^\top) \in \mathbb{R}^{n(N+1)}$ and $\mathbf{v} = (v_0^\top, \dots, v_{N-1}^\top) \in \mathbb{R}^{pN}$, and the shift matrix

$$\mathbf{Z} = \begin{bmatrix} 0_{n \times n} & & & & \\ & I_n & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I_n & 0_{n \times n} \end{bmatrix} \in \mathbb{R}^{n(N+1) \times n(N+1)},$$

we can write the system (1) compactly as

$$\mathbf{x} = \mathbf{P}_{11} \mathbf{w} + \mathbf{P}_{12} \mathbf{u}, \quad \mathbf{y} = \mathbf{C} \mathbf{x} + \mathbf{v}, \quad (2)$$

where $\mathbf{P}_{11} = (\mathbf{I} - \mathbf{Z} \mathbf{A})^{-1}$ and $\mathbf{P}_{12} = (\mathbf{I} - \mathbf{Z} \mathbf{A})^{-1} \mathbf{Z} \mathbf{B}$. In this paper we consider causal output-feedback strategies of the form

$$\mathbf{u} = \mathbf{K} \mathbf{y}, \quad \mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K},$$

where $\mathbf{S} \in \{0, 1\}^{mN \times pN}$ and $\mathcal{K} \subseteq \mathbb{R}^{mN \times pN}$ is a subspace. By denominating the $m \times p$ block matrices inside \mathbf{K} as follows:

$$\mathbf{K} = \begin{bmatrix} K_{0,0} & 0_{m \times p} & 0_{m \times p} \\ \vdots & \ddots & 0_{m \times p} \\ K_{N-1,0} & \dots & K_{N-1,N-1} \end{bmatrix}, \quad (3)$$

a sparsity constraint $\mathbf{K} \in \text{Sparse}(\mathbf{S})$ is equivalent to $K_{i,j} \in \text{Sparse}(S_{i,j})$ for every $i = 0, \dots, N-1$, $j \leq i$, where the matrix $S_{i,j}$ is the corresponding matrix block of \mathbf{S} and encodes those outputs from the time instant $j < i$ that are known to controllers at time i . In this sense, \mathbf{S} expresses spatio-temporal constraints for distributed control. The subspace \mathcal{K} can encode additional general subspace constraints, such as for instance $K_{i,i} = K_{j,j}$ for every $i, j = 0, \dots, N-1$.

Our goal is to compute $\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ that minimizes a finite-horizon cost

$$J(\mathbf{K}) := \mathbb{E}_{\mathbf{w}, \mathbf{v}} \left(\sum_{t=0}^{N-1} x_t^\top M_t x_t + u_t^\top R_t u_t \right) + x_N^\top M_N x_N, \quad (4)$$

where $M_t \succeq 0$ and $R_t \succ 0$ for every t .

Remark 1: When \mathbf{K} is not constrained, the problem of minimizing (4) is known as the Linear Quadratic Gaussian (LQG) problem. The infinite-horizon LQG problem is that of minimizing (4) for $N \rightarrow \infty$, where usually it is required that the controller is static in the sense that $u_t = Ky_t$ for some $K \in \mathbb{R}^{m \times p}$ for every t . In this paper, we address the distributed and finite-horizon version of the LQG problem, where the controller \mathbf{K} is dynamic in the sense that it is allowed to depend on the full history of the outputs. As will be clarified later in this section, considering dynamic controllers allows to link the quadratic invariance notion [4] with first-order methods.

By defining $\mathbf{M} = \text{blkdg}(M_0, M_1, \dots, M_N) \in \mathbb{R}^{n(N+1) \times n(N+1)}$, $\mathbf{R} = \text{blkdg}(R_0, \dots, R_{N-1}) \in \mathbb{R}^{mN \times mN}$, $\Sigma_w = \text{blkdg}(\Sigma_0, \Sigma_0^w, \dots, \Sigma_{N-1}^w)$, $\Sigma_v = \text{blkdg}(\Sigma_0^v, \dots, \Sigma_{N-1}^v)$, $\mu_w = [\mu_0^T \ 0 \ \dots \ 0]^T$ the cost function (4) can be written as

$$\begin{aligned} J(\mathbf{K}) = & \left\| \mathbf{M}^{\frac{1}{2}} (I - \mathbf{P}_{12} \mathbf{K} \mathbf{C})^{-1} \mathbf{P}_{11} \Sigma_w^{\frac{1}{2}} \right\|_F^2 + \\ & + \left\| \mathbf{M}^{\frac{1}{2}} \mathbf{P}_{12} \mathbf{K} (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1} \Sigma_v^{\frac{1}{2}} \right\|_F^2 + \\ & + \left\| \mathbf{R}^{\frac{1}{2}} \mathbf{K} (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1} \mathbf{C} \mathbf{P}_{11} \Sigma_w^{\frac{1}{2}} \right\|_F^2 + \\ & + \left\| \mathbf{R}^{\frac{1}{2}} \mathbf{K} (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1} \Sigma_v^{\frac{1}{2}} \right\|_F^2 + \\ & + \left\| \mathbf{M}^{\frac{1}{2}} (I - \mathbf{P}_{12} \mathbf{K} \mathbf{C})^{-1} \mathbf{P}_{11} \mu_w \right\|_2^2 + \\ & + \left\| \mathbf{R}^{\frac{1}{2}} \mathbf{K} (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1} \mathbf{C} \mathbf{P}_{11} \mu_w \right\|_2^2. \end{aligned} \quad (5)$$

A derivation of $J(\mathbf{K})$ as per (5) is reported in the Appendix.

Remark 2: Note that $J(\mathbf{K})$ is a multivariate polynomial in the entries of \mathbf{K} . This is because one can verify

$$(I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1} = \sum_{i=0}^N (\mathbf{C} \mathbf{P}_{12} \mathbf{K})^i, \quad (6)$$

due to the fact that each $pN \times pN$ block on the diagonal of $\mathbf{C} \mathbf{P}_{12} \mathbf{K}$ is the zero matrix by construction, and hence $(\mathbf{C} \mathbf{P}_{12} \mathbf{K})^i = 0_{pN \times pN}$ for every $i \geq N + 1$.

To summarize, in this paper we are interested in solving the following optimization problem:

$$\begin{aligned} \text{Problem } \mathcal{P}_K & \\ \min_{\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}} & J(\mathbf{K}). \end{aligned}$$

which might be non-convex due to J being non-convex in \mathbf{K} in general.

C. Disturbance-feedback strategies

Motivated by e.g. [18], [19], we introduce the concept of equivalent disturbance-feedback strategies.

Lemma 1: A disturbance-feedback control input of the form $\tilde{\mathbf{u}} = \mathbf{Q} \mathbf{C} \mathbf{P}_{11} \mathbf{w} + \mathbf{Q} \mathbf{v}$ yields the same state, output and input trajectories as an output-feedback control input of the form $\mathbf{u} = \mathbf{K} \mathbf{y}$ if and only if

$$\mathbf{K} = h(\mathbf{Q}, \mathbf{C} \mathbf{P}_{12}),$$

where $h : \mathbb{R}^{mN \times pN} \rightarrow \mathbb{R}^{mN \times pN}$ is the bijection defined as

$$\begin{aligned} h(\mathbf{Q}, \mathbf{C} \mathbf{P}_{12}) &= (I + \mathbf{Q} \mathbf{C} \mathbf{P}_{12})^{-1} \mathbf{Q}, \\ h^{-1}(\mathbf{K}, \mathbf{C} \mathbf{P}_{12}) &= \mathbf{K} (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1}. \end{aligned}$$

Proof: For $\mathbf{u} = \mathbf{K} \mathbf{y}$, from (2) we derive the closed-loop equations:

$$\begin{aligned} \mathbf{x} &= (I - \mathbf{P}_{12} \mathbf{K} \mathbf{C})^{-1} (\mathbf{P}_{11} \mathbf{w} + \mathbf{P}_{12} \mathbf{K} \mathbf{v}), \\ \mathbf{y} &= \mathbf{C} (I - \mathbf{P}_{12} \mathbf{K} \mathbf{C})^{-1} \mathbf{P}_{11} \mathbf{w} + (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1} \mathbf{v}, \\ \mathbf{u} &= \mathbf{K} \mathbf{C} (I - \mathbf{P}_{12} \mathbf{K} \mathbf{C})^{-1} \mathbf{P}_{11} \mathbf{w} + \mathbf{K} (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})^{-1} \mathbf{v}. \end{aligned} \quad (7)$$

By defining $\mathbf{Q} = h^{-1}(\mathbf{K}, \mathbf{C} \mathbf{P}_{12}) = \mathbf{K} (I - \mathbf{C} \mathbf{P}_{12} \mathbf{K})$, the above equations can be rewritten as

$$\begin{aligned} \mathbf{x} &= (I + \mathbf{P}_{12} \mathbf{Q} \mathbf{C}) \mathbf{P}_{11} \mathbf{w} + \mathbf{P}_{12} \mathbf{Q} \mathbf{v}, \\ \mathbf{y} &= \mathbf{C} (I + \mathbf{P}_{12} \mathbf{Q}) \mathbf{P}_{11} \mathbf{w} + (I + \mathbf{C} \mathbf{P}_{12} \mathbf{Q}) \mathbf{v}, \\ \mathbf{u} &= \mathbf{Q} \mathbf{C} \mathbf{P}_{11} \mathbf{w} + \mathbf{Q} \mathbf{v}. \end{aligned}$$

Hence, the disturbance-feedback control input $\mathbf{u} = \mathbf{Q} \mathbf{C} \mathbf{P}_{11} \mathbf{w} + \mathbf{Q} \mathbf{v}$ achieves the same trajectories as $\mathbf{u} = \mathbf{K} \mathbf{y}$. \blacksquare

Now, define the function

$$\begin{aligned} \tilde{J}(\mathbf{Q}) &= \left\| \mathbf{M}^{\frac{1}{2}} (I + \mathbf{P}_{12} \mathbf{Q} \mathbf{C}) \mathbf{P}_{11} \Sigma_w^{\frac{1}{2}} \right\|_F^2 + \\ &+ \left\| \mathbf{M}^{\frac{1}{2}} \mathbf{P}_{12} \mathbf{Q} \Sigma_v^{\frac{1}{2}} \right\|_F^2 + \left\| \mathbf{R}^{\frac{1}{2}} \mathbf{Q} \mathbf{C} \mathbf{P}_{11} \Sigma_w^{\frac{1}{2}} \right\|_F^2 + \\ &+ \left\| \mathbf{R}^{\frac{1}{2}} \mathbf{Q} \Sigma_v^{\frac{1}{2}} \right\|_F^2 + \left\| \mathbf{R}^{\frac{1}{2}} \mathbf{Q} \mathbf{C} \mathbf{P}_{11} \mu_w \right\|_2^2 + \\ &+ \left\| \mathbf{M}^{\frac{1}{2}} (I + \mathbf{P}_{12} \mathbf{Q} \mathbf{C}) \mathbf{P}_{11} \mu_w \right\|_2^2. \end{aligned} \quad (8)$$

By directly comparing the expressions for $\tilde{J}(\mathbf{Q})$ and $J(\mathbf{K})$, we obtain that:

$$\begin{aligned} \tilde{J}(h^{-1}(\mathbf{K}, \mathbf{C} \mathbf{P}_{12})) &= J(\mathbf{K}), \\ \tilde{J}(\mathbf{Q}) &= J(h(\mathbf{Q}, \mathbf{C} \mathbf{P}_{12})). \end{aligned}$$

We introduce the following Lemmas about $J(\mathbf{K})$ and $\tilde{J}(\mathbf{Q})$ which will be useful for proving our main results in Section 3 and Section IV. The proofs are reported in the Appendix.

Lemma 2: The function $\tilde{J}(\mathbf{Q})$ is strictly convex and quadratic in \mathbf{Q} .

Lemma 3: Let $\mathbf{K}_0 \in \mathbb{R}^{mN \times pN}$ and define the sublevel set at $J(\mathbf{K}_0)$ as

$$\mathcal{L} := \{\mathbf{K} \mid J(\mathbf{K}) \leq J(\mathbf{K}_0)\}.$$

The sublevel set \mathcal{L} is bounded for any \mathbf{K}_0 .

D. Quadratic invariance

Since $\tilde{J}(\cdot)$ is convex in the disturbance-feedback strategies, it can be exploited for convex computation of sparse controllers. In particular, if and only if a property denoted as Quadratic Invariance (QI) holds, it is possible to compute a solution of \mathcal{P}_K by solving a corresponding convex program in the disturbance-feedback parameter \mathbf{Q} . We recall the notion of QI and the corresponding result from [4]:

Definition 1: A subspace $\mathcal{C} \subset \mathbb{R}^{mN \times pN}$ is QI with respect to \mathbf{CP}_{12} if and only if

$$\mathbf{KCP}_{12}\mathbf{K} \in \mathcal{C}, \quad \forall \mathbf{K} \in \mathcal{C},$$

and it is strongly QI with respect to \mathbf{CP}_{12} if and only if

$$\mathbf{K}_1\mathbf{CP}_{12}\mathbf{K}_2 \in \mathcal{C}, \quad \forall \mathbf{K}_1, \mathbf{K}_2 \in \mathcal{C}.$$

Notice that a subspace is QI if it is strongly QI, but not vice-versa. Note that a sparsity subspace $\text{Sparse}(\mathbf{S})$ is QI if and only if it is strongly QI [4]. The QI property has been shown to be necessary and sufficient for convex design of globally optimal distributed controllers [4], [5]. To review this result, notice that problem \mathcal{P}_K is equivalent to

$$\min_{\mathbf{Q} \in h^{-1}(\text{Sparse}(\mathbf{S}) \cap \mathcal{K}, \mathbf{CP}_{12})} \tilde{J}(\mathbf{Q}), \quad (9)$$

where $h^{-1}(\text{Sparse}(\mathbf{S}) \cap \mathcal{K}, \mathbf{CP}_{12}) = -h(\text{Sparse}(\mathbf{S}) \cap \mathcal{K}, \mathbf{CP}_{12})$ by definition of the bijection h . The main QI result is that problem (9), equivalent to \mathcal{P}_K , is convex if and only QI holds.

Theorem 1 (QI, [4], [5]): The set $h(\text{Sparse}(\mathbf{S}) \cap \mathcal{K}, \mathbf{CP}_{12})$ is convex if and only if $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ is QI with respect to \mathbf{CP}_{12} , and if and only if it is equal to $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$.

It follows from Theorem 1 that problem \mathcal{P}_K is equivalent to a convex program, and in particular to

$$\min_{\mathbf{Q} \in \text{Sparse}(\mathbf{S})} \tilde{J}(\mathbf{Q}), \quad (10)$$

if and only if QI holds.

The rest of the paper is dedicated to developing a first-order gradient-descent method to solve \mathcal{P}_K to global optimality directly in the \mathbf{K} domain.

III. FIRST-ORDER METHOD FOR GLOBALLY OPTIMAL SPARSE CONTROLLERS GIVEN QI

Given a sparsity constraint $\mathbf{K} \in \text{Sparse}(\mathbf{S})$, we characterize the set of stationary points for problem \mathcal{P}_K as follows:

Proposition 1: Consider problem \mathcal{P}_K with $\mathbf{K} \in \text{Sparse}(\mathbf{S})$. A controller $\bar{\mathbf{K}} \in \text{Sparse}(\mathbf{S})$ is a stationary point for \mathcal{P}_K if and only if

$$\nabla_{\mathbf{K}} J(\bar{\mathbf{K}}) \in \text{Sparse}(\mathbf{S}^c), \quad (11)$$

where \mathbf{S}^c is the binary matrix that has a 0 wherever \mathbf{S} has a 1, and a 1 wherever \mathbf{S} has a 0. When (11) holds, we say that $\bar{\mathbf{K}}$ is a *structured stationary point*.

Proof: Let $\mathbf{k} = \text{vec}(\mathbf{K})$ for every \mathbf{K} and f be the function such that $f(\mathbf{k}) = J(\mathbf{K})$ for every \mathbf{K} . By definition $\bar{\mathbf{K}} \in \text{Sparse}(\mathbf{S})$ is a stationary point for \mathcal{P}_K if and only if

$$\nabla f(\bar{\mathbf{k}})^\top (\mathbf{k} - \bar{\mathbf{k}}) \geq 0, \quad \forall \mathbf{k} \in \text{Sparse}(\text{vec}(\mathbf{S})). \quad (12)$$

Now suppose that (11) does not hold or equivalently there exists i such that $\text{vec}(\mathbf{S})_i = 1$ and $\nabla f(\bar{\mathbf{k}})_i = \alpha \neq 0$. Take $\mathbf{k} = \bar{\mathbf{k}} + \alpha e_i$, where e_i is the i -th element of the standard orthonormal basis of \mathbb{R}^{mpN^2} . Since $\mathbf{k} \in \text{Sparse}(\text{vec}(\mathbf{S}))$ and

$$\nabla f(\bar{\mathbf{k}})^\top (\mathbf{k} - \bar{\mathbf{k}}) = \nabla f(\bar{\mathbf{k}})^\top (-\alpha e_i) = -\alpha^2 < 0,$$

we conclude that (12) implies (11). Vice-versa, if (11) holds, then $\nabla f(\bar{\mathbf{k}}) \in \text{Sparse}(\text{vec}(\mathbf{S}^c))$ and

$$\nabla f(\bar{\mathbf{k}})^\top (\mathbf{k} - \bar{\mathbf{k}}) = 0, \quad \forall \mathbf{k} \in \text{Sparse}(\text{vec}(\mathbf{S})).$$

Hence, (11) implies (12). \blacksquare

In general, a structured stationary point as in (11) can be a local/global minimum/maximum or a saddle point for problem \mathcal{P}_K . In the next lemma, we show that the set of stationary points for \mathcal{P}_K corresponds to that of stationary points for problem (10) when strong QI holds. The proof is based on the idea of [18, Lemma 1] and is reported in the Appendix.

Lemma 4: Suppose that \mathcal{C} is strongly QI with respect to \mathbf{CP}_{21} , and let $\bar{\mathbf{K}} \in \mathcal{C}$. Also define $\bar{\mathbf{Q}} = h^{-1}(\bar{\mathbf{K}}, \mathbf{CP}_{12})$. We have that

$$\nabla_{\bar{\mathbf{Q}}} \tilde{J}(\bar{\mathbf{Q}}) \in \mathcal{C}^\perp \iff \nabla_{\bar{\mathbf{K}}} J(\bar{\mathbf{K}}) \in \mathcal{C}^\perp.$$

A. Global optimality of gradient descent

By exploiting Lemma 4, our first result establishes that, if $\text{Sparse}(\mathbf{S})$ is QI with respect to \mathbf{CP}_{12} , any structured stationary point of \mathcal{P}_K is a global optimum.

Theorem 2: Suppose that $\text{Sparse}(\mathbf{S})$ is QI with respect to \mathbf{CP}_{12} and let $\mathbf{K}^* \in \text{Sparse}(\mathbf{S})$ be a structured stationary point of $J(\mathbf{K})$. Then,

$$\mathbf{K}^* \in \arg \min_{\mathbf{K} \in \text{Sparse}(\mathbf{S})} J(\mathbf{K}).$$

Proof: By Theorem 1, \mathcal{P}_K is equivalent to (10). Since problem (10) is convex, every $\mathbf{Q}^* \in \text{Sparse}(\mathbf{S})$ such that $\nabla \tilde{J}(\mathbf{Q}^*) \in \text{Sparse}(\mathbf{S}^c)$ (that is, \mathbf{Q}^* is a structured stationary point) is a global optimum and thus achieves the optimal cost J^* . Let $\mathbf{K}^* = h(\mathbf{Q}^*, \mathbf{CP}_{12})$. Now remember that $\text{Sparse}(\mathbf{S})$ is QI if and only if it is strongly QI [4]. By Lemma 4 we have

$$\nabla J(\mathbf{K}^*) \in \text{Sparse}(\mathbf{S}^c),$$

and hence \mathbf{K}^* is a structured stationary point for $J(\mathbf{K})$. Since $\tilde{J}(\mathbf{Q}) = J(h(\mathbf{Q}, \mathbf{CP}_{12}))$ for every \mathbf{Q} by definition, we have that $J(\mathbf{K}^*) = \tilde{J}(\mathbf{Q}^*) = J^*$ and thus

$$\mathbf{K}^* \in \arg \min_{\mathbf{K} \in \text{Sparse}(\mathbf{S})} J(\mathbf{K}).$$

By Lemma 4, there can be no other structured stationary point $\bar{\mathbf{K}} \in \text{Sparse}(\mathbf{S}^c)$ such that $J(\bar{\mathbf{K}}) = \bar{J} > J^*$; otherwise, $\bar{\mathbf{Q}} = h^{-1}(\bar{\mathbf{K}}, \mathbf{CP}_{12})$ would also be a structured stationary point for problem (10) with cost $\bar{J} > J^*$, which is a contradiction because problem (10) is convex. \blacksquare

Remark 3: We remark that Theorem 2 trivially generalizes to any constraint set $\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ where \mathcal{K} is strongly QI. Indeed, the key Lemma 4 holds for any strongly QI subspace. In Theorem 2, we decided to make the most common case of sparsity constraints explicit in the interest of clarity. Instead, if \mathcal{K} was QI, but not strongly QI, Lemma 4 would not hold because the matrix $\tilde{\mathbf{K}}$ as per (25) would not necessarily lie in \mathcal{K} . Future work will investigate whether Theorem 2 can be proven for QI subspaces that are not strongly QI.

Theorem 2 clarifies a fundamental insight: if we can find any structured stationary point of the generally non-convex

function $J(\mathbf{K})$, and QI holds, this is enough to guarantee that we have solved problem \mathcal{P}_K . Strong of this observation, we develop a gradient-descent method that solves \mathcal{P}_K to global optimality for QI sparsity constraints.

Theorem 3: Suppose $\text{Sparse}(\mathbf{S})$ is QI with respect to CP_{12} . Let $\mathbf{K}_0 \in \text{Sparse}(\mathbf{S})$ be an initial output-feedback control policy, and consider the iteration

$$\mathbf{K}_{t+1} = \mathbf{K}_t - \eta_t \nabla J(\mathbf{K}_t) \odot \mathbf{S}, \quad (13)$$

Then, $\mathbf{K}_t \in \text{Sparse}(\mathbf{S})$ for every t and there exists a choice for η_t at every t such that

$$\lim_{t \rightarrow \infty} J(\mathbf{K}_t) = J^*,$$

where J^* is the optimal value of problem \mathcal{P}_K .

The proof of Theorem 3 makes use of Lemma 3 and the following three Lemmas adapted from [20]. The adapted proof of Lemma 5 is reported in the Appendix and the proofs of Lemma 6 and Lemma 7 can be found in [20, Lemma 3.1] and [21, Proposition 5.7] respectively.

Lemma 5: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be bounded below in \mathbb{R}^n and consider the iteration

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \quad (14)$$

where η_t satisfies the Wolfe conditions:

$$f(x_t - \eta_t \nabla f(x_t)) \leq f(x_t) - c_1 \eta_t \|\nabla f(x_t)\|_2^2, \quad (15)$$

$$-\nabla f(x_t - \eta_t \nabla f(x_t))^\top \nabla f(x_t) \geq -c_2 \|\nabla f(x_t)\|_2^2, \quad (16)$$

for some $0 < c_1 < c_2 < 1$.

Let f be continuously differentiable in an open set \mathcal{U} containing the level set $\mathcal{L} = \{x : f(x) \leq f(x_0)\}$, where x_0 is the starting point of the iteration (14). Assume that the gradient ∇f is Lipschitz continuous on \mathcal{U} , that is, there exists a constant $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathcal{U}.$$

Then,

$$\lim_{t \rightarrow \infty} \nabla f(x_t) = 0_{n \times 1}.$$

Lemma 6: Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and bounded below. Then, for any $0 < c_1 < c_2 < 1$, there exist intervals of step lengths satisfying the Wolfe conditions (15)-(16).

Lemma 7: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable on an open convex set $\mathcal{U} \subseteq \mathbb{R}^n$, and suppose that the Hessian $\nabla^2 f$ is bounded on \mathcal{U} . Then ∇f is Lipschitz continuous on \mathcal{U} .

We are now ready to prove Theorem 3.

Proof: [Theorem 3] We denote $\mathbf{k} \in \mathbb{R}^{|\mathbf{S}|}$ to be the vector that contains only those entries of $\mathbf{K} \in \text{Sparse}(\mathbf{S})$ that correspond to non-zero entries of $\text{vec}(\mathbf{S})$. Let $f : \mathbb{R}^{|\mathbf{S}|} \rightarrow \mathbb{R}$ be the function such that $f(\mathbf{k}) = J(\mathbf{K})$ for every $\mathbf{K} \in \text{Sparse}(\mathbf{S})$. Then, $\nabla_{\mathbf{k}} f(\mathbf{k}) \in \mathbb{R}^{|\mathbf{S}|}$ contains only those entries of $\nabla J(\mathbf{K})$ that correspond to non-zero entries of \mathbf{S} . Hence, the iteration (13) is equivalent to

$$\mathbf{k}_{t+1} = \mathbf{k}_t - \eta_t \nabla f(\mathbf{k}_t), \quad (17)$$

which is a standard gradient-descent iteration. By definition \mathbf{k}_t corresponds to a $\mathbf{K}_t \in \text{Sparse}(\mathbf{S})$, and hence the iterates of (13) remain in $\text{Sparse}(\mathbf{S})$ if $\mathbf{K}_0 \in \text{Sparse}(\mathbf{S})$.

At every iteration t , let us choose η_t satisfying (15)-(16). Notice that, according to Lemma 6, a choice for η_t exists for every t because f is continuously differentiable and bounded below by 0. Let \mathbf{k}_0 be the initial value for the decision variables. We know by Lemma 3 that the sublevel set

$$\mathcal{L} = \{\mathbf{k} \mid f(\mathbf{k}) \leq f(\mathbf{k}_0)\},$$

is bounded. Consider an open, convex and bounded set \mathcal{U} that contains \mathcal{L} (for instance, an open ball that contains \mathcal{L}). Since $f(\mathbf{k})$ is a multivariate polynomial, every entry of its Hessian matrix $\nabla^2 f(\mathbf{k})$ is also a multivariate polynomial. Every multivariate polynomial is bounded on a bounded set, and hence $\nabla^2 f(\mathbf{k})$ is bounded on \mathcal{U} . By Lemma 7, we deduce that there exists $L > 0$ such that

$$\|\nabla f(\mathbf{k}_1) - \nabla f(\mathbf{k}_2)\| \leq L \|\mathbf{k}_1 - \mathbf{k}_2\|, \quad \forall \mathbf{k}_1, \mathbf{k}_2 \in \mathcal{U}.$$

By Lemma 5, we conclude that

$$\lim_{t \rightarrow \infty} \nabla f(\mathbf{k}_t) = 0_{n \times 1}. \quad (18)$$

Let $\mathbf{k}^* = \lim_{t \rightarrow \infty} \mathbf{k}_t$ and $\mathbf{K}^* \in \text{Sparse}(\mathbf{S})$ the corresponding output-feedback controller. Since $\nabla f(\mathbf{k}^*) = 0_{|\mathbf{S}| \times 1}$, by definition $\nabla J(\mathbf{K}^*) \odot \mathbf{S} = 0_{mN \times n(N+1)}$ or equivalently $\nabla J(\mathbf{K}^*) \in \text{Sparse}(\mathbf{S}^c)$, that is, \mathbf{K}^* is a structured stationary point. We conclude by Theorem 2 that \mathbf{K}^* is a globally optimal solution for \mathcal{P}_K because $\text{Sparse}(\mathbf{S})$ is QI with respect to CP_{12} . ■

A choice for the stepsize η_t for every t satisfying the Wolfe conditions (15)-(16) can be found by using, for instance, the following bisection Algorithm at each time t .

Algorithm 1 Bisection for the Wolfe conditions (15)-(16)

- 1: Let $0 < c_1 < c_2 < 1$, $a = 0$, $b = 2$, $continue = 1$,
 $\mathbf{p} = -\nabla f(\mathbf{k}_t)$
 - 2: **while** $f(\mathbf{k}_t + b\mathbf{p}) \leq f(\mathbf{k}_t) - c_1 b \|\mathbf{p}\|_2^2$ **do**
 - 3: $b = 2b$
 - 4: **end while**
 - 5: **while** $continue = 1$ **do**
 - 6: $\eta_t = \frac{1}{2}(a + b)$
 - 7: **if** $f(\mathbf{k}_t + \eta_t \mathbf{p}) > f(\mathbf{k}_t) - c_1 \eta_t \|\mathbf{p}\|_2^2$ **then**
 - 8: $b = \eta_t$
 - 9: **else if** $\nabla f(\mathbf{k}_t + \eta_t \mathbf{p})^\top \mathbf{p} < -c_2 \|\mathbf{p}\|_2^2$ **then**
 - 10: $a = \eta_t$
 - 11: **else**
 - 12: $continue = 0$
 - 13: **end if**
 - 14: **end while** **return** η_t
-

We refer the reader to [21, Proposition 5.5] for a proof that Algorithm 1 returns a stepsize satisfying (15)-(16) in a finite number of iterations. We conclude this Section by providing a numerical example.

B. Numerical example

Motivated by the example system of [4], we consider (1) and the cost function (5) with

$$A_t = \begin{bmatrix} 1.6 & 0 & 0 & 0 & 0 \\ 0.5 & 1.6 & 0 & 0 & 0 \\ 2.5 & 2.5 & -1.4 & 0 & 0 \\ -2 & 1 & -2 & 0.1 & 0 \\ 0 & 2 & 0 & -0.5 & 1.1 \end{bmatrix}, \quad B_t = I, \quad C_t = I,$$

$$M_t = I, \quad R_t = I, \quad \Sigma_w = I, \quad \Sigma_v = I, \quad \forall t = 0, \dots, 5,$$

and $\mu_0 = [1 \quad -1 \quad 2 \quad -3 \quad 3]^\top$, where we set a horizon of $N = 5$. Our goal is to compute a controller \mathbf{K} with a given sparsity that minimizes the cost (5). Specifically, we aim at solving \mathcal{P}_K with $\mathcal{K} = \mathbb{R}^{mN \times pN}$ and

$$\mathbf{S} = T \otimes S,$$

where $T_{i,j} = 1$ if $j \leq i$ and $T_{i,j} = 0$ otherwise, and

$$S = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

It follows that the total number of scalar variables is $|\mathbf{S}| = \frac{|\mathbf{S}|N(N+1)}{2} = 30$. It is easy to verify that $\text{Sparse}(\mathbf{S})$ is QI with respect to \mathbf{CP}_{12} , for example by using the binary test [4, Theorem 26]. By direct computation of the Hessian through the Symbolic Math Toolbox Ver. 7.1 available in MATLAB [22] we verify that $\nabla^2 J(0_{|\mathbf{S}| \times 1}) \neq 0$, where $\mathbf{k} \in \mathbb{R}^{|\mathbf{S}|}$ contains only the entries of $\mathbf{K} \in \text{Sparse}(\mathbf{S})$ such that $\text{vec}(\mathbf{S})_{i,j} = 1$. It follows that $J(\mathbf{K})$ is not convex in \mathbf{K} . Despite this non-convexity, we know by Theorem 3 that the gradient descent iteration (13) will converge to a global optimum of \mathcal{P}_K for $t \rightarrow \infty$ thanks to the QI property.

1) *Numerical results:* The gradient-descent iteration (13) was implemented in MATLAB with the stepsize being chosen according to Algorithm 1. The iteration (13) was initialized from a variety of randomly selected initial distributed controllers. Specifically, for each entry (i, j) such that $\mathbf{S}_{i,j} = 1$ we selected the entry $(\mathbf{K}_0)_{i,j}$ uniformly at random in the interval $[-10, 10]$, and set $(\mathbf{K}_0)_{i,j} = 0$ otherwise. In all instances, we converged to an optimal cost of 767.5627 in $\approx 500 - 600$ iterations, with a run time of approximately 2 seconds. The stopping criterion was selected as $\max |\nabla J(\mathbf{k}_t)| < 0.005$. To validate the global optimality result, we solved the corresponding convex program (10) in \mathbf{Q} with MOSEK [23], called through MATLAB via YALMIP [24], and obtained a minimum cost of 767.5627.

At this point, it is natural to ask a follow-up question: is the QI/strong QI property *necessary* to guarantee convergence of gradient-descent to a globally optimal distributed controller? In the following section, we provide a negative answer.

IV. UNIQUE STATIONARITY: GLOBAL OPTIMALITY BEYOND QI

It was shown in [5] that the QI condition is not only sufficient, but also necessary for convexity of the set $h(\text{Sparse}(\mathbf{S}), \mathbf{CP}_{12})$. This implies that convex programming can only be used to systematically solve \mathcal{P}_K in the disturbance-feedback parameter \mathbf{Q} if QI holds. In this paper, we have shown that descending the gradient directly in the \mathbf{K} domain also guarantees global convergence for (strongly) QI problems. One may then be tempted to guess that, similarly to [5], QI/strong QI are necessary conditions to certify global convergence of first-order methods.

In this Section, we show otherwise. We define a novel class of problems, which we denote as *Uniquely Stationary* problems (US). Every US problem can be solved to global optimality via gradient-descent starting from any initial controller in (13). Our main result is that the class of US problems is strictly larger than that of strongly QI problems, and it is not included in the class of QI problems. It follows that first-order minimization offers certificates of global optimality strictly beyond those offered by convex programming for distributed control.

A. Unique stationarity generalizes QI

We define the concept of unique stationarity of problem \mathcal{P}_K as follows.

Definition 2 (Unique Stationarity (US)): Consider problem \mathcal{P}_K subject to a general subspace constraint $\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$, where \mathbf{S} is a binary matrix and $\mathcal{K} \subseteq \mathbb{R}^{mN \times pN}$ is a subspace. We say that \mathcal{P}_K is *Uniquely Stationary* (US) if and only if for every $\bar{\mathbf{K}}$:

$$\Pi_{\mathcal{K}}(\nabla J(\bar{\mathbf{K}}) \odot \mathbf{S}) = 0 \implies \bar{\mathbf{K}} \in \arg \min_{\text{Sparse}(\mathbf{S}) \cap \mathcal{K}} J(\mathbf{K}), \quad (19)$$

where $\Pi_{\mathcal{K}}(\cdot)$ is the projection operator on the subspace \mathcal{K} .

We have the following result about US problems.

Proposition 2: Suppose that \mathcal{P}_K is US. Let $\mathbf{K}_0 \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ and consider the iteration

$$\mathbf{K}_{t+1} = \mathbf{K}_t - \eta_t \Pi_{\mathcal{K}}(\nabla J(\mathbf{K}_t) \odot \mathbf{S}). \quad (20)$$

Then, $\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ for every t and there exists a choice for η_t at every t such that

$$\lim_{t \rightarrow \infty} J(\mathbf{K}_t) = J^*,$$

where J^* is the optimal value of problem \mathcal{P}_K .

Proof: First, notice that (20) is equivalent to

$$\mathbf{K}_{t+1} = \Pi_{\text{Sparse}(\mathbf{S}) \cap \mathcal{K}}(\mathbf{K}_t - \eta_t \nabla J(\mathbf{K}_t)).$$

It follows that (20) is a standard projected gradient-descent iteration. Let $\{\mathbf{n}_i\}_{i=1}^r$ be an orthonormal basis of $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ and let $f : \mathbb{R}^{mpN^2} \rightarrow \mathbb{R}$ be the function such that $f(\text{vec}(\mathbf{K})) = J(\mathbf{K})$ for every \mathbf{K} . Also define $\bar{f} : \mathbb{R}^r \rightarrow \mathbb{R}$ as

$$\bar{f}(\boldsymbol{\alpha}) = f\left(\sum_{i=1}^r \alpha_i \mathbf{n}_i\right), \quad (21)$$

where $\alpha = [\alpha_1, \dots, \alpha_r]^\top$ and $r \leq mpN^2$ is the dimension of $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$. By (21), the constrained optimization problem of minimizing $J(\mathbf{K})$ over $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ is equivalent to the unconstrained optimization problem of minimizing $\bar{f}(\alpha)$. It follows that (20) is equivalent to the iteration

$$\alpha_{t+1} = \alpha_t - \eta_t \nabla \bar{f}(\alpha_t).$$

We have shown in Theorem 3 that the iteration above converges to $\lim_{t \rightarrow \infty} \nabla \bar{f}(\alpha_t) = 0$ by choosing η_t according to (15)-(16), for instance by using Algorithm 1. Since \mathcal{P}_K is US, every stationary point is optimal and thus achieves the cost J^* . ■

We deduce that every US problem can be solved to global optimality with a projected gradient-descent algorithm. Our main result is as follows.

Theorem 4: The class of US problems is strictly larger than the class of strongly QI problems, and it is not included in the class of QI problems.

The proof of Theorem 4 is based on constructing an example of a US problem which is neither strongly QI or QI, for which the property (19) can be certified a priori in a tractable way.

Example: Consider the system (1) and the cost function (5) with

$$A_t = \begin{bmatrix} 1 & 2 \\ -1 & -3 \end{bmatrix}, \quad B_t = I, \quad C_t = I, \quad \Sigma_0 = I, \\ M_t = I, \quad R_t = I, \quad \Sigma_t^w = 0, \quad \Sigma_v = 0, \quad \forall t = 0, 1, 2,$$

and $\mu_0 = 0$, where we set a horizon of $N = 2$. The output-feedback controller \mathbf{K} is subject to being in the form

$$\mathbf{K} = I_N \otimes K,$$

for some $K \in \mathbb{R}^{2 \times 2}$. In other words, we consider the static-controller case where $u_t = Ky_t$, as was also considered in the works [15], [17] for the infinite-horizon setup. We remark that in the finite-horizon framework considered here, it is not necessary that $(A+BK)$ is Hurwitz, since the finite-horizon cost $J(\mathbf{K})$ is finite for every \mathbf{K} . Additionally, we require that K is decentralized, or equivalently $K \in \text{Sparse}(I_2)$. The resulting subspace constraint for problem \mathcal{P}_K is expressed as

$$\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K} = \{\mathbf{K} = I_N \otimes \text{diag}(x, y), \quad x, y \in \mathbb{R}\}.$$

By computing $\mathbf{KCP}_{12}\mathbf{K}$ for a generic $\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ it is easy to verify that $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ is neither strongly QI or QI with respect to \mathbf{CP}_{12} . Hence, a convex program equivalent to \mathcal{P}_K in the \mathbf{Q} domain does not exist by Theorem 1. Nonetheless, we have the following result:

Proposition 3: The problem \mathcal{P}_K considered in this example is US. Specifically, \mathcal{P}_K is convex in \mathbf{K} .

Proof: For any $\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$, the cost function (5) is equivalent to

$$J(\mathbf{K}) = J(x, y) = 2x^4 + 6x^3 + 14x^2 + 10xy - 22x + 2y^4 - 18y^3 + 67y^2 - 110y + 87.$$

The expression above was obtained by using the Symbolic Math Toolbox in MATLAB [22]. The Hessian of $J(x, y)$ is

$$H(x, y) = \begin{bmatrix} 24x^2 + 36x + 28 & 10 \\ 10 & 24y^2 - 108y + 134 \end{bmatrix} \\ = \begin{bmatrix} 24(x + \frac{3}{4})^2 + \frac{29}{2} & 10 \\ 10 & 24(y - \frac{9}{4})^2 + \frac{25}{2} \end{bmatrix}.$$

Clearly, $H(x, y) \succ 0$ for all $x, y \in \mathbb{R}$, and hence $J(\mathbf{K})$ is convex over $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$. It follows that \mathcal{P}_K is convex and hence US, despite not being QI. We remark that, for this example, checking whether \mathcal{P}_K is US is done in polynomial time. Indeed, a sufficient condition for US in this case is that the diagonal entries of the Hessian of $J(\mathbf{K})$ are sum-of-square (SOS) polynomials over $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$. Deciding whether a polynomial is a SOS is tractable [25]. ■

We are now ready to prove Theorem 4.

Proof: First, we have proven in Theorem 1 that if \mathcal{P}_K is strongly QI, then every stationary point is a global optimum. See also Remark 3. Hence, (19) holds and every strongly QI problem is US. Second, by Proposition 3 there exists an instance of a US problem \mathcal{P}_K which is neither strongly QI or QI. This proves that the class of US problems is both strictly larger than strong QI problems and not included in QI problems. ■

Theorem 4 shows that the notion of US genuinely extends QI. This is possible because QI is only necessary for global optimality certificates of convex programming in the \mathbf{Q} domain. However, we have shown that QI might miss instances of \mathcal{P}_K which may even be convex in the original \mathbf{K} coordinates.

Remark 4: The reader might be familiar with the recent results of [6], [7], where it was shown that for a special instance of \mathcal{P}_K one can construct a convex restriction in the \mathbf{Q} domain which contains a global optimum despite a lack of QI. One may argue that such an example shows that [6], [7] go beyond the QI notion. However, as was also clarified in [6], the techniques of [6], [7] do not offer a global optimality certificate beyond QI. This is because in [6], [7] global optimality is certified only after computing the optimal solution analytically, which may defy the purpose of solving an optimization problem numerically in the first place.

Instead, in this paper, we have provided an example of one non-QI problem for which a global optimality certificate is obtained without knowing its solution analytically. Additionally, similarly to the binary test of QI for sparsity constraints [4, Theorem 26], the US property was proven in a tractable way.

B. Tests for unique stationarity

The reader might have noticed that the US property, while having a theoretical interest, is not useful in practice in the form (19). This is because, in general, one can only prove (19) by knowing the set of global optima a priori. For this reason, it is necessary to identify sufficient conditions for US that are not based on directly checking that every stationary point is a global optimum. While noting that more general tests should be envisioned in future research, we provide

our initial results. A first test for US follows naturally as a corollary of Theorem 2:

Corollary 1 (Theorem 2): Suppose that $\mathbf{K}_1 \mathbf{C} \mathbf{P}_{12} \mathbf{K}_2 \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ for every $\mathbf{K}_1, \mathbf{K}_2 \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$. Then, \mathcal{P}_K is US.

Proof: The required property is strong QI. If \mathbf{K} is a stationary point of \mathcal{P}_K then by definition

$$\Pi_{\text{Sparse}(\mathbf{S}) \cap \mathcal{K}}(\nabla J(\mathbf{K})) = \Pi_{\mathcal{K}}(\nabla J(\mathbf{K}) \odot \mathbf{S}) = 0.$$

By Theorem 1 and Remark 3, it follows that the above implies \mathbf{K} is a global optimum. Hence, US as per (19) holds. ■

When $\mathcal{K} = \mathbb{R}^{mN \times pN}$, it was shown [4, Theorem 26] that QI of the sparsity subspace $\text{Sparse}(\mathbf{S})$ can be tested in polynomial time by checking

$$(\mathbf{S} \text{ Struct}(\mathbf{C} \mathbf{P}_{12}) \mathbf{S})_{i,j} \neq 0 \implies \mathbf{S}_{i,j} = 1, \quad \forall i, j.$$

A second sufficient test for US beyond QI is to check whether \mathcal{P}_K is convex, that is to verify that $J(\mathbf{K})$ is convex on $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$.

Proposition 4: Let $\mathbf{k} = \text{vec}(\mathbf{K})$, $\{\mathbf{n}_i\}_{i=1}^r$ be an orthonormal basis of $\text{Sparse}(\mathbf{S}) \cap \mathcal{K}$, f be the function such that $f(\mathbf{k}) = J(\mathbf{K})$ for every \mathbf{K} , and define $\bar{f} : \mathbb{R}^r \rightarrow \mathbb{R}$ as per (21). Define $H = \nabla^2 \bar{f}(\alpha)$ and let \bar{H}_g be the submatrix of \bar{H} obtained by removing its last $r - g$ rows and columns. Then, if the determinant of H_g is a positive multivariate polynomial for every $g = 1, \dots, r$, the problem \mathcal{P}_K is convex and thus US.

Proof: Notice that for every $\mathbf{K} \in \text{Sparse}(\mathbf{S}) \cap \mathcal{K}$ there exists $\alpha \in \mathbb{R}^r$ such that $\mathbf{k} = \sum_{i=1}^r \alpha_i \mathbf{n}_i$ and vice-versa. Hence, \mathcal{P}_K is equivalent to the unconstrained problem $\min_{\alpha} \bar{f}(\alpha)$. The function \bar{f} is convex if and only if its Hessian H is positive definite for every α , or equivalently the polynomial $\det(H_g)$ is globally positive for every g . The statement follows. ■

Deciding positivity of multivariate polynomials is NP-hard in general, but it can be performed in finite time [26]. When the determinants of H_g as per the proposition above are all SOS, as in our Example above, then the US property can be decided in polynomial time [25].

V. CONCLUSIONS

We have studied convergence to a global optimum of gradient descent first-order methods for the distributed LQG problem in finite-horizon. When the strong QI property holds, a steepest descent algorithm is guaranteed to converge to a global optimum. Moreover, we have characterized the class of US problems, which is strictly larger than strong QI problems and not included in QI problems, for which projected gradient descent converges to a global optimum. Our results show that first-order methods in the \mathbf{K} domain are superior to convex programming in the \mathbf{Q} domain in terms of generality of the corresponding global optimality certificates. Additionally, first-order methods can be used to learn globally optimal distributed controllers when the system and the cost function are unknown, as was recently

shown in [12], [14] for the non-distributed case. A follow-up work will discuss this application of our methods.

This work opens up the possibility of finding novel classes of distributed control problems, for which an a priori and possibly tractable test of the US property is available, beyond QI and beyond testing convexity of \mathcal{P}_K in the \mathbf{K} domain. For instance, one can study conditions such that $\bar{f}(\alpha)$ is *gradient dominated* [27], a condition more general than convexity which ensures convergence of gradient descent and thus implies US. It is also important to extend this work to the infinite-horizon and continuous-time cases.

ACKNOWLEDGEMENTS

We thank Tyler Summers and Ilnura Usmanova for useful discussions.

APPENDIX

A. Derivation of the cost function $J(\mathbf{K})$

Note that the cost (4) is equivalent to

$$J(\mathbf{K}) = \mathbb{E}_{\mathbf{w}, \mathbf{v}} (\mathbf{x}^T \mathbf{M} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u}). \quad (22)$$

Now consider the control input $\mathbf{u} = \mathbf{K} \mathbf{y}$. The closed-loop state, output and input trajectories are given in (7), where \mathbf{x} and \mathbf{u} are expressed as a function of \mathbf{w} and \mathbf{v} . Substitute (7) into (22). By using the fact that for any matrix \mathbf{X} we have

$$\begin{aligned} \mathbb{E}_{\mathbf{w}} (\mathbf{w}^T \mathbf{X} \mathbf{w}) &= \text{Trace}(\mathbf{X} \Sigma_w) + \boldsymbol{\mu}_w^T \mathbf{X} \boldsymbol{\mu}_w, \\ \mathbb{E}_{\mathbf{v}} (\mathbf{v}^T \mathbf{X} \mathbf{v}) &= \text{Trace}(\mathbf{X} \Sigma_v), \\ \mathbb{E}_{\mathbf{w}, \mathbf{v}} (\mathbf{w}^T \mathbf{X} \mathbf{v}) &= 0, \end{aligned}$$

and remembering that $\|\mathbf{X}\|_F^2 = \text{Trace}(\mathbf{X}^T \mathbf{X})$ we obtain the expression (5).

B. Proof of Lemma 2

By using several relationships to compute derivatives with respect to matrices from [28] and the fact that $\text{vec}(AXB) = (B^T \otimes A) \text{vec}(x)$, where \otimes is the Kronecker product, we obtain that

$$\begin{aligned} &\nabla \left(\text{vec} \left(\nabla \tilde{J}(\mathbf{Q}) \right) \right) \\ &= 2 \left[\underbrace{(\Sigma_v + \mathbf{C} \mathbf{P}_{11} \Sigma_w \mathbf{P}_{11}^T \mathbf{C}^T)}_{\succ 0} \otimes \underbrace{(\mathbf{R} + \mathbf{P}_{12}^T \mathbf{M} \mathbf{P}_{12})}_{\succ 0} \right. \\ &\quad \left. + \underbrace{\mathbf{C} \mathbf{P}_{11} \boldsymbol{\mu}_w \boldsymbol{\mu}_w^T \mathbf{P}_{11}^T \mathbf{C}^T}_{\succeq 0} \otimes \underbrace{(\mathbf{R} + \mathbf{P}_{12}^T \mathbf{P}_{12})}_{\succ 0} \right] \succ 0. \end{aligned}$$

because $\mathbf{R}, \Sigma_v \succ 0$ and $\mathbf{M} \succeq 0$ by hypothesis. The positive definite matrix above is the Hessian of $\tilde{J}(\mathbf{Q})$ when the entries of \mathbf{Q} are stacked into a single column. It follows that $\tilde{J}(\mathbf{Q})$ is a quadratic form that is strictly convex everywhere.

C. Proof of Lemma 3

Since $\tilde{J}(\mathbf{Q})$ is strictly convex everywhere by Lemma 2, its sublevel set

$$\tilde{\mathcal{L}} := \{\mathbf{Q} \mid \tilde{J}(\mathbf{Q}) \leq J(\mathbf{K}_0)\},$$

is bounded for any \mathbf{K}_0 [29, Ch. 9.1.2]. Since $\tilde{J}(\mathbf{Q}) = J(h(\mathbf{Q}, \mathbf{CP}_{12}))$ for every \mathbf{Q} we have

$$\mathcal{L} = h(\tilde{\mathcal{L}}, \mathbf{CP}_{12}).$$

Now notice that, similar to (6)

$$h(\mathbf{Q}, \mathbf{CP}_{12}) = (I + \mathbf{QCP}_{12})^{-1}\mathbf{Q} = \sum_{i=0}^N (-1)^i (\mathbf{QCP}_{12})^i \mathbf{Q}.$$

The expression above clarifies that if all the entries of \mathbf{Q} are finite, so are all the entries of $h(\mathbf{Q}, \mathbf{CP}_{12})$. In particular, we conclude that \mathcal{L} is bounded if and only if $\tilde{\mathcal{L}}$ is bounded. Since $\tilde{\mathcal{L}}$ is bounded for any \mathbf{K}_0 , the result follows.

D. Proof of Lemma 4

\Leftarrow) In the interest of readability, in this proof we omit the second argument of the function $h(\cdot, \cdot)$, which is assumed to always be fixed to \mathbf{CP}_{12} . Assume that $\nabla_{\mathbf{K}} J(\bar{\mathbf{K}}) \in \mathcal{C}^\perp$, but $\nabla_{\mathbf{Q}} \tilde{J}(\bar{\mathbf{Q}}) \notin \mathcal{C}^\perp$. Then, there exists $\tilde{\mathbf{Q}} \in \mathcal{C}$ with $\tilde{\mathbf{Q}} \neq 0$ such that:

$$\lim_{\epsilon \rightarrow 0} \frac{\tilde{J}(\bar{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}}) - \tilde{J}(\bar{\mathbf{Q}})}{\epsilon} = k \neq 0,$$

that is, there exists a direction $\tilde{\mathbf{Q}}$ in \mathcal{C} along which the gradient is not null at $\bar{\mathbf{Q}}$. Equivalently, since $h(\cdot)$ is invertible,

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \frac{\tilde{J}(h^{-1}(h(\bar{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}}))) - \tilde{J}(\bar{\mathbf{Q}})}{\epsilon} &= \\ = \lim_{\epsilon \rightarrow 0} \frac{J(h(\bar{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}})) - J(\bar{\mathbf{K}})}{\epsilon} &= k \neq 0 \end{aligned} \quad (23)$$

Now, observe

$$\begin{aligned} h(\bar{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}}) &= (I + (\bar{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}})\mathbf{CP}_{12})^{-1}(\bar{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}}) \quad (24) \\ &= (I + \tilde{\mathbf{Q}}\mathbf{CP}_{12} + \epsilon \tilde{\mathbf{Q}}\mathbf{CP}_{12})^{-1}(\tilde{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}}) \\ &= \left[(I + \tilde{\mathbf{Q}}\mathbf{CP}_{12})^{-1} + \epsilon (I + \tilde{\mathbf{Q}}\mathbf{CP}_{12})^{-1} \tilde{\mathbf{Q}}\mathbf{CP}_{12} \right. \\ &\quad \left. \cdot (I + \tilde{\mathbf{Q}}\mathbf{CP}_{12})^{-1} + \mathcal{O}(\epsilon^2) \right] (\tilde{\mathbf{Q}} + \epsilon \tilde{\mathbf{Q}}) = \bar{\mathbf{K}} + \epsilon \tilde{\mathbf{K}} + \mathcal{O}(\epsilon^2), \end{aligned}$$

where

$$\tilde{\mathbf{K}} = (I + \bar{\mathbf{Q}}\mathbf{CP}_{12})^{-1}(\tilde{\mathbf{Q}} + \tilde{\mathbf{Q}}\mathbf{CP}_{12}(I + \bar{\mathbf{Q}}\mathbf{CP}_{12})^{-1}\tilde{\mathbf{Q}}). \quad (25)$$

Since $\bar{\mathbf{Q}}\mathbf{CP}_{12}$ is strictly lower-triangular by construction, it is easy to verify that

$$(I + \bar{\mathbf{Q}}\mathbf{CP}_{12})^{-1} = \sum_{i=0}^N (-1)^i (\bar{\mathbf{Q}}\mathbf{CP}_{12})^i,$$

and by repeated application of the strong QI property we deduce that $\tilde{\mathbf{K}} \in \mathcal{C}$. By substituting the above derivations into (23), we obtain

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \frac{J(\bar{\mathbf{K}} + \epsilon \tilde{\mathbf{K}} + \mathcal{O}(\epsilon^2)) - J(\bar{\mathbf{K}})}{\epsilon} &= \\ = \lim_{\epsilon \rightarrow 0} \frac{J(\bar{\mathbf{K}} + \epsilon \tilde{\mathbf{K}}) - J(\bar{\mathbf{K}})}{\epsilon} &= k \neq 0. \end{aligned}$$

Since $\tilde{\mathbf{K}} \in \mathcal{C}$, this contradicts $\nabla_{\mathbf{K}} J(\bar{\mathbf{K}}) \in \mathcal{C}^\perp$. \Rightarrow) can be proven analogously.

E. Proof of Lemma 5

From (16) and (14) we have

$$\nabla f(x_{t+1})^\top \nabla f(x_t) \leq c_2 \|\nabla f(x_t)\|_2^2,$$

while the Lipschitz condition implies

$$\nabla f(x_{t+1})^\top \nabla f(x_t) \geq (1 - \eta_t L) \|\nabla f(x_t)\|_2^2,$$

and hence

$$\eta_t \geq \frac{1 - c_2}{L}.$$

By substituting the above inequality into (15) we obtain

$$f(x_{t+1}) \leq f(x_t) - c_1 \frac{1 - c_2}{L} \|\nabla f(x_t)\|_2^2.$$

By summing the above expression over all indices $j = 1, \dots, t$ we obtain

$$f(x_{t+1}) \leq f(x_0) - c_1 \frac{1 - c_2}{L} \sum_{j=0}^t \|\nabla f(x_j)\|_2^2. \quad (26)$$

Since f is bounded below, we have that $f(x_0) - f(x_{t+1})$ is bounded above by some positive constant for all t . Hence, we conclude from (26) that

$$\sum_{t=0}^{\infty} \|\nabla f(x_t)\|_2^2 < \infty,$$

which implies

$$\lim_{t \rightarrow \infty} \nabla f(x_t) = 0_{n \times 1}.$$

REFERENCES

- [1] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, no. 1, pp. 131–147, 1968.
- [2] C. H. Papadimitriou and J. Tsitsiklis, "Intractable problems in control theory," *SIAM jour. on contr. and opt.*, vol. 24, no. 4, pp. 639–654, 1986.
- [3] D. Youla, H. Jabr, and J. Bongiorno, "Modern Wiener-Hopf design of optimal controllers—Part II: The multivariable case," *IEEE Transactions on Automatic Control*, vol. 21, no. 3, pp. 319–338, 1976.
- [4] M. Rotkowitz and S. Lall, "A characterization of convex problems in decentralized control," *IEEE Transactions on Automatic Control*, vol. 51, no. 2, pp. 274–286, 2006.
- [5] L. Lessard and S. Lall, "Quadratic invariance is necessary and sufficient for convexity," in *American Control Conference (ACC), 2011*. IEEE, 2011, pp. 5360–5362.
- [6] L. Frieri, Y. Zheng, A. Papachristodoulou, and M. Kamgarpour, "Sparsity invariance for convex design of distributed controllers," *arXiv preprint arXiv:1906.06777*, 2019.
- [7] Y.-S. Wang, N. Matni, and J. C. Doyle, "A system level approach to controller synthesis," *IEEE Transactions on Automatic Control*, 2019.
- [8] G. Fazelnia, R. Madani, A. Kalbat, and J. Lavaei, "Convex relaxation for optimal distributed control problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 206–221, 2017.
- [9] Y. Wang, J. A. Lopez, and M. Sznajder, "Convex optimization approaches to information structured decentralized control," *IEEE Transactions on Automatic Control*, 2018.
- [10] K. Dvijotham, E. Todorov, and M. Fazel, "Convex structured controller design in finite horizon," *IEEE Transactions on Control of Network Systems*, vol. 2, no. 1, pp. 1–10, 2015.
- [11] F. Lin, M. Fardad, and M. R. Jovanovic, "Augmented lagrangian approach to design of structured optimal state feedback gains," *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2923–2929, 2011.
- [12] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," *arXiv preprint arXiv:1801.05039*, 2018.

- [13] B. Gravell, Y. Guo, and T. Summers, "Sparse optimal control of networks with multiplicative noise via policy gradient," *arXiv preprint arXiv:1905.13547*, 2019.
- [14] B. Gravell, P. Mohajerin Esfahani, and T. Summers, "Global convergence of policy gradient methods for the linear quadratic regulator," *arXiv preprint arXiv:1905.13547*, 2019.
- [15] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.
- [16] K. Mårtensson and A. Rantzer, "Gradient methods for iterative distributed control synthesis," in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009, pp. 549–554.
- [17] H. Feng and J. Lavaei, "On the exponential number of connected components for the feasible set of optimal decentralized control problems," in *Proceedings of the 2019 American Control Conference*, 2019, p. 8.
- [18] M. Colombino, R. S. Smith, and T. H. Summers, "Mutually quadratically invariant information structures in two-team stochastic dynamic games," *IEEE Transactions on Automatic Control*, vol. 63, no. 7, pp. 2256–2263, 2017.
- [19] L. Frieri and M. Kamgarpour, "The value of communication in synthesizing controllers given an information structure," *arXiv preprint arXiv:1711.05324*, 2017.
- [20] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.
- [21] F. J. Aragón, M. A. Goberna, M. A. López, and M. M. Rodríguez, *Nonlinear optimization*. Springer, 2019.
- [22] *MATLAB R2016b and Symbolic Math Toolbox Version 7.1*. Natick, Massachusetts: The MathWorks Inc., 2016.
- [23] MOSEK Aps, "The MOSEK optimization toolbox for MATLAB manual. Version 8.1." 2017.
- [24] J. Löfberg, "YALMIP : A Toolbox for Modeling and Optimization in MATLAB," in *In Proc. of the CACSD Conf.*, Taipei, Taiwan, 2004.
- [25] P. A. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, California Institute of Technology, 2000.
- [26] E. Becker, V. Powers, and T. Wormann, "Deciding positivity of real polynomials," *Contemporary Mathematics*, vol. 253, pp. 19–24, 2000.
- [27] B. T. Polyak, "Gradient methods for the minimisation of functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 864–878, 1963.
- [28] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.
- [29] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.