Regularization for Design

Nikolai Matni and Venkat Chandrasekaran

***Abstract*— An algorithmic bridge is starting to be established between sparse reconstruction theory and distributed control theory. For example, ℓ_1-regularization has been suggested as an appropriate means for co-designing sparse feedback gains and consensus topologies subject to performance bounds. In recent work, we showed that ideas from atomic norm minimization could be used to simultaneously co-design a distributed optimal controller and the communication delay structure on which it is to be implemented. While promising and successful, these results lack the same theoretical support that their sparse reconstruction counterparts enjoy – as things stand, these methods are at best viewed as principled heuristics. In this paper, we describe theoretical connections between sparse reconstruction and systems design by developing approximation bounds for control co-design problems via convex optimization.**

I. INTRODUCTION

As we move into the era of large scale systems, such as the new smart grid, clean slate internet and automated highway systems, the design of high performing controllers will become more and more challenging. For example, distributed controllers are characterized by the structure imposed on them by information sharing constraints – this structure can take the form of sparsity, in which each controller has access to only a subset of local measurements, or of delay, in which each controller eventually has access to all measurements, but with a delay dictated by a communication network.

In general, such structured optimal control problems can be very difficult (c.f. [1] for a canonical example) – however, with the identification of Quadratic Invariance (QI) [2] as an appropriate condition for the convexification of such problems, much progress has been made. We refer the reader to [3], and the references therein, for a survey of recent results in the area. Furthermore, even when the underlying problem is not QI, progress has been made in designing structured linear feedback controllers; representative examples include sparsity inducing control [4], [5], convex relaxations of rank constrained problems [6], [7], the minimization of convex surrogates to traditional performance metrics [8], spatial truncation [9], positive systems [10], [11], and localized distributed control [12], [13], [14]. . For the purposes of this paper, the salient feature of all of these results is that imposing and/or inducing structural constraints on the design variable, i.e. the controller, can be done in a *convex* manner.

What is often taken for granted in these results (exceptions include [4], [5], [9]) is the existence of information sharing constraints to begin with. Indeed, in many systems engineering settings, there is the possibility of designing how information is shared between controllers. For example, in the delay constrained setting, if the speed of communication between controllers is twice as fast as dynamics propagate through the physical plant, then this suggests the possibility of adding shortcuts between second neighbors. Designing such information constraints, and searching for an "optimal" topology, that is to say one that optimally trades off between communication structure complexity and closed loop performance, is in its full generality most likely an intractable combinatorial problem.

An approach which has seen much success in similar problems in other fields has been to employ convex relaxations in order to approximately recover such solutions. The idea of using convex relaxations in optimization problems, in particular in the setting of attempting to recover structured solutions, has a rich and fruitful history in the machine learning and statistics communities. In particular, it is often known *a priori* that the solution to an optimization problem should be structurally "simple."

It has been shown that this simple structure can often be approximately, and sometimes exactly, recovered by minimizing an appropriately chosen convex penalty function. Well known examples include the ℓ_1-norm to induce sparse solutions, and the nuclear norm to induce low-rank solutions (e.g. [15], [16], [17]). In [18], this notion of "simplicity" was formalized and generalized in terms of *atomic norms*. We refer the interested reader to [19] for a review of the current state of the art, both theoretically and computationally, of using regularizers to induce structure in a statistical setting.

These developments have not gone unnoticed within the control community, and indeed the natural connection between such sparse model selection ideas and system identification has been made. The use of atomic norms and regularizers to identify systems of low Hankel order have been particularly fruitful, c.f. [20], [21], [22], and similar ideas have been suggested for linear regression based methods [23]. However, one can argue that these system identification problems are essentially model selection problems, but with a specific type of structure, namely that of a linear time invariant (LTI) system's impulse response.

Representatives of the use of such ideas in the control literature for the *design* of information sharing constraints have only recently begun to emerge. These include the use

N. Matni is with the Department of Control and Dynamical Systems, California Institute of Technology, Pasadena, CA. nmatni@caltech.edu.

V. Chandrasekaran is with the Department of Computational and Mathematical Sciences, California Institute of Technology, Pasadena, CA.. venkatc@caltech.edu.

This research was in part supported by NSF Career award CCF-1350590, and by the NSF, AFOSR, ARPA-E, and the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

of ℓ_1-regularization to design sparse $\mathcal{H}_2$ optimal feedback gains [5], sparse treatment therapies [24], sparse consensus [25], [26] and synchronization [27] topologies , and the use of a particular type of atomic norm to design communication delay constraints that are well suited to $\mathcal{H}_2$ distributed optimal control [28], [29]. One can make the argument that these results form an algorithmic link between the results in sparse reconstruction theory and co-design for optimal control. The computational methods have been ported, with great success, but these methods do not yet enjoy the same theoretical support that their sparse reconstruction counterparts do.

The goal of this paper is to show that there is indeed a way to make natural theoretical links between these two areas. We will argue that *identifying* appropriately structured estimators in a machine learning, system identification or statistical context is in fact conceptually and mathematically parallel to *designing* appropriately structured controllers in an optimal control setting. By appropriately re-interpreting the parameters and measures of success of these estimation problems to be compatible with a design problem, we will be able to leverage the results from the machine learning community that give these methods their legitimacy. We will focus in particular on the so-called group norm (and its variant that allows overlap [30]), and repurpose existing results from the sparse reconstruction literature [19] towards our goals. In doing so, we will be able to provide sufficient conditions under which the co-design of a controller and its structure succeeds.

This paper is organized as follows: in Section II, we give a brief overview of atomic norms, and give some examples of how they can be used for design in control, focussing in particular on the group norm. In Section III, we pause to make an explicit theoretical link between identification and design, and then with this link in mind, present a stylized example, what we term "actuator regularization," to be used to illustrate the ideas that we will be presenting. Continuing our focus on the group norm, in Section V, we formally introduce group norm regularized optimizations, and provide sufficient conditions for their success. We then return to our stylized example in Section VI, and show via a numerical study the usefulness of these techniques from a design perspective. Finally, we end with conclusions and directions for future work in Section VII. All proofs can be found in the [31].

II. Atomic Norms in Control

A. Atomic norms and structured solutions

As mentioned in the introduction, it is often known *a priori* that the solution to an optimization problem should be "simple," and that this simple structure can be promoted through the use of an appropriate *convex* penalty. This notion of solutions with simple structure, in the context of linear inverse problems, has been formalized and generalized in terms of *atomic norms* [18].

In particular, if it is known that the true solution X_* to a set of linear equations $y = AX + \nu$, for some bounded noise term $||\nu|| \leq \delta$ (we use $\|\cdot\|$ to denote the Euclidean norm), should consist of a linear combination of a small number of "atoms", then it is shown that one should seek the solution that minimizes an appropriately defined atomic norm, subject to consistency constraints. Specifically, if one assumes that

$$X_* = \sum_{i=1}^{r} c_i a_i, \; a_i \in \mathcal{A}, \; c_i \geq 0$$

for $\mathcal{A}$ a set of appropriately scaled and centered atoms, and r a small number relative to the ambient dimension, then solving

$$\text{minimize}_X \|X\|_{\mathcal{A}} \quad \text{s.t. } \|y - AX\|^2 \leq \delta^2 \tag{1}$$

with the atomic norm $\|\cdot\|_{\mathcal{A}}$ given by the gauge function

$$\begin{aligned} ||X||_{\mathcal{A}} : &= \inf\{t \geq 0 \mid X \in t\text{conv}(\mathcal{A})\} \\ &= \inf\{\textstyle\sum_{a \in \mathcal{A}} |c_a| \mid X = \sum_{a \in \mathcal{A}} c_a a\} \end{aligned} \tag{2}$$

results in solutions that both satisfy the consistency constraint $\|y - AX\|^2 \leq \delta^2$, and are sparse at the atomic level (i.e. are a linear combination of a small number of elements $a \in \mathcal{A}$).

The geometric justification behind the success of these methods is that the unit-ball of an atomic norm is appropriately "pointy" in high dimensions, and thus solutions are likely to be at singularities (i.e. edges or corners) of the norm-ball, inducing the desired simple structure.

B. Applications to distributed control

As distributed control problems are naturally characterized by the structural constraints imposed on the controller, this is a natural area to which regularization for co-design ideas can be applied.

1) Static controllers: Arguably the simplest controller/constraint co-design problem is one in which the controller to be designed is a static feedback gain, and there are no additional design constraints on the communication structure beyond the desire for sparsity. Then the natural atomic norm to use as a regularizer is the ℓ_1-norm – indeed this is the approach that has been taken in [5]. Note that similar ideas could be applied to the optimization problems formulated for positive systems in [10] and [11].

Suppose now that we ask more from our design – not only do we wish for a sparse feedback gain K, but we also require that its sparsity pattern correspond to that of the adjacency matrix of a strongly connected graph. It has been shown that indeed the ℓ_1/ℓ_2-norm with overlap, or group norm with overlap, promotes such a sparsity structure [30]. Thus, through simple modification of the atomic norm used in the regularizer, we can achieve more or less structure in our controller, and additionally, we can fine tune this structure to our design needs.

2) Dynamic controllers: More care must be taken when designing regularizers for dynamic controllers, as now the structure induced must be consistent both spatially and temporally. As a representative example, we consider the task of inducing structure that is consistent with how information propagates across a communication graph – indeed it was shown in [28], [29] that such a structure can be induced via a carefully chosen atomic norm, allowing for the co-design of delay patterns that are well suited to $\mathcal{H}_2$ distributed optimal control subject to strongly connected communication graphs [32].

C. Other applications

This idea is also applicable to co-design problems outside the scope of distributed control. For example, a sparse communication graph for consensus can be designed using ℓ_1-regularization [25], [26]. In Section IV, we introduce a novel control co-design task that falls outside of the realm of any examples discussed – that of designing actuation and sensing schemes (to be made precise later). As it is a conceptually simple problem to formulate, and does not require any of the technical machinery of distributed control (such as QI), we will use it as a stylized example throughout the remainder of the paper. However, we first pause to make an explicit, and we argue natural, theoretical link between sparse reconstruction and controller co-design.

III. Design vs. Identification

Let y, $\epsilon \in \mathcal{Y}$, $x \in \mathcal{X}$, with $\mathcal{X}$, $\mathcal{Y}$ Hilbert spaces, and let $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator from $\mathcal{X}$ to $\mathcal{Y}$. We consider a generic convex optimization problem

$$\min_{x \in \mathcal{X}} C\left(y - \mathcal{L}(x)\right) + \lambda \|x\|_{\mathcal{A}} \tag{3}$$

where $\|\cdot\|_{\mathcal{A}}$ is an appropriately chosen atomic norm, C is a convex cost function, and λ is a regularization parameter. We assume that there is an underlying desired solution x^*, related to y via

$$y = \mathcal{L}(x^*) + \epsilon. \tag{4}$$

In the following, we discuss how this optimization problem can be interpreted from both a sparse reconstruction/statistical perspective, and a controller and structure co-design perspective.

A. A statistical interpretation

In a statistical setting, C is often taken to be a risk or loss function, y and $\mathcal{L}$ are observations and measurements, respectively, ϵ corresponds to measurement/estimation noise/error, and x is an estimator to be designed that explains the observed data y given the measurements $\mathcal{L}$. Typically, the regularizer $\|\cdot\|_{\mathcal{A}}$ is chosen to leverage *a priori* knowledge that the estimator x should be sparse at the atomic level.

The objective of this minimization is to use the observation and measurement data $(y, \mathcal{L})$ and the prior knowledge on the structure of x to obtain an estimator that not only explains the current data properly, but that also *predicts* future observations given future measurements. In trying to achieve this task, the often competing metrics of bias and variance can be traded off against each other, and it is not necessarily a requirement to recover the "true" atomic support of x, so long as the designed estimator can be interpreted, and is sufficiently predictive.

B. A control theoretic interpretation

In a control co-design setting, C is taken to be a performance metric (such as the $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norms), y corresponds to the open loop system, x the controller to be designed, and $\mathcal{L}$ the mapping that takes the controller to its effect on the closed loop system. In this case, $\epsilon = y - \mathcal{L}(x)$ corresponds to the closed loop system, and $C(y - \mathcal{L}(x)) = C(\epsilon)$ the closed loop performance of the system. In particular note that if the state space parameters of the system are known, then y and $\mathcal{L}$ can be computed. Typically the regularizer $\|\cdot\|_{\mathcal{A}}$ is chosen in accordance to *design constraints and objectives*.

The objective of this optimization is then to find a controller that minimizes the closed loop norm of the system and that has desirable structure. In particular, there is no notion of prediction, bias or variance in this setting, and we argue that recovering the atomic support (i.e. the information sharing constraint) of the controller x is of paramount importance, as this informs the structural design of the controller.

C. A natural synthesis

It is worth focusing on the "ϵ" term in both of the above settings – in particular, we see a mathematical equivalence between measurement noise in the statistical setting, and the closed loop response of the system on the control theoretic side. Intuitively, this suggests that controllers achieving smaller closed loop norms (i.e. those whose identification is subject to less "measurement noise") are easier to recover than those achieving poorer closed loop performance. In what follows, we will show that this intuition indeed holds true.

Further, although not the subject of this paper, the compatibility of these two problems within the same framework suggests that there may be a natural synthesis of the fields of statistics and control. Indeed if the analysis used on both types of problems can be appropriately combined, this would yield a unified and principled framework for analyzing joint system identification, control, and adaptation schemes.

IV. A Stylized Example

With the observations and questions of the previous section in mind, we now introduce our stylized "actuator regularization" problem.

A. $\mathcal{H}_2$ Preliminaries and Notation

In the following, if $\mathcal{M}$ is a subspace of an inner product space, we denote the orthogonal projection onto $\mathcal{M}$ by $\mathbb{P}_{\mathcal{M}}$.

Let $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ be the unit disc of complex numbers, and let $\bar{\mathbb{D}}$ be its closure. A function $G : (\mathbb{C} \bigcup \{\infty\}) \backslash \bar{\mathbb{D}} \to \mathbb{C}^{p \times q}$ is in $\mathcal{H}_2$ if it can be expanded as

$$G(z) = \sum_{i=0}^{\infty} \frac{1}{z^i} G_i$$

where $G_i \in \mathbb{C}^{p \times q}$ and $\sum_{i=0}^{\infty} \mathrm{Tr}(G_i G_i^*) < \infty$. Define the conjugate of G by

$$G(z)^{\sim} = \sum_{i=0}^{\infty} z^i G_i^*$$

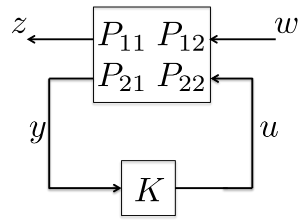


Fig. 1: The generalized plant model of Section IV

$\mathcal{H}_2$ is a Hilbert space with inner product given by

$$\begin{aligned} < G, H > &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Tr}(G(e^{j\theta})H(e^{j\theta})^{\sim})d\theta \\ &= \sum_{i=0}^{\infty} \mathrm{Tr}(G_i H_i^*), \end{aligned}$$

where the last equality follows from Parseval's identity.

B. $\mathcal{H}_2$ *optimal control*

Let P be a stable discrete-time plant given by

$$P = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right] = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \tag{5}$$

with inputs of dimension p_1, p_2 and outputs of dimension q_1, q_2. The control input $u \in \mathbb{R}^{p_2}$, disturbance $w \in \mathbb{R}^{p_1}$, measured output $y \in \mathbb{R}^{q_2}$ and controlled output $z \in \mathbb{R}^{q_1}$ are related via the diagram in Figure 1, where each P_{ij} is given by

$$P_{ij} = C_i(zI - A)^{-1}B_j + D_{ij}. \tag{6}$$

In this example we our restrict attention to stable plants, and look to design a controller K so as to minimize the closed loop $\mathcal{H}_2$ norm of the system:

$$\begin{aligned} \underset{K}{\text{minimize}} \quad & ||P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}||_{\mathcal{H}_2} \\ & \text{s.t.}\, K(I - P_{22}K)^{-1} \in \mathcal{H}_2 \end{aligned} \tag{7}$$

where the constraint is sufficient (under mild technical assumptions) to ensure internal stability of the system.

As is standard, we now pass to the Youla parameterization to rewrite (7) as

$$\begin{aligned} \underset{Q}{\text{minimize}} \quad & ||P_{11} + P_{12}QP_{21}||_{\mathcal{H}_2}. \\ & \text{s.t.}\, Q \in \mathcal{H}_2 \end{aligned} \tag{8}$$

C. *Actuator Regularization*

Suppose for simplicity that B_2 is a diagonal matrix with non-zero elements along its diagonal – in this case, by constraining a row of K, or equivalently a row of Q, to be zero, we are effectively removing the corresponding actuator from the system's design. Note that under dual assumptions, constraining a column would correspond to removing a sensor from the system's design.

Remark 1: This problem could also be interpreted as an actuator/sensor placement problem for a distributed system under supervisory control.

Rather than constraining rows to be zero, we will look to "regularize actuators" through the use of the group-norm. In particular, let the atomic set be given by

$$\mathcal{A} \;:=\; \{Q^a \in \mathcal{H}_2 \,|\, Q^a_t = e_a(v^a_t)^\top, \forall t \geq 0, \|Q^a\|_{\mathcal{H}_2} = 1\}, \tag{9}$$

i.e. those stable transfer functions with a single non-zero row, normalized to have unit $\mathcal{H}_2$ norm.

For $Q \in \mathcal{H}_2^{p_2 \times q_2}$, this induces the atomic-norm

$$\|Q\|_{\mathcal{A}} = \sum_{a=1}^{q_2} \|(e^a)^\top Q\|_{\mathcal{H}_2}, \tag{10}$$

and thus induces sparsity at the actuator level. We can then write a regularized variant of (8) as

$$\begin{aligned} \underset{Q}{\text{minimize}} \quad & ||P_{11} + P_{12}QP_{21}||^2_{\mathcal{H}_2} + \lambda\|Q\|_{\mathcal{A}} \\ & \text{s.t.}\, Q \in \mathcal{H}_2 \end{aligned} \tag{11}$$

where $\lambda \geq 0$ is now a regularization parameter.

Remark 2: Through appropriate modifications, simultaneous actuator/sensor co-design could be done via an analogous group atomic norm built from the columns and rows – care needs to be taken in constructing such a regularizer however, due to the overlap between atoms. In the next section we show how this can be accommodated through the *group norm with overlap* [30].

Unfortunately, as stated, this problem is infinite dimensional, and there are no elegant ways of reducing this regularized problem to a finite dimensional one. Thus in order to perform the computation, we fix a time horizon T, and a controller order N, and approximate (11) with

$$\begin{aligned} \underset{Q}{\text{minimize}} \quad & \textstyle\sum_{t=0}^{T} ||y_t - G_t(Q)||_2^2 + \lambda\left(\sum_{a=1}^{q_2} \|(e^a)^\top Q\|_F\right) \\ \text{s.t.} \quad & Q = \begin{bmatrix} Q_0 & Q_1 & \dots & Q_N \end{bmatrix} \end{aligned} \tag{12}$$

with

$$G_t(Q) = \sum_{\substack{j,k,l \geq 0 \\ j+k+l=t}} H_j Q_k J_l \tag{13}$$

where $y_0 = 0$, $H_0 = D_{12}$, $J_0 = D_{21}$, and $y_t = C_1A^{t-1}B_1$, $H_t = C_1A^{t-1}B_2$, $J_t = C_2A^{t-1}B_1$ $\forall t \geq 1$.

We can write this in an even more concise form by defining the stacked vector $Y_T = (y_t)_{t=0}^T$ and the linear operator $\mathcal{L}_T(Q) = (G_t(Q))_{t=0}^T$, allowing us to reduce the problem to one of the same form as (3)

$$\begin{aligned} \underset{Q}{\text{minimize}} \quad & ||Y_T - \mathcal{L}_T(Q)||_F^2 + \lambda\left(\textstyle\sum_{a=1}^{q_2} \|(e^a)^\top Q\|_F\right) \\ \text{s.t.} \quad & Q = \begin{bmatrix} Q_0 & Q_1 & \dots & Q_N \end{bmatrix} \end{aligned} \tag{14}$$

Thus we see an explicit example of the control theoretic interpretation given to (3) in Section III – indeed Y_T corresponds to a stacked vector of the open loop response, $\mathcal{L}_T$ maps the effect of the controller Q to the closed loop impulse response, and letting $\epsilon(Q) := Y_T - \mathcal{L}_T(Q)$, we have that $\|\epsilon(Q)\|_F$ is the finite horizon $\mathcal{H}_2$ cost of the system using controller Q. The question now becomes, how can we measure the success of the optimization (14)?

As will be shown, the atomic norm considered in this section is a special case of the group norm. The next section defines different measures of approximation quality for a group norm regularized optimization, and provides sufficient conditions for the optimization to yield such an approximation.

V. The Group Norm

A. Preliminaries

Consider a set of groups $\mathcal{G}$, with $|\mathcal{G}| = G$, where each group $g \in \mathcal{G}$ can be identified with a subspace $S_g \subset R^{m\times p}$, with $\text{card} S_g = \dim S_g = |g|$. Define the group embedding operator $\mathcal{E}_g : \mathbb{R}^{|g|} \to S_g$ as the operator that appropriately embeds a vector $v_g \in \mathbb{R}^{|g|}$ such that $\mathcal{E}_g(v_g) \in S_g$. Its adjoint operator $\mathcal{E}_g^+$, as defined with respect to the trace inner-product on $\mathbb{R}^{m\times p}$, is then the operator that extracts the values of a vector $x \in \mathbb{R}^{m\times p}$ that lie in S_g and then appropriately maps them to $v_g \in \mathbb{R}^{|g|}$.

With these definitions in mind, let $v = (v_g)_{g\in\mathcal{G}} \in \mathbb{R}^q$, with $q = \sum_{g\in\mathcal{G}} |g|$, and define the group addition operator $\mathcal{A}_\mathcal{G} : \mathbb{R}^q \to \mathbb{R}^{m\times p}$ as

$$\mathcal{A}_\mathcal{G}(v) := \sum_{g\in\mathcal{G}} \mathcal{E}_g(v_g). \tag{15}$$

It is then immediate that its adjoint, once again with respect to the trace inner-product, is then given by

$$\mathcal{A}_\mathcal{G}^+(x) = \left(\mathcal{E}_g^+(x)\right)_{g\in\mathcal{G}} \tag{16}$$

Example 1: Let $m = p = 2$, $G = 2$, and suppose that $q = 4$ with

$$g_1 = \begin{bmatrix} * & 0 \\ * & 0 \end{bmatrix},\ g_2 = \begin{bmatrix} 0 & 0 \\ * & * \end{bmatrix} \tag{17}$$

where $*$ denote place holders for arbitrary real numbers.

Let $v_1 = [1, 2]^\top$ and $v_2 = [\pm 2, 3]^\top$. Then $v = [1, 2, \pm 2, 3]^\top$ and

$$\mathcal{A}_\mathcal{G}(v) = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \pm 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 \pm 2 & 3 \end{bmatrix}, \tag{18}$$

and

$$\mathcal{A}_\mathcal{G}^+\mathcal{A}_\mathcal{G}(v) = \begin{bmatrix} 1 & 2\pm 2 & 2\pm 2 & 3 \end{bmatrix}^\top. \tag{19}$$

This also illustrates the complications that can arise from allowing overlap between groups – in particular that there can be constructive $(2 + 2 = 4)$ or destructive $(2 - 2 = 0)$ interference between them.

We define the *group norm*, with respect to a set of groups $\mathcal{G}$, of a vector $v = (v_g)_{g\in\mathcal{G}} \in \mathbb{R}^q$ as

$$\|v\|_\mathcal{G} := \sum_{g\in\mathcal{G}} \|v_g\| \tag{20}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. The dual norm, which we denote by $\|\cdot\|_{\mathcal{G},\infty}$, is then given by

$$\|v\|_{\mathcal{G},\infty} := \sup_{g\in\mathcal{G}} \|v_g\| \tag{21}$$

B. Problem Formulation

We consider the task of approximating a vector v^* satisfying

$$y = \mathcal{L} \circ \mathcal{A}_\mathcal{G}(v^*) + \epsilon, \tag{22}$$

where we assume that $v^* = (v_g^*)_{g\in\mathcal{G}^*}$, $\mathcal{G}^* \subset \mathcal{G}$, is sparse at the group level, via the optimal solution $\hat{v}$ of the convex program

$$\text{minimize}_v \tfrac{1}{2}\|y - \mathcal{L} \circ \mathcal{A}_\mathcal{G}(v)\|_F^2 + \lambda\|v\|_\mathcal{G}. \tag{23}$$

Beyond the traditional measure of approximation error $\|\hat{v} - v^*\|_{\mathcal{G},\infty}$, which we will be able to control, we also consider three increasingly demanding measures of accuracy:

Definition 1: A vector $\hat{v}$ is *$\mathcal{A}_\mathcal{G}$-support accurate* with respect to v^* if $\text{supp}\,(\mathcal{A}_\mathcal{G}(\hat{v})) \subseteq \text{supp}\,(\mathcal{A}_\mathcal{G}(v^*))$.

Definition 2: A vector $\hat{v}$ is *$\mathcal{G}$-support accurate* with respect to v^* if $\text{gsupp}\,(\hat{v}) = \mathcal{G}^*$, where $\text{gsupp}\,(v) := \{g \in \mathcal{G} : \|v_g\| > 0\}$, that is to say the group support of v.

Additionally, in the following, we will overload notation and also use $\mathcal{G}^*$ to refer to the subspace spanned by all v such that $\text{gsupp}\,(v) \subseteq \mathcal{G}^*$, and similarly we will refer to its orthogonal complement $(\mathcal{G}^*)^\perp$ as the subspace spanned by all v such that $\text{gsupp}\,(v) \bigcap \mathcal{G}^* = \emptyset$. Finally,

Definition 3: A vector $\hat{v}$ is *$(\mathcal{G}, \delta)$-accurate* with respect to v^* if it is $\mathcal{G}$-support accurate, and for each $g \in \mathcal{G}^*$, it is true that

$$\left\langle \frac{\hat{v}_g}{\|\hat{v}_g\|}, \frac{v_g^*}{\|v_g^*\|} \right\rangle \geq \delta. \tag{24}$$

Notice that if $\hat{v}$ satisfies Definition 3, then it also satisfies Definition 2, but the converse is in general not true. Similarly, if $\hat{v}$ satisfies Definition 2, then it also satisfies Definition 1, but the converse is once gain generally not true.

Whereas Definitions 1 and 2 are straightforward to interpret, Definition 3 deserves some further discussion. Essentially, we ask that our approximation not only have the correct group support, but also that each active group vector $\hat{v}_g$ point approximately in the same direction as the corresponding true vector v_g^*. We argue that this is a more natural notion of success in our setting than statistical "sparsistency" (c.f. [33], [34]), which asks that correct support and sign be recovered in the approximation. In particular, if we interpret these vectors as control gains, then an approximation that is $(\mathcal{G}, \delta)$-accurate yields a controller that is both qualitatively and quantitatively similar in performance to the underlying optimal controller v^*.

C. $(\mathcal{G}, \delta)$-accuracy of the group norm without overlap

The ideas presented in this section, based on the notion of group incoherence, are not new and have been stated in various forms in the literature. (see the the recent review [19], and the references therein).

We analyze the performance of (23) under the assumption that there is no overlap between groups. In this case, $\mathcal{A}_\mathcal{G}$ is easily seen to be a bijection and a group-isometry (such that $\mathcal{A}_\mathcal{G}^+\mathcal{A}_\mathcal{G}$ is the identity operator), greatly simplifying the analysis – in fact, as we will discuss at the end of the section, no such non-asymptotic structural guarantees exist when there is overlap between groups.

Let $\mathcal{G}^*$ denote the set of non-zero groups in $v^* = \left(v_g^*\right)_{g\in\mathcal{G}}$, such that $v_g^* \neq 0$ if and only if $g \in \mathcal{G}^*$, and let

$$S^* = \oplus_{g\in\mathcal{G}^*} S_g,$$

where the S_g are the subspaces of $\mathbb{R}^{m\times p}$ that each $\mathcal{E}_g$ maps to. We can define S^* as a direct sum of these subspaces due to the assumption of *no overlap* between groups.

We assume that for all $\Delta \in \mathcal{G}^*$,

$$\min_{\|\Delta\|_{\mathcal{G},\infty}=1} \|\mathcal{P}_{\mathcal{G}^*}\mathcal{A}_{\mathcal{G}}^+\mathcal{L}_{|S^*}^+\mathcal{L}_{|S^*}\mathcal{A}_{\mathcal{G}}\mathcal{P}_{\mathcal{G}^*}\Delta\|_{\mathcal{G},\infty} \geq \alpha, \tag{25}$$

$$\max_{\|\Delta\|_{\mathcal{G},\infty}\leq 1} \|\mathcal{P}_{(\mathcal{G}^*)^\perp}\mathcal{A}_{\mathcal{G}}^+\mathcal{L}_{|(S^*)^\perp}^+\mathcal{L}_{|S^*}\mathcal{A}_{\mathcal{G}}\mathcal{P}_{\mathcal{G}^*}\Delta\|_{\mathcal{G},\infty} \leq \gamma, \tag{26}$$

and that there exists $\nu \in [0, 1)$ such that

$$\frac{\gamma}{\alpha} \leq \nu \tag{27}$$

Condition (25), often referred to as a self-incoherence condition, asks that $\mathcal{L}^+\mathcal{L}$ be bijective when restricted to the support of the active groups S^* – this condition limits the amount of *destructive* interference between active groups. Condition (26), often known as a mutual-incoherence condition, on the other hand, asks that $\mathcal{L}^+\mathcal{L}$ have small gain when viewed as a mapping from $S^* \to (S^*)^\perp$ – this condition limits the amount of *constructive* interference between active and inactive groups. Thus ν in (27) is a total measure of the incoherence, or interference, between groups.

We can now state our first consistency result:

Theorem 1: Let conditions (25), (26) and (27) hold, and suppose additionally that $\|\mathcal{A}_{\mathcal{G}}^+\mathcal{L}^+\epsilon\|_{\mathcal{G},\infty} \leq (\kappa-1)\lambda$ for some $1 \leq \kappa < 2/(\nu+1)$. Then

1) The solution $\hat{v}$ to optimization (23) is $\mathcal{A}_{\mathcal{G}}$-support accurate, and
2) $\|\hat{v} - v^*\|_{\mathcal{G},\infty} \leq \lambda\,(\kappa/\alpha)$.

Note that this theorem states that indeed $\mathcal{A}_{\mathcal{G}}$-support accuracy is possible for *any* sized ϵ by choosing λ large enough. As will be shown, however, this comes at the price of risking not identifying smaller sized groups, which in effect, can be "hidden" by the error. In a control context, this reads that if a particular controller v^* achieves a large closed loop norm ϵ, then optimization (23) will only recover the dominant components of the control action, as the remaining components may not be large enough to be detected. Alternatively, if $\mathcal{G}$-support and $(\mathcal{G},\delta)$-accuracy are desired, the underlying controller v^* must achieve a good closed loop performance ϵ with respect to the size of its control components – what this means will be formalized in the following corollaries.

Corollary 1: Suppose the conditions of Theorem 1 hold, and additionally that $\|v_g^*\| > \lambda\,(\kappa/\alpha)$ for all $g \in \mathcal{G}^*$. Then $\hat{v}$ is $\mathcal{G}$-support accurate. If in addition, each v_g^* has entry wise magnitude $|v_g^{*j}| > \lambda\,(\kappa/\alpha)$, then $\hat{v}$ is sparsistent as well.

In order to prove that $\hat{v}$ is $(\mathcal{G},\delta)$-accurate, we need to exploit the fact that the ℓ_2 unit ball is a manifold with curvature – doing so leads to the following lemma.

Lemma 1: Let $\Delta_g = v_g^* - \hat{v}_g$ be such that $\|\Delta_g\| \leq C\lambda$. Then, if $\|v_g^*\| \geq \frac{5C\lambda}{1-\delta^2}$, we have that $\left\langle v_g^*, \hat{v}_g\right\rangle \geq \delta\|\hat{v}_g\|\|v_g^*\|$.

Lemma 1 is easily interpreted geometrically. Suppose that we fix a perturbation size $\|\Delta_g\|$, but allow v_g^* to grow, and examine the angle between v_g^* and $\hat{v}_g := v_g^* + \Delta_g$. Then this angle will become negligible as $\|v_g^*\|$ becomes large. The following is then an immediate consequence of Theorem 1 and Lemma 1:

Corollary 2: Suppose that the conditions of Theorem 1 hold, and additionally that $\|v_g^*\| > \frac{5\kappa\lambda}{\alpha(1-\delta^2)}$ for all $g \in \mathcal{G}^*$. Then $\hat{v}$ is $(\mathcal{G},\delta)$-accurate.

As can be seen in both of the previous corollaries, the minimum sizes $\|v_g^*\|$ of the components of the controller that are guaranteed to be identified are lower bounded by λ, which in turn is lower bounded by a function of ϵ. As a consequence, it can be seen that *controllers leading to smaller closed-loop norms are easier to identify,* as the sufficient conditions of the corollaries are easier to satisfy.

D. Incoherence and gain estimates

We now specialize the discussion to when $y, \epsilon \in \mathbb{R}^m$, $x \in \mathbb{R}^q$ and $\mathcal{L}\circ\mathcal{A}_{\mathcal{G}}$ can be identified with a matrix $A \in \mathbb{R}^{m\times q}$, and provide estimates for α and γ (and hence ν).

In particular, we assume, permuting columns if necessary, that A admits a block-column partition

$$A = \begin{bmatrix} A_{\mathcal{G}^*} & A_{(\mathcal{G}^*)^\perp}\end{bmatrix} \tag{28}$$

with $A_{\mathcal{G}^*} = (A_g)_{g\in\mathcal{G}^*}$, $A_{(\mathcal{G}^*)^\perp} = (A_j)_{j\in(\mathcal{G}^*)^\perp}$.

Suppose that the following two conditions are satisfied for all $g \neq j \in \mathcal{G}$

$$\min_{\|x\|=1} \|A_g^\top A_g x\| \geq \sigma_{gg} > 0 \tag{29}$$

$$\|A_g^\top A_j\| \leq \sigma_{gj}. \tag{30}$$

Then one can show that

$$\alpha \geq \sigma_{gg} - (|\mathcal{G}^*| - 1)\sigma_{gj}, \quad \gamma \leq |\mathcal{G}^*|\sigma_{gj}. \tag{31}$$

If $\sigma_{gg} > (|\mathcal{G}^*| - 1)\sigma_{gj}$, we can then pick

$$\nu = \frac{|\mathcal{G}^*|\sigma_{gj}}{\sigma_{gg} - (|\mathcal{G}^*| - 1)\sigma_{gj}}. \tag{32}$$

It is informative to insert this bound into the conditions on κ and ν in Theorem 1. With a little bit of algebra, we are able to identify an upper bound on the number of active groups $|\mathcal{G}^*|$ that our sufficient conditions can provide guarantees for:

$$|\mathcal{G}^*| < (2-\kappa)\frac{\sigma_{gg}+\sigma_{gj}}{2\sigma_{gj}} \leq \frac{\sigma_{gg}+\sigma_{gj}}{2\sigma_{gj}}, \tag{33}$$

where the final inequality follows from recalling that $\kappa \geq 1$.

E. Extensions to the group norm with overlap

Although traditional estimation (or approximation in our terminology) errors exist for the group norm with overlap [19], no non-asymptotic structural recovery results exist. From a co-design perspective, this is somewhat dissatisfying, as we are not able to guarantee any of the measures of success that we have defined as they are all predicated on the structure (support) of the resulting approximation. However, due to the empirical success that the group norm with overlap has seen in practice [28][30], and its asymptotic structural guarantees [35], we are confident that these technical challenges can be overcome leading to stronger guarantees – this is the subject of current work.

VI. A NUMERICAL STUDY

We now return to our stylized example from Section IV. We first use the fact that $\|X\|_F = \|\text{vec}(X)\|$ to rewrite optimization (14) as

$$\begin{array}{rl} \underset{Q}{\text{minimize}} & ||\text{vec}(Y_T) - M \cdot \text{vec}(Q)||^2 + \dots \\ & \lambda \left(\sum_{a=1}^{q_2} \|\text{vec}((e^a)^\top Q)\| \right) \\ \text{s.t.} & Q = \begin{bmatrix} Q_0 & Q_1 & \dots & Q_N \end{bmatrix} \end{array} \tag{34}$$

where M is a block lower Toeplitz matrix, with

$$M_{ij} = M_{i-j} = \sum_{\substack{k,l \geq 0 \\ k+l=i-j}} J_k^\top \otimes H_l, \tag{35}$$

for all $i \leq j$, and 0 otherwise.

Unfortunately, we are as of yet unable to find interesting bounds on ν in terms of a system's state space parameters. In light of this, this section will use brute force computations to understand the limitations of the sufficient conditions that we have been able to provide, and to show that even when they do not hold, the group norm regularized optimization yields $\mathcal{A}_\mathcal{G}$-support and $\mathcal{G}$-support accurate approximations.

With these objectives in mind, we fix a time horizon of $T = 20$, and consider the following random ensemble of state space parameters. All matrices are in $\mathbb{R}^{10\times 10}$. We set $D_{12} = D_{21} = C_1 = C_2 = B_1 = B_2 = I$, and for each problem instance we let $A = (A_{ij})$ be such that each A_{ij} is drawn independently uniformly at random from the interval $[-.1, .1]$. These latter bounds on the elements of A are such that A is Hurwitz with probability one (this follows from Gershgorin's disk theorem).

Our first numerical study will be to show how σ_{gg}, σ_{gj} and the upper bound (33) vary as the controller order N increases. In particular, we randomly generated 1000 instances from the aforementioned ensemble and computed the quantities of interest for $N \in \{1, 2, 3, 4\}$. The mean values, as well as their standard deviations, are shown in Figures 2(a) and 2(b). As can be seen, the usefulness of our bounds decreases quickly as N increases; however, as we will see, our algorithm still performs very well on these design tasks. This would suggest that the sufficient condition of incoherence between groups stated in equations (29) and (30) are perhaps not the right ones to consider for control systems – indeed, due to the LTI structure of the system, there is "temporal" coherence between columns, as is evident from the block-lower Toeplitz structure of M.

We first begin by verifying that indeed our sufficient conditions provide us with the desired accuracy guarantees. We fix our controller order to $N = 1$, and aim to design a 3-group sparse controller. We ran 100 examples, beginning with $\lambda = \|Y_T\|_F$ (so as to begin with 0 active groups), and decreased λ's value in .1 increments until the number of active groups was less than or equal to 3. In order to validate our results, we conducted a brute force search over all 3-group sparse controllers in order to identify the desired controller Q^* that we were trying to approximate. Indeed in all 100 trials, we were able to recover an $\mathcal{A}_\mathcal{G}$-support accurate approximation, and 34 trials recovered a $\mathcal{G}$-support accurate estimate, all with respect to Q^*. Furthermore, for all n-group sparse $\mathcal{A}_\mathcal{G}$-support accurate estimates, with $n < 3$, we verified that they were indeed $\mathcal{G}$-support accurate with respect to the best n-group sparse controller. This is precisely the kind of behavior that is desired from a design perspective (and that is predicted by our analysis) – by increasing λ, we recover $\mathcal{G}$-support accurate approximations of structurally simpler and simpler optimal controllers.

We then repeat the experiments for $N = 3$, and despite not being able to verify any of our sufficient conditions, we observe the exact same behavior as in the $N = 1$ case – all approximations are $\mathcal{A}_\mathcal{G}$-support accurate with respect to the best 3-group sparse controller, and $\mathcal{G}$-support accurate with respect to their respective n-group sparse controllers. The number of active groups, with respect to λ, is plotted in Figure 2(c) for a specific instance of this experiment in which a $\mathcal{G}$-support accurate approximation with respect to the optimal 3-group sparse controller was recovered.

VII. CONCLUSION

We argued that that there is a natural theoretical link between sparse reconstruction and control structure co-design problems. With a particular focus on the group norm, we provided conditions for the success of a controller structure co-design optimization, and approximation bounds of its solution with respect to the underlying optimal controller. Then, through the use of an actuator regularization example, we illustrated the usefulness of these ideas, and observed that as the regularization penalty was increased, the recovered solution was $\mathcal{G}$-support accurate with respect to structurally simpler and simpler optimal controllers.

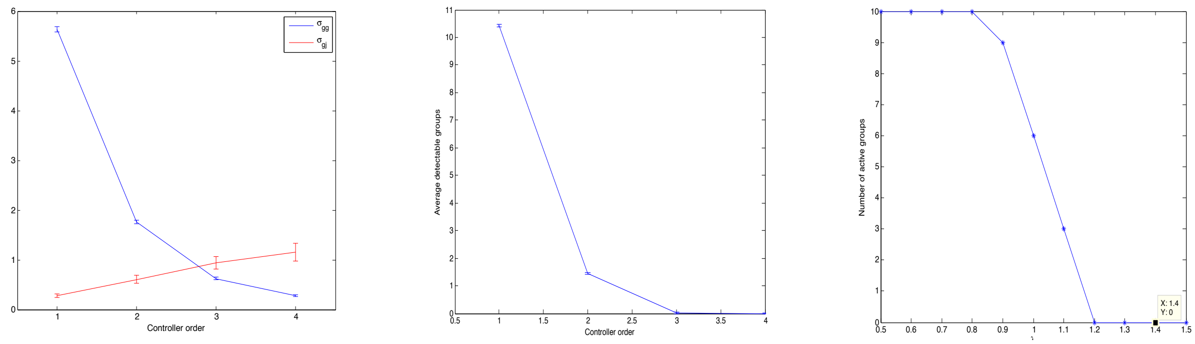
Directions for future work include developing analogous results for the group norm with overlap and for alternative cost functions (such as those expressed via linear or semi-definite constraints), making connections between the solutions to infinite dimensional optimizations and their finite dimensional approximations, and looking to exploit the structure of the design matrices to yield tighter sufficient conditions. A broader question is whether the connections made in Section III can be used to formulate a unified framework for analyzing joint identification and control schemes.

REFERENCES

[1] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, no. 1, pp. 131–147, 1968.

[2] M. Rotkowitz and S. Lall, "A characterization of convex problems in decentralized control," *Automatic Control, IEEE Transactions on*, vol. 51, no. 2, pp. 274–286, 2006.

[3] A. Mahajan, N. Martins, M. Rotkowitz, and S. Yuksel, "Information structures in optimal decentralized control," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, 2012, pp. 1291–1306.



(a) Mean and standard deviation of σ_{gg} and σ_{gj} as a function of controller order N.

(b) The maximum number of groups our sufficient conditions can provide guarantees as a function of controller order N

(c) The number of active groups recovered as a function of λ for $N = 3$. Note that each recovered group pattern was $\mathcal{G}$-support accurate.

Fig. 2: Plots from our numerical experiments on the actuator regularization problem.

[4] M. Fardad, F. Lin, and M. R. Jovanović, "Sparsity-promoting optimal control for a class of distributed systems," in *Proceedings of the 2011 American Control Conference*, 2011, pp. 2050–2055.

[5] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Trans. Automat. Control*, vol. 58, no. 9, pp. 2426–2431, 2013.

[6] G. Fazelnia, R. Madani, and J. Lavaei, "Convex relaxation for optimal distributed control problem," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, Dec 2014.

[7] A. Kalbat, R. Madani, G. Fazelnia, and J. Lavaei, "Efficient convex relaxation for stochastic optimal distributed control problem," in *Communication, Control, and Computing, IEEE 52nd Annual Allerton Conference on*, 2014.

[8] K. Dvijotham, E. Theodorou, E. Todorov, and M. Fazel, "Convexity of optimal linear controller design," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, Dec 2013, pp. 2477–2482.

[9] N. Motee and Q. Sun, "Sparsity measures for spatially decaying systems," *arXiv preprint arXiv:1402.4148*, 2014.

[10] A. Rantzer, "Distributed control of positive systems," *arXiv preprint arXiv:1203.0047*, 2012.

[11] T. Tanaka and C. Langbort, "The bounded real lemma for internally positive systems and h-infinity structured static state feedback," *IEEE transactions on automatic control*, vol. 56, no. 9, pp. 2218–2223, 2011.

[12] Y.-S. Wang, N. Matni, S. You, and J. C. Doyle, "Localized distributed state feedback control with communication delays," in *American Control Conference (ACC), 2014*. IEEE, 2014, pp. 5748–5755.

[13] Y.-S. Wang, N. Matni, and J. C. Doyle, "Localized LQR optimal control," *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 2014.

[14] Y.-S. Wang and N. Matni, "Localized distributed optimal control with output feedback and communication delays," *Communication, Control, and Computing, IEEE 52nd Annual Allerton Conference on*, 2014.

[15] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, June 2012.

[16] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, 2006.

[17] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Aug. 2010.

[18] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, pp. 805–849, 2012.

[19] M. J. Wainwright, "Structured regularizers for high-dimensional problems: Statistical and computational issues," *Annual Review of Statistics and Its Application*, vol. 1, pp. 233–253, 2014.

[20] P. Shah, B. N. Bhaskar, G. Tang, and B. Recht, "Linear system identification via atomic norm regularization," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, 2012, pp. 6265–6270.

[21] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *American Control Conference, 2001. Proceedings of the 2001*, vol. 6. IEEE, 2001, pp. 4734–4739.

[22] N. Matni and A. Rantzer, "Low-rank and low-order decompositions for local system identification," *CoRR*, vol. arXiv:1403.7175, 2014. [Online]. Available: http://arxiv.org/abs/1403.7175

[23] L. Ljung, "Some classical and some new ideas for identification of linear systems," *Journal of Control, Automation and Electrical Systems*, vol. 24, no. 1-2, pp. 3–10, 2013.

[24] V. Jonsson, A. Rantzer, and R. M. Murray, "A scalable formulation for engineering combination therapies for evolutionary dynamics of disease," in *The IEEE American Control Conference (ACC), 2014, To appear*, 2014.

[25] N. Dhingra, F. Lin, M. Fardad, and M. R. Jovanovic, "On identifying sparse representations of consensus networks," in *3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems, Santa Barbara, CA*, 2012, pp. 305–310.

[26] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.

[27] M. Fardad, F. Lin, and M. R. Jovanović, "Design of optimal sparse interconnection graphs for synchronization of oscillator networks," *arXiv preprint arXiv:1302.0449*, 2013.

[28] N. Matni, "Communication delay co-design in $\mathcal{H}_2$ decentralized control using atomic norm minimization," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, Dec 2013, pp. 6522–6529.

[29] ——, "Communication delay co-design in $\mathcal{H}_2$ distributed control using atomic norm minimization," *arXiv preprint arXiv:1404.4911*, 2014.

[30] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 433–440.

[31] N. Matni and V. Chandrasekaran, "Regularization for design," *CoRR*, vol. arXiv:1404.1972, 2014. [Online]. Available: http://arxiv.org/abs/1404.1972

[32] A. Lamperski and J. C. Doyle, "The $\mathcal{H}_2$ control problem for decentralized systems with delays," *CoRR*, vol. abs/1312.7724, 2013. [Online]. Available: http://arxiv.org/abs/1312.7724

[33] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Annals of statistics*, vol. 37, no. 6B, p. 4254, 2009.

[34] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu, *et al.*, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[35] G. Obozinski, L. Jacob, and J.-P. Vert, "Group lasso with overlaps: the latent group lasso approach," *arXiv preprint arXiv:1110.0413*, 2011.