Group LASSO Approach for Estimation and Source Localization of Forced Oscillations

Rajasekhar Anguluri, *Member, IEEE,* Lalitha Sankar, *Senior Member, IEEE,* and Oliver Kosut, *Member, IEEE*

***Abstract*—We consider a group LASSO (linear absolute shrinkage and selection operator) based estimator for jointly estimating the initial state and multiple forced oscillations (FOs), including their locations, in a power system. Through rigorous theoretical arguments, our main goal in this paper is to convey that the proposed estimator is desirable when the number of sources of FOs is sparse. Based on the observability and forced response matrices of a linearized power system model, we first develop a mutual incoherence condition. Using this condition, we derive ℓ_2-error bound on the estimation accuracy and group support recovery condition for accurately locating the sources of FO. We also establish relationship between the group LASSO based location recovery condition to that of left invertibility condition—a classical concept that determines the existence of unknown input estimators for linear systems. Finally, we validate the performance of the proposed approach via numerical simulations on IEEE 68-bus and WECC 179-bus power systems.**

***Index Terms*—Forced oscillations, group LASSO, source location, estimation.**

I. Introduction

Related work:

- Dissipating Energy Method (Chavlier, Senroy, Mani)
- Data-Driven approaches (Huang, etc)
- Statistical approaches
- frequency based approaches

The main contributions of this paper are as follows:

1) Assuming a linearized electro-mechanical model for a power system, we propose a group LASSO optimization program to simultaneously estimate the initial state and multiple FO inputs, and to recover the locations of FOs.
2) Based on the mutual incoherence condition for the joint observability and forced response matrix, we provide support recovery condition for group LASSO to accurately locate the sources of FO with high probability, and derive probabilistic bound on the estimation error.
3) We establish connection between group LASSO based support recovery conditions to that of the left-invertibility condition—a classical concept in dynamical system inversion literature—which guarantees the existence of estimators for recovering unknown inputs in linear systems.
4) Using the proposed group LASSO approach, we demonstrate that multiple forced oscillations present in IEEE 68-bus and WECC 179-bus benchmark power systems can be efficiently estimated including their source locations.

The authors are with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: {rangulur,lalithasankar,okosut}@asu.edu).

The remainder of the paper is organized as follows. Section II introduces the power system model and description of the group lasso framework. In Section III and IV, we state and prove consistency results on estimation and source location of forced oscillations, respectively. In Section V, we present the results of simulations that validate our theoretical analysis. Section VI concludes the paper.

II. Problem Setup and Preliminary Notions

A. Power System Dynamics and Measurement Models

For a multi-machine power system, the small-signal model near an operating point can be modelled using the following continuous time invariant linear dynamical system:

$$\dot{\mathbf{x}}(t) = \widetilde{\mathbf{A}}\mathbf{x}(t) + \widetilde{\mathbf{B}}\widetilde{\mathbf{u}}(t), \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector (generator rotor angle and speed deviations) and $\widetilde{\mathbf{A}}$ is the state matrix of appropriate dimension. The forced oscillations input $\widetilde{\mathbf{u}}(t) = [\tilde{u}_1(t), \ldots, \tilde{u}_m(t)] \in \mathbb{R}^m$ is injected by m potentially malfunctioned[1] generators, whose locations are encoded in the matrix $\widetilde{\mathbf{B}}$. Although in the FOs source localization literature $\mathbf{u}(t)$ is assumed to be periodic, we make no assumptions about its periodicity, except for being a deterministic signal.

Using the standard state-space discretization technique [], the discrete system dynamics of (1) along with the measurement model is given by

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\widetilde{\mathbf{u}}[k], \\ \mathbf{y}[k] &= \mathbf{C}\mathbf{x}[k] + \mathbf{v}[k], \end{aligned} \tag{2}$$

where $\mathbf{y}[k] \in \mathbb{R}^r$ is the noisy measurement of the state $\mathbf{x}[k]$, with noise $\mathbf{v}[k] \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Further, $\mathbf{A} = e^{\widetilde{\mathbf{A}}\Delta t}$ and $\mathbf{B} = (\int_{\tau=0}^{\Delta t} e^{\widetilde{\mathbf{A}}t}dt)\widetilde{\mathbf{B}}$, where Δt is the sampling time, and the output matrix $\mathbf{C} = [\mathbf{e}_1, \ldots, \mathbf{e}_r]^\mathsf{T}$ describes PMU locations in the network. Finally, we assume that the initial state $\mathbf{x}[0] := \mathbf{x}_0$ is unknown.

Let $\mathbf{y} = \begin{bmatrix}\mathbf{y}^\mathsf{T}[1] & \ldots & \mathbf{y}^\mathsf{T}[N]\end{bmatrix}^\mathsf{T} \in \mathbb{R}^{Nr}$ denote the measurements collected over the time horizon $k = 1, \ldots, N$. Analogously, define $\widetilde{\mathbf{u}} = \begin{bmatrix}\widetilde{\mathbf{u}}^\mathsf{T}[0] & \ldots & \widetilde{\mathbf{u}}^\mathsf{T}[N-1]\end{bmatrix}^\mathsf{T} \in \mathbb{R}^{Nm}$ and $\mathbf{v} = \begin{bmatrix}\mathbf{v}^\mathsf{T}[1] & \ldots & \mathbf{v}^\mathsf{T}[N]\end{bmatrix}^\mathsf{T} \in \mathbb{R}^{Nr}$. Then, from (2) $\mathbf{y}$ can be expressed as

$$\mathbf{y} = \mathbf{O}\mathbf{x}_0 + \mathbf{J}\widetilde{\mathbf{u}} + \mathbf{v}, \tag{3}$$

[1]Here, the term "malfunction" refers to a specific failed component of the generator (e.g., governor control, excitation system etc.), but not the overall generator.

where the observability matrix $\mathbf{O} \in \mathbb{R}^{Nr \times n}$ and the impulse response matrix $\widetilde{\mathbf{J}} \in \mathbb{R}^{Nr \times Nm}$ are defined as

$$\mathbf{O} := \begin{bmatrix} \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^N \end{bmatrix}; \widetilde{\mathbf{J}} := \begin{bmatrix} \mathbf{CB} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{CAB} & \mathbf{CB} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{CA}^{N-1}\mathbf{B} & \mathbf{CA}^{N-2}\mathbf{B} & \dots & \mathbf{CB} \end{bmatrix}.$$

B. Group LASSO for Initial State and Input Estimation

We assume that m_0 out of m sources are injecting the forced inputs in to the system (2), where $m_0 << m$. This is a valid assumption, since in many widely reported incidents on FOs, only a few malfunctioned controllers (one for every generator) turned out to be the sources.

In order to capture this sparsity structure among the source locations, we reshape the input $\widetilde{\mathbf{u}}$ as a concatenation of m sub-vectors, where each subvector contain the time stacked inputs at a given location. Formally, let $\mathbf{u}_j = [\widetilde{u}_j[0], \dots, \widetilde{u}_j[N-1]]^\mathsf{T}$ denote the input injected by the j-th source, $j \in \{1, \dots, m\}$, over a time interval of length N. Then, one can define a permutation/reshaping matrix $\mathbf{P}$ such that $\widetilde{\mathbf{u}} = \mathbf{P}[\mathbf{u}_j^\mathsf{T}, \dots, \mathbf{u}_m^\mathsf{T}]^\mathsf{T}$. Define $\mathbf{J} := \widetilde{\mathbf{J}}\mathbf{P} = [\mathbf{J}_1, \dots, \mathbf{J}_m]$. To emphasize that the vector $(\mathbf{x}_0, \mathbf{u})$ is the true parameter, we denote it by $(\mathbf{x}_0^*, \mathbf{u}^*)$, where $\mathbf{u}^* = [(\mathbf{u}_1^*)^\mathsf{T}, \dots, (\mathbf{u}_m^*)^\mathsf{T}]^\mathsf{T}$. Then, from (3) we have

$$\begin{aligned} \mathbf{y} &= \mathbf{O}\mathbf{x}_0^* + \widetilde{\mathbf{J}}\mathbf{P}\mathbf{u}^* + \mathbf{v} \\ &= \mathbf{O}\mathbf{x}_0^* + \sum_{j=1}^{m} \mathbf{J}_j \mathbf{u}_j^* + \mathbf{v}. \end{aligned} \tag{4}$$

Since $\mathbf{u}_j^* \neq \mathbf{0}$ if and only if the j-th source injects input, from (4) and the sparsity assumption ($m_0 << m$), it is clear that only few non-zero $\boldsymbol{\beta}_j$'s contribute to the measurement $\mathbf{y}$. Finally, notice that $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

We now state the group LASSO optimization problem to estimate the group sparse vector $(\mathbf{x}_0^*, \mathbf{u}^*)$:

$$\begin{bmatrix} \widehat{\mathbf{x}}_0 \\ \widehat{\mathbf{u}} \end{bmatrix} = \underset{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{u} \in \mathbb{R}^T}}{\arg\min} \left\{ \frac{1}{2T} \|\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \mathbf{J}\mathbf{u}\|_2^2 + \lambda_T \sum_{j=1}^{m} \|\mathbf{u}_j\|_2 \right\}. \tag{5}$$

Here, $T = Nm$ denotes the dimension of $\mathbf{y}$, and the tuning parameter $\lambda_T > 0$ penalizes unknown parameters in the model (4). We make few comments about the optimization program (5). First, it is a convex optimization problem, and there exist several computationally efficient algorithms for obtaining $\widehat{\boldsymbol{\beta}}$, including the coordinated LARS algorithm. We provide more details about numerical algorithms in Section V. Second, although the penalty is over the input vector $\mathbf{u}$, the optimization framework jointly recovers the initial state and the input. Third, due to the mixed ℓ_1/ℓ_2-norm penalty, i.e., sum of ℓ_2 norm of the subvectors $\mathbf{u}_j$s, the estimates $\widehat{\mathbf{u}}_j$ including $\widehat{\mathbf{x}}_0$ are not sparse; see []. Thus, the estimates returned by (5) are useful in applications where the vector parameters are fewer in number but each vector is dense—e.g., initial state deviations and sustained oscillations in power systems

★ WRITE ABOUT IDENTIFIABILITY ASSUMPTION/CONDITION

ADD FIGURE/BLOCK DIAGRAM OF THE ESTIMATION SETUP ★

We quantify the performance of an estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}})$ using the following error metrics:

- $\widehat{\boldsymbol{\beta}}$ is said to be ℓ_2-consistent if $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 \leq o(T)$ with probability at least $1 - c_1 \exp(-c_2 T)$, for some constants $c_1, c_2 > 0$.
- $\widehat{\mathbf{u}}$ is said to be group selection consistent if $\text{gsupp}(\widehat{\mathbf{u}}) = \text{gsupp}(\mathbf{u}^*)$ with probability at least $1 - c_1 \exp(-c_2 T)$, for some constants $c_1, c_2 > 0$; here, $\text{gsupp}(\mathbf{u}) = \{j : \mathbf{u}_j \neq \mathbf{0}, j = 1, \dots, m\}$.

Here, the ℓ_2-error bound ensures that the joint estimate $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}})$ can be made arbitrarily close to the true parameter $(\mathbf{x}_0^*, \mathbf{u}^*)$ by increasing the number of samples $T = Nm$. Instead, the group selection consistency guarantees that as $T \to \infty$, $\widehat{\mathbf{u}}$ correctly identifies the sources of FOs. non-asymptotic nature In the following sections we derive sufficient conditions, which guarantee that $\widehat{\boldsymbol{\beta}}$ obtained from (5) is both ℓ_2-estimation and group selection consistent. Moreover, our results also provide lower bounds on the measurements T required to achieve certain level of performance.

III. Initial State and Input Recovery: Estimation Consistency

Let $\widetilde{\mathcal{G}} = \{1, \dots, m\}$ be the indices of FO source locations, and $\mathcal{G} = \widetilde{\mathcal{G}} \cup \{0\}$.

IV. Initial State and Input Recovery: Location Selection Consistency

Assumption IV.1. ***(Group normalization and identifiability conditions)*** *Let $\mathbf{M}_S = \begin{bmatrix} \mathbf{O} & \mathbf{J}_S \end{bmatrix}$, where $\mathbf{J}_S$ is a sub-matrix of $\mathbf{J} = [\mathbf{J}_1, \dots, \mathbf{J}_m]$ constructed by concantenating column blocks $\mathbf{J}_j$, $j \in S$.*

(A1) The column block matrices $\mathbf{O}$ and $\mathbf{J}_j$ satisfies the group normalization condition, i.e.,

$$\max\left\{ \frac{\|\mathbf{O}\|_2}{\sqrt{T}}, \frac{\|\mathbf{J}_1\|_2}{\sqrt{T}}, \dots, \frac{\|\mathbf{J}_m\|_2}{\sqrt{T}} \right\} \leq C. \tag{6}$$

(A2) The minimum singular value of the scaled submatrix $\mathbf{M}_S$ is bounded below:

$$\gamma_{\min}\left(\frac{1}{T} \mathbf{M}_S^\mathsf{T} \mathbf{M}_S \right) \geq c_{\min} > 0. \tag{7}$$

(A3) There exists some $\alpha \in (0, 1)$, referred to as "block-mutual incoherence" parameter, such that

$$\max_{j \in S^c} \left\| \left(\mathbf{J}_j^\mathsf{T} \mathbf{M}_S\right) \left(\mathbf{M}_S^\mathsf{T} \mathbf{M}_S\right)^{-1} \right\|_2 \leq 1 - \alpha. \tag{8}$$

Let $\boldsymbol{\Gamma}_j$ satisfies $\mathbf{u}_j^* = \boldsymbol{\Gamma}_j \begin{bmatrix} (\mathbf{x}_0^*)^\mathsf{T} & (\mathbf{u}^*)^\mathsf{T} \end{bmatrix}^\mathsf{T}$, for any $j \in S$. Define

$$f(\mathbf{M}_S) = \max_{j \in S} \sum_{l \in S} \left\| \boldsymbol{\Gamma}_j \left(\mathbf{M}_S^\mathsf{T} \mathbf{M}_S\right)^{-1} \boldsymbol{\Gamma}_l^\mathsf{T} \right\|_2 \tag{9}$$

The above function extend matrix $|||\cdot|||_\infty$ norm to the case of block matrix; here $|||A|||_\infty = \max_{i=1,\dots,s} |A_{i,j}|$.

Theorem* IV.2.** ***(Location recovery) *Consider the linear model (4) satisfying assumptions (A1)–(A3). Suppose that the regularization parameter λ_T satisfies*

$$\lambda_T \geq \frac{2C\sigma}{(1-\alpha)} \left(\sqrt{\frac{N}{T}} + \sqrt{\frac{\log(m - m_0)}{T}} \right). \tag{10}$$

Then with probability at least $1-\frac{2m}{m_0(m-m_0)}$*, the group lasso estimate* (5) *has the following properties:*

1) *(No false inclusion) The optimal solution* $\widehat{\mathbf{u}}$ *is unique with its group support contained with in the true support* S *(i.e.,* $S(\widehat{\mathbf{u}}) \subseteq S(\mathbf{u}^*)$*).*
2) *(No false exclusion) Let*

$$g(\lambda_T) = \frac{2\sigma}{\sqrt{c_{min}}}\left(\sqrt{\frac{N}{T}} + \sqrt{\frac{\log m_0}{T}}\right) + \widetilde{f}(\mathbf{M}_S)\lambda_T T, \tag{11}$$

If the minimum ℓ_2 *norm of the true input* $\mathbf{u}^*$ *on its group support* S *is bounded below as* $\min_{j\in S}\left\|\mathbf{u}_j^*\right\|_2 \geq g(\lambda_T)$*. Then* $\widehat{\mathbf{u}}$ *recovers the correct support, i.e.,* $S(\widehat{\mathbf{u}}) = S(\mathbf{u}^*)$*.*

***Theorem* IV.3.** *(ℓ_2-error bound) Consider the linear model* (4) *satisfying assumptions (A1)–(A3). Let* λ_T *satisfies* (10)*. Then with probability at least* $1-\frac{2(m+1)}{(m_0+1)(m-m_0)}$*, the group lasso estimate satisfy*

$$\left\|\begin{bmatrix}\mathbf{x}_0^* - \widehat{\mathbf{x}}_0\\ \mathbf{u}^* - \widehat{\mathbf{u}}\end{bmatrix}\right\|_2 \leq \frac{2\sigma\sqrt{m_0+1}}{\sqrt{c_{min}}}\left[\sqrt{\frac{N}{T}} + \sqrt{\frac{\log(m_0+1)}{T}}\right] + \lambda_T\frac{\sqrt{m_0}}{c_{min}}.$$

V. Simulations

VI. Appendix

Additional Notation: Let $\mathbf{b} = [\mathbf{b}_1, \ldots, \mathbf{b}_q] \in \mathbb{R}^{Mq}$, where $\mathbf{b}_j \in \mathbb{R}^M$. For any set $G \subseteq \{1, \ldots, q\}$, we let $\mathbf{b}_G$ be the vector obtained by concatenating subvectors $\mathbf{b}_j$ such that $j \in G$, and similarly, for $\mathbf{b}_{G^c}$. Let $\mathcal{R}(\mathbf{b}_G) = \sum_{j\in G}\|\mathbf{b}_j\|_2$. Note that with this notation, the penalty in (5) becomes $\lambda_T\mathcal{R}(\mathbf{u})$. It is a standard result in optimization theory that the dual norm of $\mathcal{R}(\mathbf{b}_G)$, denoted as $\mathcal{R}^*(\mathbf{b}_G)$, is

$$\mathcal{R}^*(\mathbf{b}_G) := \sup_{\mathcal{R}(c)\leq 1}\mathbf{c}^\mathsf{T}\mathbf{b}_G = \max_{j\in G}\|\mathbf{b}_j\|_2. \tag{12}$$

Proof. The proof relies on the primal-dual witness (PDW) construction technique advocated by [?]. The PDW technique consists of following three steps[2]:

(a) Set $\widehat{\mathbf{u}}_{S^c} = 0$.
(b) Determine $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_S)$ by solving the following oracle sub-problem:

$$\underset{\substack{\mathbf{x}\in\mathbb{R}^n\\ \mathbf{u}_S\in\mathbb{R}^T}}{\arg\min}\left\{\frac{1}{2T}\|\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \mathbf{J}_S\mathbf{u}_S\|_2^2 + \lambda_T\mathcal{R}(\mathbf{u}_S)\right\}. \tag{13}$$

Substituting $\mathbf{x}_0 = \widehat{\mathbf{x}}_0$ and $\mathbf{u} = [\widehat{\mathbf{u}}_S^\mathsf{T}\,\mathbf{0}^\mathsf{T}]^\mathsf{T}$ in (28), and solve for the vectors $\widehat{\mathbf{z}}_S \in \partial\mathcal{R}(\widehat{\mathbf{u}}_S)$ and $\widehat{\mathbf{z}}_{S^c} \in \partial\mathcal{R}(\widehat{\mathbf{u}}_{S^c})$.
(c) Check if the strict dual feasibility condition $\mathcal{R}^*(\widehat{\mathbf{z}}_{S^c}) < 1$ holds.

The PDW construction is said to be successful if the vector $\widehat{\mathbf{z}}_{S^c}$ satisfies the strict dual feasibility condition. Suppose that the PDW construction succeeds. In the view of Lemma **??**, we observe that $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_S, \mathbf{0})$ obtained from steps (a) and (b) is the unique optimal solution of (5). Thus, all optimal input vectors are supported on the set S, i.e., $S(\widehat{\mathbf{u}}) \subset S(\mathbf{u}^*)$.

We now prove that $\mathcal{R}^*(\widehat{\mathbf{z}}_{S^c}) < 1$ with probability at least $1 - 2/\overline{s}$, where $\overline{s} = |S^c| = m - m_0$. We begin with fact that $\widehat{\mathbf{u}}_{S^c} = \widehat{\mathbf{u}}_{S^c}^* = 0$. Invoking this fact, along with the model (4), in (27a) and (28) results in the following block matrix form of KKT condition:

$$\frac{1}{T}\underbrace{\begin{bmatrix}\mathbf{O}^\mathsf{T}\mathbf{O} & \mathbf{O}^\mathsf{T}\mathbf{J}_S\\ \mathbf{J}_S^\mathsf{T}\mathbf{O} & \mathbf{J}_S^\mathsf{T}\mathbf{J}_S\end{bmatrix}}_{\mathbf{M}_S^\mathsf{T}\mathbf{M}_S}\underbrace{\begin{bmatrix}\mathbf{x}_0^* - \widehat{\mathbf{x}}_0\\ \mathbf{u}_S^* - \widehat{\mathbf{u}}_S\end{bmatrix}}_{\boldsymbol{\Delta}} + \frac{1}{T}\underbrace{\begin{bmatrix}\mathbf{O}^\mathsf{T}\\ \mathbf{J}_{S^c}^\mathsf{T}\end{bmatrix}}_{\mathbf{M}_S^\mathsf{T}}\mathbf{v} = \lambda_T\begin{bmatrix}\mathbf{0}\\ \widehat{\mathbf{z}}_S\end{bmatrix},$$

$$\frac{1}{T}\underbrace{\begin{bmatrix}\mathbf{J}_{S^c}^\mathsf{T}\mathbf{O} & \mathbf{J}_{S^c}^\mathsf{T}\mathbf{J}\end{bmatrix}}_{\mathbf{J}_{S^c}^\mathsf{T}\mathbf{M}_S}\boldsymbol{\Delta} + \frac{1}{T}\mathbf{J}_{S^c}^\mathsf{T}\mathbf{v} = \lambda_T\widehat{\mathbf{z}}_{S^c}. \tag{14}$$

From (14) and the assumed invertibility condition (7), $\boldsymbol{\Delta}$ can be solved as

$$\boldsymbol{\Delta} = -(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\mathbf{M}_S^\mathsf{T}\mathbf{v} + \lambda_T T(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\begin{bmatrix}\mathbf{0}\\ \widehat{\mathbf{z}}_S\end{bmatrix}. \tag{15}$$

Substituting this equation into (14) and simplifying yields us

$$\widehat{\mathbf{z}}_{S^c} = \underbrace{\mathbf{J}_{S^c}^\mathsf{T}\mathbf{M}_S(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\begin{bmatrix}\mathbf{0}\\ \widehat{\mathbf{z}}_S\end{bmatrix}}_{\boldsymbol{\mu}} + \underbrace{\mathbf{J}_{S^c}^\mathsf{T}\left[\mathbf{I} - \mathbf{M}_S(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\mathbf{M}_S^\mathsf{T}\right]\left(\frac{\mathbf{v}}{\lambda_T T}\right)}_{\boldsymbol{\eta}_{S^c}}. \tag{16}$$

By the triangle inequality, it follows that $\mathcal{R}^*(\widehat{\mathbf{z}}_{S^c}) \leq \mathcal{R}^*(\boldsymbol{\mu}) + \mathcal{R}^*(\boldsymbol{\eta}_{S^c})$. From the mutual incoherence condition (8), we deduce that $\mathcal{R}^*(\widehat{\mathbf{z}}_{S^c}) \leq \alpha + \mathcal{R}^*(\boldsymbol{\eta}_{S^c})$. Suppose that $\mathcal{R}^*(\boldsymbol{\eta}_{S^c}) \leq 0.5(1-\alpha)$. Then, it is easy to see that $\mathcal{R}^*(\widehat{\mathbf{z}}_{S^c}) \leq 0.5(1+\alpha) < 1$, which establishes the strict dual feasibility condition. Thus, it suffices to prove that $\mathcal{R}^*(\boldsymbol{\eta}_{S^c}) \leq 0.5(1-\alpha)$ holds with the claimed probability.

Define $\Pi_{S^\perp} = \mathbf{I} - \mathbf{M}_S(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\mathbf{M}_S^\mathsf{T}$. Then, $\boldsymbol{\eta}_{S^c} = \mathbf{J}_{S^c}^\mathsf{T}\Pi_{S^\perp}(\mathbf{v}/\lambda_T T)$, and from (12), we note that

$$\mathcal{R}^*(\boldsymbol{\eta}_{S^c}) = \max_{j\in S^c}\left\|\mathbf{J}_j^\mathsf{T}\Pi_{S^\perp}(\mathbf{v}/\lambda_T T)\right\|_2. \tag{17}$$

Following similar steps in [], an elementary calculation shows that $\left\|\mathbf{J}_j^\mathsf{T}\Pi_{S^\perp}(\mathbf{v}/\lambda_T T)\right\|_2$ is a Lipschitz with constant $C/\lambda\sqrt{T}$, where C is the group normalization bound defined in (A1). Hence, from Gaussian concentration of measure for Lipschitz functions [39], we have

$$\mathbb{P}\left(\left\|\mathbf{J}_j^\mathsf{T}\Pi_{S^\perp}\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2 \geq \mathbb{E}\left[\left\|\mathbf{J}_j^\mathsf{T}\Pi_{S^\perp}\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2\right] + \delta\right) \leq 2\exp\left(-\frac{T\lambda_T^2\delta^2}{2C^2\sigma^2}\right).$$

Instead, using the Sudakov-Fernique comparison principle [], we can bound expectation term as

$$\mathbb{E}\left[\left\|\mathbf{J}_j^\mathsf{T}\Pi_{S^\perp}\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2\right] \leq \frac{2C\sigma}{\lambda}\sqrt{\frac{N}{T}},$$

[2]The PDW construction is not a numerical recipe for solving the group LASSO (5) problem, since solving the sub-problem in step (b) requires us to know about the *unknown* support set S. However, this proof technique helps us to solve the required result.

where N is the row dimension of $\mathbf{J}_i^{\mathsf{T}}$. From the above two inequalities, we may conclude that

$$\mathbb{P}\left(\left\|\mathbf{J}_j^{\mathsf{T}}\Pi_{S^\perp}\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2 \geq 2\frac{C\sigma}{\lambda}\sqrt{\frac{N}{T}}+\delta\right) \leq 2\exp\left(-\frac{T\lambda_T^2\delta^2}{2C^2\sigma^2}\right). \quad (18)$$

Set $\delta^2 = 4\frac{\log \overline{s} C^2\sigma^2}{\lambda_T^2 T}$, where $\overline{s} := m - m_0 = |S^c|$. Then, from (10), we have $(1-\alpha) \geq \frac{2C\sigma}{\lambda_T}(\sqrt{N/T}+\sqrt{\log\overline{s}/T})$. Thus,

$$\mathbb{P}\left(\left\|\mathbf{J}_j^{\mathsf{T}}\Pi_{S^\perp}\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2 \geq 1-\alpha\right) \leq 2\exp\left(-2\log\overline{s}\right).$$

Apply union bound over all indices in S^c to note that

$$\mathbb{P}\left(\max_{j\in S^c}\left\|\mathbf{J}_j^{\mathsf{T}}\Pi_{S^\perp}\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2 \geq 1-\alpha\right) \leq 2\exp\left(-\log\overline{s}\right).$$

Finally, from (17), we have

$$\begin{aligned}\mathbb{P}[\mathcal{R}^*(\widehat{\mathbf{z}}_{S^c}) \leq 0.5(1-\alpha)] &\geq \mathbb{P}[\mathcal{R}^*(\widehat{\mathbf{z}}_{S^c}) \leq (1-\alpha)] \\ &\geq 1-2/\overline{s}, \end{aligned} \quad (19)$$

which establishes the required strict dual feasibility.

Part 2: We begin by bounding the dual norm of $\mathbf{u}_S^* - \widehat{\mathbf{u}}_S$ Let $\mathbf{\Gamma}_j$ be as in the statement of Theorem IV.2, and $\mathbf{\Delta}$ be as in (14). Then, from (15), note the following:

$$\mathbf{u}_j^* - \widehat{\mathbf{u}}_j = \mathbf{\Gamma_j}\Delta = \underbrace{-\mathbf{\Gamma_j}(\mathbf{M}_S^{\mathsf{T}}\mathbf{M}_S)^{-1}\mathbf{M}_S^{\mathsf{T}}\mathbf{v}}_{\boldsymbol{r}_{j,1}} + \lambda_T T \underbrace{\mathbf{\Gamma_j}\left(\mathbf{M}_S^{\mathsf{T}}\mathbf{M}_S\right)^{-1}\begin{bmatrix}\mathbf{0}\\ \widehat{\mathbf{z}}_S\end{bmatrix}}_{\mathbf{r}_{j,2}}. \quad (20)$$

By the triangle inequality, we have $\left\|\mathbf{u}_j^* - \widehat{\mathbf{u}}_j\right\|_2 \leq \|\boldsymbol{r}_{j,1}\|_2 + \|\boldsymbol{r}_{j,2}\|_2$. Taking maximum on both sides yields us

$$\underbrace{\max_{j\in S}\left\|\mathbf{u}_j^* - \widehat{\mathbf{u}}_j\right\|_2}_{\mathcal{R}^*(\mathbf{u}_S^* - \widehat{\mathbf{u}}_S)} \leq \underbrace{\max_{j\in S}\|\boldsymbol{r}_{j,1}\|}_{\mathcal{R}^*(\boldsymbol{r}_1)} + \underbrace{\lambda_T T \max_{j\in S}\|\boldsymbol{r}_{j,2}\|}_{\mathcal{R}^*(\boldsymbol{r}_2)}, \quad (21)$$

where $\boldsymbol{r}_i$, $\in \{1,2\}$, is formed by concatenating the vectors $\boldsymbol{r}_{i,j}$ for $j \in S$. Let $j^* = \arg\min_{j\in S}\left\|\mathbf{u}_j^*\right\|_2$, and suppose that $g(\lambda_T) \geq \mathcal{R}^*(\boldsymbol{r}_1) + \mathcal{R}^*(\boldsymbol{r}_2)$. Consider the following:

$$\left\|\mathbf{u}_j^*\right\|_2 - \|\widehat{\mathbf{u}}_j\|_2 \overset{(21)}{\leq} \mathcal{R}^*(\boldsymbol{r}_1) + \mathcal{R}^*(\boldsymbol{r}_2) \leq g(\lambda_T) < \left\|\mathbf{u}_{j^*}^*\right\|_2.$$

For the final inequality we invoked theorem hypothesis in part (2). Rearranging above inequality yields us $\|\widehat{\mathbf{u}}_j\|_2 \geq \left\|\mathbf{u}_j^*\right\|_2 - \left\|\mathbf{u}_{j^*}^*\right\|_2 > 0$, for all $j \in S$. Thus, for the LASSO estimate (5), $\widehat{\mathbf{u}}_j \neq 0$ for all $j \in S$. Instead, in part (1) we showed that $\widehat{\mathbf{u}}_j = 0$ for all $j \in S^c$. Hence, it follows that $S(\widehat{\mathbf{u}}) = S(\mathbf{u}^*)$.

We proceed to show that $g(\lambda_T) \geq \mathcal{R}^*(\boldsymbol{r}_1) + \mathcal{R}^*(\boldsymbol{r}_2)$ with high probability. Notice that both $\mathcal{R}^*(\boldsymbol{r}_1)$ and $\mathcal{R}^*(\boldsymbol{r}_2)$ are random; however, $\mathcal{R}^*(\boldsymbol{r}_2)$ is bounded above by a deterministic constant, almost surely. In fact,

$$\begin{aligned}\frac{\mathcal{R}^*(\boldsymbol{r}_2)}{\lambda_T T} &= \max_{j\in S}\left\|\sum_{l\in S}\mathbf{\Gamma}_j(\mathbf{M}^{\mathsf{T}}\mathbf{M})^{-1}\mathbf{\Gamma}_l^{\mathsf{T}}\widehat{\mathbf{z}}_l\right\|_2 \\ &\leq \max_{j\in S}\sum_{l\in S}\left\|\mathbf{\Gamma}_j(\mathbf{M}^{\mathsf{T}}\mathbf{M})^{-1}\mathbf{\Gamma}_l^{\mathsf{T}}\right\|_2\|\widehat{\mathbf{z}}_l\|_2 \\ &\leq \underbrace{\max_{j\in S}\sum_{l\in S}\left\|\mathbf{\Gamma}_j(\mathbf{M}^{\mathsf{T}}\mathbf{M})^{-1}\mathbf{\Gamma}_l^{\mathsf{T}}\right\|_2}_{\xi_2}. \end{aligned} \quad (22)$$

The final inequality follows because $\|\widehat{\mathbf{z}}_l\|_2 \leq 1$, for all $l \in S$; follows from KKT conditions.

We now upper bound $\mathcal{R}^*(\mathbf{r}_1) = \max_{j\in S}\|\mathbf{r}_{j,1}\|_2$. From Lemma VI.1, for a $j \in S$, notice that $\|\mathbf{r}_{j,1}\|_2$ is a Lipschitz with constant $1/\sqrt{Tc_{\min}}$. Thus, once again, by Gaussian concentration of measure phenomenon [], we obtain

$$\mathbb{P}\left(\|\mathbf{r}_{j,1}\|_2 \geq \mathbb{E}[\|\mathbf{r}_{j,1}\|_2] + \gamma\right) \leq 2\exp\left(-\frac{\gamma^2 T c_{\min}}{2\sigma^2}\right), \quad (23)$$

for any $\gamma > 0$. Instead, from the Sudakov-Fernique comparsion principle [], it can be shown that $\mathbb{E}[\|\mathbf{r}_{j,1}\|_2] \leq 2\sigma\sqrt{N/Tc_{\min}}$. Invoking this fact in (23) yields

$$\mathbb{P}\left(\|\mathbf{r}_{j,1}\|_2 \geq 2\sigma\sqrt{\frac{N}{Tc_{\min}}}+\gamma\right) \leq 2\exp\left(-\frac{\gamma^2 T c_{\min}}{2\sigma^2}\right).$$

Now, take union bound over all elements in S, and set $\gamma^2 = \frac{4\sigma^2\log m_0}{Tc_{\min}}$, where $|S| = m_0$, to conclude that

$$\mathbb{P}\left(\mathcal{R}^*(\mathbf{r}_2) \geq \frac{2\sigma}{\sqrt{c_{\min}}}\left(\sqrt{\frac{N}{T}}+\sqrt{\frac{\log m_0}{T}}\right)\right) \leq \frac{2}{m_0}. \quad (24)$$

Thus, from (11), (22), and (24), we can assert that

$$\mathbb{P}\left[\mathcal{R}^*(\mathbf{r}_1)+\mathcal{R}^*(\mathbf{r}_2) \leq g(\lambda_T)\right] \geq 1 - \frac{2}{m_0}$$

Finally, to obtain the probability claimed in Theorem IV.2, proceed as follows. Let $G_1 = \{\mathcal{R}^*(\widehat{\mathbf{z}}_S^c) \leq 0.5(1-\alpha)\}$ and $G_2 = \{\mathcal{R}^*(\mathbf{r}_1)+\mathcal{R}^*(\mathbf{r}_2) \leq g(\lambda_T)\}$ be two events. Then,

$$\mathbb{P}(G_1 \cap G_2) = 1 - \mathbb{P}(G_1^c \cup G_2^c) \geq 1 - [\mathbb{P}(G_1^c) + \mathbb{P}(G_2^c)].$$

Substituting (19) and (22) in the above inequality and simplifying yields the desired result. □

Proof. Suppose that the LASSO estimate (5) satisfies $\widehat{\mathbf{u}}_{S^c} = \mathbf{0}$. Then, from the KKT conditions (27a) and (28) and the fact $\mathbf{u}_{S^c}^* = \mathbf{0}$, we note[3] that

$$\underbrace{\begin{bmatrix}\mathbf{x}_0^* - \widehat{\mathbf{x}}_0\\ \mathbf{u}_S^* - \widehat{\mathbf{u}}_S\end{bmatrix}}_{\mathbf{\Delta}} = \underbrace{-(\mathbf{M}_S^{\mathsf{T}}\mathbf{M}_S)^{-1}\mathbf{M}_S^{\mathsf{T}}\mathbf{v}}_{\mathbf{r}_1} + \underbrace{\lambda_T T(\mathbf{M}_S^{\mathsf{T}}\mathbf{M}_S)^{-1}\begin{bmatrix}\mathbf{0}\\ \widehat{\mathbf{z}}_S\end{bmatrix}}_{\mathbf{r}_2},$$

Furthermore,

$$\underbrace{\begin{bmatrix}\mathbf{x}_0^* - \widehat{\mathbf{x}}_0\\ \mathbf{u} - \widehat{\mathbf{u}}\end{bmatrix}}_{\mathbf{\Delta}_{\text{full}}} = \begin{bmatrix}\mathbf{\Delta}\\ \mathbf{0}\end{bmatrix},$$

and, as a result, $\|\mathbf{\Delta}_{\text{full}}\|_2 = \|\mathbf{\Delta}\|_2$.

It now suffices to upper bound $\|\mathbf{\Delta}\|_2$. First, observe that

$$\begin{aligned}\|\mathbf{\Delta}\|_2 &\leq \|\mathbf{r}_1\|_2 + \|\mathbf{r}_2\|_2 \\ &\leq \|\mathbf{r}_1\|_2 + \lambda_T\left\|\left(\mathbf{M}_S^{\mathsf{T}}\mathbf{M}_S/T\right)^{-1}\right\|_2\|\widehat{\mathbf{z}}_S\|_2^2\end{aligned}$$

[3]A similar computation is also used in prove part (1) of Theorem IV.2.

$$\leq \|\mathbf{r}_1\|_2 + \frac{\lambda_T}{c_{min}}\sqrt{m_0} \tag{25}$$

For the final inequality, we invoked Assumption (A3) and the fact that $\|\widehat{\mathbf{z}}_S\|_2^2 = \sum_{j\in S}\|\widehat{\mathbf{z}}_j\|_2^2 \leq |S| = m_0$.

To bound $\|\mathbf{r}_1\|_2$, proceed as follows. Define $\mathbf{u}_0 = \mathbf{x}_0$. Using the arguments presented in the proof of part (2), Theorem IV.2, we can show that

$$\max_{j\in S\cup\{0\}} \left\|\mathbf{u}_j^* - \widehat{\mathbf{u}}_j\right\|_2 \geq \underbrace{\frac{2\sigma}{\sqrt{c_{\min}}}\left(\sqrt{\frac{N}{T}} + \sqrt{\frac{\log(m_0+1)}{T}}\right)}_{\zeta}$$

holds with probability at most $2/(m_0+1)$. On the other hand, since $\|\mathbf{r}_1\|_2/\sqrt{m_0+1} \leq \max_{j\in S\cup\{0\}}\left\|\mathbf{u}_j^* - \widehat{\mathbf{u}}_j\right\|_2$, from the above inequality, we have $\mathbb{P}\left(\|\mathbf{r}_1\|_2 \geq \zeta\sqrt{m_0+1}\right) \leq 2/(m_0+1)$. Finally, from (25), we have

$$\mathbb{P}\left(\|\mathbf{\Delta}\|_2 \leq \zeta\sqrt{m_0+1} + \frac{\lambda_T\sqrt{m_0}}{c_{min}}\right) \geq 1 - \frac{2}{m_0+1}. \tag{26}$$

Note that the above bound is valid because we assumed that $\widehat{\mathbf{u}}_{S^c} = \mathbf{0}$, i.e., $S(\widehat{\mathbf{u}}) \subseteq S(\mathbf{u}^*)$.

To obtained probability claimed in the statement of Theorem IV.3, proceed as follows. Define the events $G_1 = \{S(\widehat{\mathbf{u}}) \subseteq S(\mathbf{u}^*)\}$ and $G_2 = \{\|\mathbf{\Delta}\|_2 \leq \zeta\sqrt{m_0+1} + \lambda_T\sqrt{m_0}/c_{min}\}$. Since λ_T satisfies (10), from (19), we have $\mathbb{P}(G_1^c) \leq 2/(m - m_0)$. Putting together the pieces, we can see that

$$\begin{aligned}\mathbb{P}(G_1 \cap G_2) = 1 - \mathbb{P}(G_1^c \cup G_2^c) &\geq 1 - [\mathbb{P}(G_1^c) + \mathbb{P}(G_2^c)] \\ &\geq 1 - \frac{2(m+1)}{(m_0+1)(m-m_0)}.\end{aligned}$$

The proof is now complete. □

***Lemma* VI.1.** *The function* $\mathbf{v} \mapsto \left\|\mathbf{\Gamma_j}(\mathbf{M_S^\mathsf{T}M_S})^{-1}\mathbf{M_S^\mathsf{T}v}\right\|_\mathbf{2}$ *is a Lipschitz with constant* $1/\sqrt{Tc_{min}}$.

Proof. Let $\mathbf{X}_j = \mathbf{\Gamma}_j(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\mathbf{M}_S^\mathsf{T}$. For any pair of vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^T$, we have

$$\begin{aligned}\left|\,\|\mathbf{X}_j\mathbf{v}_1\|_2 - \|\mathbf{X}_j\mathbf{v}_2\|_2\,\right| &\leq \|\mathbf{X}_j(\mathbf{v}_1 - \mathbf{v}_2)\|_2 \\ &\leq \|\mathbf{X}_j\|_2\,\|(\mathbf{v}_1 - \mathbf{v}_2)\|_2\,.\end{aligned}$$

The above inequality establishes that $\|\mathbf{X}_j\mathbf{v}\|_2$ is a Lipschitz with constant $\|\mathbf{X}_j\|_2$, which can be be upper bounded as

$$\begin{aligned}\|\mathbf{X}_j\|_2^2 &= \lambda_{\max}\left(\mathbf{X}_j\mathbf{X}_j^\mathsf{T}\right) \\ &= \lambda_{\max}\left(\mathbf{\Gamma}_j(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\mathbf{M}_S^\mathsf{T}\mathbf{M}_S(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\mathbf{\Gamma}_j^\mathsf{T}\right) \\ &= \lambda_{\max}\left(\mathbf{\Gamma}_j(\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\mathbf{\Gamma}_j^\mathsf{T}\right) \\ &\leq \lambda_{\max}\left((\mathbf{M}_S^\mathsf{T}\mathbf{M}_S)^{-1}\right) = 1/Tc_{\min}.\end{aligned}$$

The inequality follows from Cauchy Interlace Theorem. □

Proposition VI.2. ***(KKT Conditions)*** *A necessary and sufficient condition for* $[\mathbf{x}_0, \mathbf{u}_1, \ldots, \mathbf{u}_m]$ *to be solution of* (5) *is*

$$-\frac{1}{T}\mathbf{O}^\mathsf{T}[\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \mathbf{J}\mathbf{u}] = \mathbf{0}, \tag{27a}$$

$$-\frac{1}{T}\mathbf{J}^\mathsf{T}[\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \mathbf{J}\mathbf{u}] + \lambda_T\mathbf{z} = \mathbf{0}, \tag{27b}$$

for all $i \in \{1, \ldots, m\}$. *Here,* $\mathbf{z} = [\mathbf{z}_1^\mathsf{T}, \ldots, \mathbf{z}_m^\mathsf{T}]^\mathsf{T} \in \partial\mathcal{R}(\mathbf{u})$, *where* $\partial\mathcal{R}(\mathbf{u})$ *denotes the sub-gradient of* $\widehat{\mathbf{u}}$, *and* $\mathbf{z}_i$ *is the subgradient of* $\|\mathbf{u}_i\|_2$ *evaluated at* $\mathbf{u}_i$, *which is given by*

$$\mathbf{z}_i = \begin{cases} \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_2}, & if\ \mathbf{u}_i = \mathbf{0} \\ \in \{\mathbf{q} : \|\mathbf{q}\|_2 \leq 1\}, & if\ \mathbf{u}_i \neq \mathbf{0} \end{cases}$$

Proof. See []. □

To prove Theorem IV.2, we need the following block matrix form of KKT condition (27b).

Corollary VI.3. ***(Partitioned form of KKT conditions)*** *For any set* $G \subset \{1, \ldots, m\}$, *let* $\mathbf{J} = \begin{bmatrix}\mathbf{J}_G & \mathbf{J}_{G^c}\end{bmatrix}$ *and* $\mathbf{u} = \begin{bmatrix}\mathbf{u}_G & \mathbf{u}_{G^c}\end{bmatrix}$. *Then, the equation* (27b) *can be rewritten as*

$$-\frac{1}{T}\begin{bmatrix}\mathbf{J}_G^\mathsf{T} \\ \mathbf{J}_{G^c}^\mathsf{T}\end{bmatrix}[\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \mathbf{J}_G\mathbf{u}_G - \mathbf{J}_{G^c}\mathbf{u}_{G^c}] + \lambda_T\begin{bmatrix}\mathbf{z}_G \\ \mathbf{z}_{G^c}\end{bmatrix} = \mathbf{0}, \tag{28}$$

where $\mathbf{z}_G$ *is obtained by concatenating sub-vectors* $\mathbf{z}_j$, $j \in G$, *and similarly, for* $\mathbf{z}_{G^c}$.

REFERENCES