Scalable and Robust State Estimation from Abundant but Untrusted Data

Ming Jin, Igor Molybog, Reza Mohammadi-Ghazi, and Javad Lavaei

***Abstract*—Power system state estimation is an important problem in grid operation that has a long tradition of research since 1960s. Due to the nonconvexity of the problem, existing approaches based on local search methods are susceptible to spurious local minima, which could endanger the reliability of the system. In general, even in the absence of noise, it is challenging to provide a practical condition under which one can uniquely identify the global solution due to its NP-hardness. In this study, we propose a linear basis of representation that succinctly captures the topology of the network and enables an efficient two-stage estimation method in case the amount of measured data is not too low. Based on this framework, we propose an identifiability condition that numerically depicts the boundary where one can warrant efficient recovery of the unique global minimum. Furthermore, we develop a robustness metric called "mutual incoherence," which underpins theoretical analysis of global recovery condition and statistical error bounds in the presence of both dense noise and bad data. The method demonstrates superior performance over existing methods in terms of both estimation accuracy and bad data robustness in an array of benchmark systems. Above all, it is scalable to large systems with more than 13,000 buses and can achieve accurate estimation within a minute.**

***Index Terms*—Power system state estimation, statistical analysis, robust learning, smart grid**

I. Introduction

Power system state estimation (SE) is conducted on a regular basis (e.g., every few minutes) to monitor the state of the grid by collecting and filtering a wealth of sensor data from transmission and distribution infrastructures [1], [2]. The state estimate presents system operators with essential information about the real-time operating status to improve situational awareness, make economic decisions, and take contingency actions in response to potential threats that could engender the grid reliability [3].

Due to the nonlinearity of the alternating-current (AC) grid physics, solving the set of power flow equations that arise from sensor measurements is known to be NP-hard for both transmission and distribution networks [4], [5]. As a result, there is a long tradition of studying this problem [2], [6]–[16]. At a high level, these methods are evaluated against multiple key criteria, including **(i)** accuracy (e.g., linearization/approximation of the nonlinear law of physics and its side effect on losing important information), **(ii)** robustness (e.g., to random/adversarial bad data, model mismatch, topological errors), and **(iii)** scalability (i.e., computational/memory requirements to solve for large-scale systems). We provide a summary of the existing methods below, and refer the reader to [17] and [15] for a more comprehensive review.

†This work was supported by the ONR grant N00014-17-1-2933 and NSF Award 1807260, ARO grant W911NF-17-1-0555, and AFOSR grant FA9550-17-1-0163. The authors are with the Department of Industrial Engineering and Operation Research, University of California, Berkeley, CA 94710. Emails: {jinming,igormolybog,mohammadi,lavaei}@berkeley.edu

A. Background and related work

The current practice in the power industry relies on a set of linearization and/or Newton's methods that are originally developed in 1960s [2], [6], [7]. The Newton's method has been employed to solve the nonlinear least square (NLS) SE and has quadratic convergence whenever the initial point is sufficiently close to the true solution [6]. However, the estimator is prone to outliers and sparse noise/errors, which can arise from sensor faults, topological errors [18]–[21], or adversarial attack [22]–[24]. To deal with large and sparse noise, one common approach is to perform bad data detection (BDD) on residual errors [25], [26]. This method relies on the statistical assumptions of the random errors (e.g., mean-zero and independent Gaussian distributions) and is only effective when the estimation from the Newton algorithm is close enough to the ground truth [2]. Alternatively, by redesigning the cost functions, robust estimators such as the least-absolute value (LAV) (a.k.a., ℓ_1 loss), the least median of squares, or Huber's estimator have been employed [2], [8], [9], [27]–[30]. A major drawback of the above local search methods is the vulnerability to spurious local minima, which are those points that satisfy first- and second-order optimality conditions but are not the global minimum [30]–[32]. This is a major issue that can potentially cause a tremendous danger to the operation of the system, and it is difficult to distinguish a spurious local minimum that is consistent with data from the ground truth. Even though some recent works have shed light on the possibility of the non-existence of local minima in certain scenarios [33], the conditions are difficult to verify for SE [30].

Apart from local search algorithms mentioned above, several advanced optimization techniques have been proposed, such as particle swarm optimization [34], holomorphic embedding load flow method [35], homotopy continuation methods [36], feasible point pursuit [37], composite optimization [38], iterative mixed objective convex program [39], and algorithms for solving variational inequalities involving monotone operators [31]. A comprehensive review of these methods can be found in [15], [17].

The technique of convexification and semidefinite programming (SDP) relaxation is a powerful tool to tackle polynomial optimization, which arise from several areas such as graph

theory, signal processing, and power systems [40]–[45]. Recently, the SDP relaxation technique has been applied to SE following its success for the optimal power flow problem [46], which demonstrate satisfactory numerical performance even in the presence of topological errors and bad data [11], [13], [16], [20], [47]. Also, [24] analyzes the vulnerabilities of AC SE against potential cyber attacks. While SDP relaxation is a promising approach with numerical success, this method requires that the solution to be rank-1 to recover the true state. Since the SDP relaxation is not always exact, one needs to add an extra rank penalty to the objective function (e.g., nuclear norm [11] or custom-designed norm [13], [16]), which would make the solution near-global optimal. The existing theoretical analyses of the SDP technique have not considered the bad data detection [13], [16]. Furthermore, the positive semidefinite constraint of the SDP technique limits the applicability of this method to large-scale problems, since the most common conic numerical algorithms scale on the order of $O(n^6)$, where n is the number of variables.

B. Contributions

We propose a method to solve large-scale AC SE with linear programming (LP) or quadratic programming (QP) that finds the correct state and is robust to sparse bad data, provided that the amount of measured data is relatively high. A new basis of representation is proposed, which is related to, but different from, the two dominant complex number representations used in power flow equations (i.e., polar coordinates and rectangular representation). This basis fully captures the properties of the power grid topology, which leads to efficient SE algorithms. Furthermore, we develop a SE identifiability condition to guarantee that no spurious local minimum exists in SE. We also perform theoretical analysis on the recovery condition of the true state in the presence of sparse bad data with statistical bounds on the estimation error.

The paper is organized as follows. The linear basis of representation is introduced in Sec. II-B, together with the measurement models and some key definitions to facilitate theoretical analysis. The two-stage estimator is introduced in Sec. III, whose performance is analyzed in Sec. IV. Sec. V includes numerical evaluations of the proposed methods on benchmark systems. Conclusion is drawn in Sec. VI. All proofs have been delegated to the appendix for the interested readers without interrupting the flow of the presentation.

II. Power System AC-model

A. Notations

Vectors are shown by bold letters, and matrices are shown by bold and capital letters. Let x_i denote the i-th element of vector $\boldsymbol{x}$. We use $\mathbb{R}$ and $\mathbb{C}$ to show the sets of real and complex numbers. The set of indices $\{1, 2, ..., m\}$ is denoted by $[m]$. The cardinality $|\mathcal{J}|$ of a set $\mathcal{J}$ is the number of elements in the set. The support $\mathrm{supp}(\boldsymbol{x})$ of a vector $\boldsymbol{x}$ is the set of indices of the nonzero entries of $\boldsymbol{x}$. For a set $\mathcal{J} \subset [m]$, we use $\mathcal{J}^c = [m] \setminus \mathcal{J}$ to denote its complement. We use $\boldsymbol{A}_{\mathcal{J}}$ to denote the submatrix formed by the rows of $\boldsymbol{A}$ indexed by $\mathcal{J}$. The symbol $(\cdot)^\top$ represents the transpose operator. We use $\Re(\cdot)$, $\Im(\cdot)$ and $\mathrm{Tr}\,(\cdot)$ to denote the real part, imaginary part and trace of a scalar/matrix. The imaginary unit is denoted as i. The notations $\angle x$ and $|x|$ indicate the angle and magnitude of a complex scalar. For a convex function $g(\boldsymbol{x})$, we use $\nabla g(\boldsymbol{x})$ to denote its subgradient. The inner product between two vectors is denoted by $\langle \cdot, \cdot \rangle$. The notations $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$ represent the 1-norm, 2-form and ∞-norm of $\boldsymbol{x}$. We use $\mathbb{E}$ to denote the expectation operator of a random variable.

B. Power system modeling

We model the electric grid as a graph $\mathcal{G} \coloneqq \{\mathcal{N}, \mathcal{L}\}$, where $\mathcal{N} \coloneqq [n_b]$ and $\mathcal{L} \coloneqq [n_l]$ represent its set of buses and branches. Each branch $\ell \in \mathcal{L}$ that connects bus k and bus j is characterized by the branch admittance $y_\ell = g_\ell + \mathrm{i} b_\ell$ and the shunt admittance $y_\ell^{\mathrm{sh}} = g_\ell^{\mathrm{sh}} + \mathrm{i} b_\ell^{\mathrm{sh}}$, where g_ℓ (resp., g_ℓ^{sh}) and b_ℓ (resp., b_ℓ^{sh}) denote the (shunt) conductance and susceptance, respectively. Since $g_\ell^{\mathrm{sh}} \ll b_\ell^{\mathrm{sh}}$ in practice, we set it to zero in the subsequent description. In addition, to avoid duplicate definitions, each line $\ell \coloneqq (k, j)$ is assigned with a unique direction from bus k (i.e., *from* end, given by $f(\ell) \coloneqq k$) to bus j (i.e., *to* end, given by $t(\ell) \coloneqq j$). We also use $\ell : \{k, j\}$ to denote a line ℓ with the direction of either (k, j) or (j, k).

The power system state is described by the complex voltage vector $\boldsymbol{v} = \left[v_1, ..., v_{n_b}\right]^\top \in \mathbb{C}^{n_b}$, where $v_k \in \mathbb{C}$ is the complex voltage at bus $k \in \mathcal{N}$ with magnitude $|v_k|$ and phase $\theta_k \coloneqq \angle v_k$. Given the complex voltages, by Ohm's law, the complex current injected into line $\ell : \{k, j\}$ at bus k is given by:

$$i_{kj} = y_\ell (v_k - v_j) + \frac{\mathrm{i}}{2} b_\ell^{\mathrm{sh}} v_k.$$

Defining $\theta_{kj} \coloneqq \theta_k - \theta_j$, one can write the power flow from bus k to bus j as

$$p_{kj}^{(\ell)} = |v_k|^2 g_\ell - |v_k||v_j|(g_\ell \cos\theta_{kj} - b_\ell \sin\theta_{kj}),$$
$$q_{kj}^{(\ell)} = -|v_k|^2 (b_\ell + \tfrac{1}{2} b_\ell^{\mathrm{sh}}) + |v_k||v_j|(b_\ell \cos\theta_{kj} - g_\ell \sin\theta_{kj}),$$

and active (reactive) power injections at bust k,

$$p_k = \sum_{\ell : \{k,j\}} p_{kj}^{(\ell)}, \qquad q_k = \sum_{\ell : \{k,j\}} q_{kj}^{(\ell)}. \tag{1}$$

The above formulas are based on polar coordinates of complex voltages, where measurements are nonlinear functions of voltage magnitudes and phases. Another popular representation uses rectangular coordinates of complex numbers, where measurements are expressed as quadratic functions of the real and imaginary parts of voltages (see [48, Chap. 1] for more details).

C. Linear basis of representation

In this paper, we introduce a new basis of representation, where measurements can be expressed as *linear combinations* of the quantities derived from bus voltages. Specifically, for a given system $\mathcal{G}$, we introduce two groups of variables:

1) voltage magnitude square, $x_k^{\mathrm{mg}} \coloneqq |v_k|^2$, for each bus $k \in \mathcal{N}$, and
2) real and imaginary parts of complex products, denoted as $x_\ell^{\mathrm{re}} \coloneqq \Re(v_i v_j^*)$ and $x_\ell^{\mathrm{im}} \coloneqq \Im(v_i v_j^*)$, respectively, for

each line $\ell = (i, j)$. Note that there is only one set of variables x_ℓ^{re} and x_ℓ^{im} for each line.

Using this representation, we can re-derive various types of power and voltage measurements (without noise) as follows:

- *Voltage magnitude square:* The voltage magnitude square at bus $k \in \mathcal{N}$ is simply x_k^{mg} by definition.
- *Branch power flows:* For each line $\ell = (i, j)$, the real and reactive power flows from bus i to bus j and in the reverse direction are given by:

$$\begin{aligned} p_{ij}^{(\ell)} &= g_\ell x_i^{\text{mg}} - g_\ell x_\ell^{\text{re}} - b_\ell x_\ell^{\text{im}} \\ q_{ij}^{(\ell)} &= -(b_\ell + \tfrac{1}{2} b_\ell^{\text{sh}}) x_i^{\text{mg}} + b_\ell x_\ell^{\text{re}} - g_\ell x_\ell^{\text{im}} \\ p_{ji}^{(\ell)} &= g_\ell x_j^{\text{mg}} - g_\ell x_\ell^{\text{re}} + b_\ell x_\ell^{\text{im}} \\ q_{ji}^{(\ell)} &= -(b_\ell + \tfrac{1}{2} b_\ell^{\text{sh}}) x_j^{\text{mg}} + b_\ell x_\ell^{\text{re}} + g_\ell x_\ell^{\text{im}} \end{aligned}$$

- *Nodal power injection:* The power injection at bus node k consists of real and reactive powers, where:

$$p_k = \sum_{k \in \ell} g_\ell x_k^{\text{mg}} - \sum_{k \in \ell} g_\ell x_\ell^{\text{re}} - \left(\sum_{f(\ell)=k} b_\ell - \sum_{t(\ell)=k} b_\ell \right) x_\ell^{\text{im}}$$

$$q_k = -\left(\sum_{k \in \ell} b_\ell + \tfrac{1}{2} b_\ell^{\text{sh}} \right) x_k^{\text{mg}} + \sum_{k \in \ell} b_\ell x_\ell^{\text{re}} - \left(\sum_{f(\ell)=k} g_\ell - \sum_{t(\ell)=k} g_\ell \right) x_\ell^{\text{im}},$$

where $\sum_{k \in \ell}$ is the sum over all lines $\ell \in \mathcal{L}$ that are connected to k, $\sum_{f(\ell)=k}$ is the sum over all lines ℓ where $f(\ell) = k$, and similarly, $\sum_{t(\ell)=k}$ is the sum over all lines ℓ where $t(\ell) = k$. Equivalently, we can use (1) to combine the branch power flows defined above.

Thus, each customary measurement in power systems that belongs to one of the above *measurement types* can be represented by a linear function[1]:

$$m_i(\boldsymbol{x}) = \boldsymbol{a}_i^\top \boldsymbol{x}_\natural, \tag{2}$$

where $\boldsymbol{a}_i \in \mathbb{R}^{n_x}$ is the vector for the i-th noiseless measurement and $\boldsymbol{x}_\natural = (\{x_k^{\text{mg}}\}_{k \in \mathcal{N}}, \{x_\ell^{\text{im}}, x_\ell^{\text{im}}\}_{\ell \in \mathcal{L}}) \in \mathbb{R}^{n_x}$ is the regression vector. By collecting all the sensor measurements in a vector $\boldsymbol{m} \in \mathbb{R}^{n_m}$, we have

$$\boldsymbol{m} = \boldsymbol{A} \boldsymbol{x}_\natural, \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{n_m \times n_x}$ is the sensing matrix with rows $\boldsymbol{a}_i^\top$ for $i \in [n_m]$. Fig. 1 illustrates the sensing equation (3) for a simple 3-bus system.

It is worth mentioning that the linear basis introduced above is different from DC modeling of measurements, because the expression is *exact* for the AC model. This parametrization is inspired by the semidefinite relaxation approach for power system optimization [11], [13], [16], [20], [47], and it efficiently exploits the sparsity of the network (more on this in Sec. IV-A).

[1]It is straightforward to include linear PMU measurements in our analysis as well using the relation $\tan \theta_{ij} = x_\ell^{\text{im}} / x_\ell^{\text{re}}$ for each line $\ell = (i, j)$, assuming we have a pair of PMUs on each end of a branch.

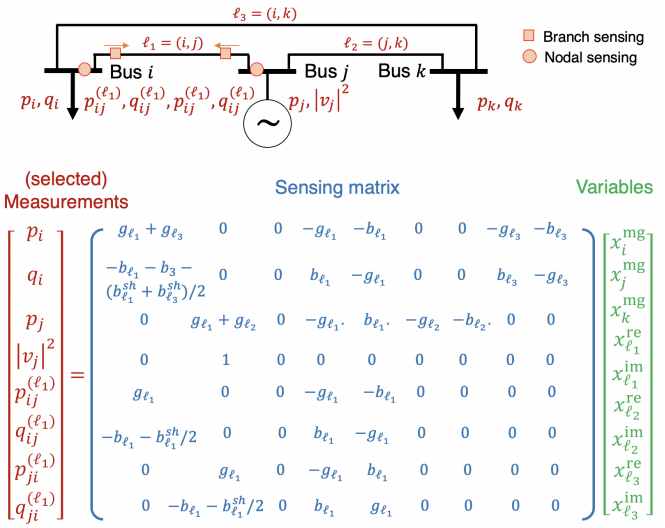


Fig. 1: Illustration of the sensing equation (3) for a 3-bus system. A selected set of measurements are considered, namely nodal injections at buses i and j, voltage magnitude square at bus j, and branch power flows along line $\ell_1 = (i, j)$. Note that one can choose the set of regression variables based on the availability of measurements, as long as each measurement can be fully represented by the chosen set of variables. For instance, we can omit the variables $x_{\ell_3}^{\text{re}} := \Re(v_i v_k^*)$ and $x_{\ell_3}^{\text{im}} := \Im(v_i v_k^*)$ by simultaneously excluding measurements $p_{ij}^{(\ell_3)}, q_{ij}^{(\ell_3)}, p_{ji}^{(\ell_3)}, q_{ji}^{(\ell_3)}$ and p_i, q_i, p_k, q_k, since they all rely on the omitted variables.

D. Measurement model

To perform SE, the supervisory control and data acquisition (SCADA) system collects measurements on power flows and complex voltages at key locations instrumented with sensors. This process is subject to both ubiquitous sensor noise and randomly occurring sensor faults. We consider the measurement model as follows:

$$\boldsymbol{y} = \boldsymbol{A} \boldsymbol{x}_\natural + \mathbf{w}_\natural + \boldsymbol{b}_\natural, \tag{4}$$

where $\boldsymbol{A} \in \mathbb{R}^{n_m \times n_x}$ and $\boldsymbol{x}_\natural \in \mathbb{R}^{n_x}$ are the sensing matrix and the true regression vector in (3), $\mathbf{w}_\natural \in \mathbb{R}^{n_m}$ denotes random noise, and $\mathbf{b}_\natural \in \mathbb{R}^m$ is the bad data error that accounts for sensor failures or adversarial attacks [24]. Let $\mathcal{J} := \text{supp}(\boldsymbol{b}) \subset [n_m]$ denote the support of the bad data $\boldsymbol{b}$. We introduce the following properties to characterize the sensing matrix $\boldsymbol{A}$.

Definition 1 (Lower eigenvalue)**.** *Let $\boldsymbol{Q}_\mathcal{J} := \begin{bmatrix} \boldsymbol{A} & \boldsymbol{I}_\mathcal{J}^\top \end{bmatrix}$, where $\boldsymbol{I}_\mathcal{J}$ consists of the $\mathcal{J}$ rows of the identity matrix $\boldsymbol{I} \in \mathbb{R}^{n_m \times n_m}$, and let $\boldsymbol{A}_{\mathcal{J}^c}$ be the submatrix of $\boldsymbol{A}$ with rows indexed by $\mathcal{J}^c$. Then, the lower eigenvalue $C_{\min}(\mathcal{J})$ for a given corruption support $\mathcal{J}$ is defined as the lower bound:*

$$\min \left\{ \lambda_{\min} \left(\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J} \right), \lambda_{\min} \left(\boldsymbol{A}_{\mathcal{J}^c}^\top \boldsymbol{A}_{\mathcal{J}^c} \right) \right\}, \tag{5}$$

where $\lambda_{\min}(\boldsymbol{X})$ denotes the smallest eigenvalue of $\boldsymbol{X}$.

The value $C_{\min}(\mathcal{J})$ characterizes the influence of bad data

on the identifiability of $\boldsymbol{x}_\natural$. If $C_{\min}(\mathcal{J})$ is strictly positive, and one can accurately detect the support of bad data (a.k.a., support recovery), then it would be possible to obtain a good estimation of $\boldsymbol{x}_\natural$ with only the clean data in $\mathcal{J}^c$. Typically, the bad data due to sensor faults are randomly located, so if only a small amount of sensors are grossly corrupted (i.e., $|\mathcal{J}| < |\mathcal{J}^c|$), then the first term in (5) will be smaller than the second term. As we will see in Sec. IV, the first term is relevant for the case with dense noise $\boldsymbol{w}_\natural$.

The next property turns out to be critical for BDD.

Definition 2 (Mutual incoherence)**.** *Given a set $\mathcal{J} \subset [m]$ and its complement $\mathcal{J}^c \coloneqq [m] \setminus \mathcal{J}$, let the pseudoinverse of $\boldsymbol{A}_{\mathcal{J}^c}$ be denoted as $\boldsymbol{A}_{\mathcal{J}^c}^{+} = (\boldsymbol{A}_{\mathcal{J}^c}^\top \boldsymbol{A}_{\mathcal{J}^c})^{-1} \boldsymbol{A}_{\mathcal{J}^c}^\top$. Then, the mutual incoherence parameter $\rho(\mathcal{J})$ is defined to be:*

$$\rho(\mathcal{J}) = \|\boldsymbol{A}_{\mathcal{J}^c}^{\top +} \boldsymbol{A}_{\mathcal{J}}^\top\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the matrix infinity norm (i.e., the maximum absolute column sum of the matrix).

The name "mutual incoherence" originates from the compressed sensing literature [49]–[52]. However, our definition is different because it measures the alignment of the sensing directions of the corrupted measurements (i.e., $\boldsymbol{A}_{\mathcal{J}}$) with those of the clean data (i.e., $\boldsymbol{A}_{\mathcal{J}^c}$). If these directions are misaligned (a.k.a., incoherent), then the value $\rho(\mathcal{J})$ is low and therefore it is likely to uncover the support of bad data. In general, the smaller the number of bad data measurement is, the more likely that $\rho(\mathcal{J})$ is small. As our analysis will show, if this value is strictly less than 1, then we can provably recover the support of the bad data.

Because the sensor data are of different types and scales, we make a normalization assumption.

Definition 3 (Measurement normalization)**.** *Each row of $\boldsymbol{A}$ is normalized as*

$$\|\boldsymbol{a}_i\|_2^2 = 1, \qquad \forall i \in [n_m] \tag{6}$$

where $\boldsymbol{a}_i$ is the i-th row of $\boldsymbol{A}$.

This condition is straightforward to implement in practice, since one can arbitrarily rescale the given coefficients of each measurement equation. This is also known as preconditioning, which assists with both the numerical stability and the statistical performance of regression.

III. Two-Stage State Estimation

This section describes the proposed two-stage state estimation method, where both stages are linear or quadratic regression problems.

A. Stage 1: Estimation of $\boldsymbol{x}_\natural$

In the first stage, the goal is to estimate $\boldsymbol{x}_\natural$ from a set of noisy and corrupted measurements $\boldsymbol{y}$. We consider two cases separately. In the first case, the dense noise is negligible, i.e., $\boldsymbol{w}_\natural = \boldsymbol{0}$, and we only need to consider the sparse measurement corruption $\boldsymbol{b}$. Under some conditions to be specified in Sec. IV, one can exactly recover the underlying vector $\boldsymbol{x}_\natural$.

Case 1: Sparse corruption but no dense noise (i.e., $\boldsymbol{w} = \boldsymbol{0}$)

In this case, the measurements are given by $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_\natural + \boldsymbol{b}_\natural$. To estimate $\boldsymbol{x}_\natural$, we solve the following program:

$$\min_{\boldsymbol{x}\in\mathbb{R}^{n_x}, \boldsymbol{b}\in\mathbb{R}^{n_m}} \|\boldsymbol{b}\|_1, \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{y}. \tag{S1-L1}$$

Briefly, if the lower eigenvalue is bounded away from 0 (i.e., $C_{\min}(\mathcal{J}) > 0$) and the mutual incoherence is less than 1 (i.e., $\rho(\mathcal{J}) < 1$), then we can faithfully recover $\boldsymbol{x}_\natural$ and $\boldsymbol{b}_\natural$ from the above program.

Case 2: Sparse corruption and dense noise

In this case, the dense noise cannot be ignored, and the measurements are given by (4). We perform the estimation by solving the following LASSO-style optimization:

$$\min_{\boldsymbol{b}\in\mathbb{R}^{n_m}, \boldsymbol{x}\in\mathbb{R}^{n_x}} \tfrac{1}{2n_m}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_1, \tag{S1-LASSO}$$

where $\lambda > 0$ is the regularization coefficient. Due to the existence of dense noise, it is no longer possible to exactly recover the true $\boldsymbol{x}_\natural$; however, if the magnitudes of the dense noise are small, then we can still have good statistical bounds on the estimation error.

To detect and remove bad data, we first estimate the linear basis by solving either (S1-L1) or (S1-LASSO). This automatically produces a bad data vector estimation $\boldsymbol{b}$. If the value of any entry of $\boldsymbol{b}$ is larger than a threshold, namely 0.1 in the experiments, then we classify the corresponding measurement as "bad data." If there is a topology error, then the bad data tend to be localized on a line or a bus. We can observe the topological distribution of bad data vector to determine if there are such occurrences.

B. Stage 2: Recovery of $\boldsymbol{v}$

The goal of the second stage is to recover the underlying system voltage $\boldsymbol{v}$ from the estimation $\hat{\boldsymbol{x}}$ from stage 1. First, we transform $\hat{\boldsymbol{x}}$ into estimations of voltage magnitudes and phase differences:

- The voltage magnitude at each bus $k \in \mathcal{N}$ is estimated as $|\hat{v}_k| = \sqrt{\hat{x}_k^{\text{mg}}}$;
- The phase difference along each line $\ell = (i, j)$ is estimated as $\hat{\theta}_{ij} = \arctan \hat{x}_\ell^{\text{im}} / \hat{x}_\ell^{\text{re}}$.

To obtain the phase estimation at each bus, we solve the least-squares problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{n_b}} \sum_{\ell=(i,j)} (\theta_i - \theta_j - \hat{\theta}_{ij})^2, \tag{S2-θ}$$

which has a closed-form solution. To delve into this, let $\boldsymbol{\theta}_\Delta$ be a collection of $\hat{\theta}_{ij}$, and $\boldsymbol{L} \in \mathbb{R}^{n_\ell \times n_b}$ be a sparse matrix with $L(\ell, i) \coloneqq 1$ and $L(\ell, j) \coloneqq -1$ for each line $\ell = (i, j)$ and zero elsewhere. Then, the solution for (S2-θ) is given by:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{L}^\top \boldsymbol{L})^{-1} \boldsymbol{L}^\top \boldsymbol{\theta}_\Delta. \tag{7}$$

Finally, we can reconstruct $\hat{\boldsymbol{v}}$ by definition:

$$\hat{v}_k = |\hat{v}_k| e^{\mathrm{i}\hat{\theta}_k}, \qquad k \in \mathcal{N}. \tag{8}$$

If the regression vector from stage 1 is exact, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{x}_\natural$, then we can accurately recover the system state $\hat{\boldsymbol{v}} = \boldsymbol{v}$. Even if the $\hat{\boldsymbol{x}}$ is not exact, the second stage estimator (S2-θ) has strong properties to control the estimation error and ensure that the errors in $\hat{\theta}_{ij}$ will not propagate along the branches.

IV. Theoretical Analysis

This section presents several theoretical analysis for the proposed framework. First, a condition for AC SE identifiability is presented. Then, we discuss the conditions under which accurate recovery of the true state is guaranteed. Furthermore, we present a novel statistical analysis of the recovery condition using concentration bounds.

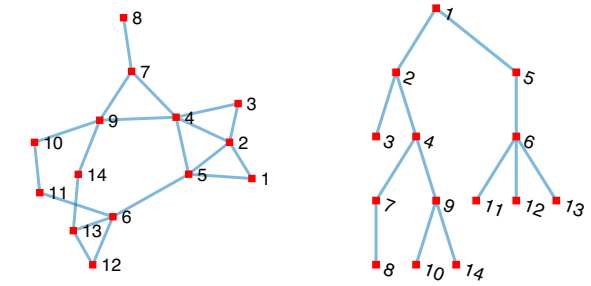
A. Identifiability condition

Due to the nonconvexity of NLS, the existence of spurious local minima in SE is well-recognized, which makes it difficult to analyze whether the true state can be uniquely identified based on a given set of clean measurements (i.e., $\boldsymbol{w} = \boldsymbol{b} = \boldsymbol{0}$). Because SE can be formulated as a quadratic sensing problem, the results from the low-rank compressed sensing community seem to be directly applicable, which rely on a condition called restricted isometry property (RIP) (e.g., see [33], [53], [54]). The main result from this line of research indicate that if RIP of the sensing system is small enough, then every local minimum is also a global minimum [33], [54]. However, numerical results indicate that the condition is often too stringent to be satisfied for SE. It is also possible to characterize an "essentially strongly convex region" around the true solution, where any initial point converges to the true solution by local search [32], or to delineate a recovery region where the rank penalty leads to exact rank-1 solution [13], [16]. However, they all depend on the location of the true solution and the condition is hard to check numerically. The following theorem provides a condition similar to the DC-approximation results but for the AC SE. Without loss of generality, assume that the power network $\mathcal{G}$ is connected.

Theorem 4. *In the absence of noise (i.e., $\boldsymbol{w} = \boldsymbol{b} = \boldsymbol{0}$), one can uniquely identify the true state of the power grid if there exists a spanning tree $\mathcal{T}_{\text{span}}$ such that the set of measurements on the spanning tree (not including nodal injections) forms a matrix $\boldsymbol{A}_{\text{span}}$ that has full column rank (i.e., the null space of $\boldsymbol{A}_{\text{span}}$ is zero).*

We do not include nodal injections in order to avoid dealing with variables on lines that do not have any measurements. For a given set of clean measurements, if the set satisfies the above identifiability condition, then there is one and only one global optimal for the Stage 1 estimator. Furthermore, this global optimal can be used in Stage 2 to recover the true state of the system.

Theorem 4 is inspired by the literature on false data injection attacks (FDIA) [22], [24], but the main difference is that we work through a completely different basis of variables. In FDIA, usually one works with the complex voltage directly. In that case, it is cumbersome to derive a necessary/sufficient condition for identifiability, because it is often algorithm dependent. In contrast, due to the proposed linear basis, it is possible to provide a numerically-verifiable condition for checking whether the underlying set of clean measurements are enough to identify the actual state of the system. A spanning tree of a graph can be found in linear time by either depth-first search or breadth-first search. For instance, Fig. 2 illustrates a spanning tree of the IEEE 14 bus system. Moreover, it is a strong result in the sense that it does not depend on the numerical value of the true state (i.e., universally applicable).



(a) full network (b) spanning tree

Fig. 2: Network topology of the IEEE 14 bus system. A sufficient condition for identifiability is the existence of a spanning tree $\mathcal{T}_{\text{span}}$ (shown on the right) where the measurements form a nonsingular sensing matrix $\boldsymbol{A}_{\text{span}}$.

TABLE I: Empirical evaluation of the identifiability condition (Y/N) in Theorem 4. Measurement sets include: M1: $\{|v_k|\}_{k\in\mathcal{N}}$ and $\{p_{ij}^{(\ell)}, q_{ij}^{(\ell)}\}_{\ell\in\mathcal{T}_{\text{span}}}$; M2: $\{|v_k|\}_{k\in\mathcal{N}}$ and $\{p_{ij}^{(\ell)}, p_{ji}^{(\ell)}\}_{\ell\in\mathcal{T}_{\text{span}}}$; M3: $\{|v_k|\}_{k\in\mathcal{N}}$ and $\{q_{ij}^{(\ell)}, q_{ji}^{(\ell)}\}_{\ell\in\mathcal{T}_{\text{span}}}$; M4: $\{p_{ij}^{(\ell)}, p_{ji}^{(\ell)}, q_{ij}^{(\ell)}\}_{\ell\in\mathcal{T}_{\text{span}}}$; M5: $\{p_{ij}^{(\ell)}, q_{ji}^{(\ell)}, q_{ij}^{(\ell)}\}_{\ell\in\mathcal{T}_{\text{span}}}$. Note that the number of measurements is bounded by $3 \times n_b$.

	M1	M2	M3	M4	M5
30 Bus	Y	N	N	N	Y
57 Bus	Y	N	N	N	Y
118 Bus	Y	N	N	N	Y
300 Bus	Y	N	N	N	Y
1354 Bus	Y	N	N	N	Y
2848 Bus	Y	N	N	N	Y

Note that "identifiability" is a stronger condition than "observability." The latter is usually based on DC-approximation (see [55] and [2, Chap. 4]), where a system is observable as long as the measurement Jacobian is nonsingular. However, for AC SE, "observability" implies the *existence* of a method (with potentially exponential-time complexity) to infer the state uniquely, but polynomial-time methods (such as local search) may have a number of spurious local minima [13], [24]); however, "identifiability" indicates that the given method is guaranteed to efficiently recover the unique system state. In other words, "identifiability" is a sufficient condition for "observability."

Remark 5. *In general, there are $n_b - 1$ edges for a spanning tree with n_b nodes; hence, by the construction of the linear representation in Sec. II-B, there are in total $3 \times n_b - 2$ number of variables on the spanning tree. Therefore, as long as there are at least $3 \times n_b$ independent measurements of branch flows or voltage magnitudes, we can achieve identifiability. This is empirically evaluated for several IEEE standard systems in Table I. It is well known that the traditional setting of power flow analysis with $2 \times n_b$ measurements may have many spurious local minima [14], [30], [32]. However, it follows from the proposed linear basis that the network becomes identifiable with a limited but right set of measurements (e.g., M1 and M5 in Table I).*

Remark 6 (Significance of Theorem 4)**.** *Local minima may not be a serious concern in normal situations where the state does not change rapidly and is within a normal operation region. However, the problem becomes crucial during adverse conditions where some states fall out of the normal operation region and need to be identified in real-time. In this case, if the initial point used by a local search algorithm is not close enough to the actual state, then the algorithm could arrive at a point that is a local minimum with no physical meaning.*

Mathematically speaking, SE can be regarded as an overdetermined power flow (PF) in the noiseless case. It is known that PF can have an exponential number of solutions for highly meshed networks [4], [14]. So, the SE in the noiseless case can have any number of solutions from 1 to a very large number. First, it is not known when the number of solutions is unique and that depends on how overdetermined the SE is. Second, although there was not much known about local solutions of optimal power flow (OPF) till 2009, there have been a lot of studies in the past 10 years showing how non-convex the problem is. The computational complexity of SE is an overlooked problem. The complexity of OPF is due to the nonlinearity of PF, and since PF is integral to SE, it could be highly nonconvex. Recent papers show that the problem could have many spurious solutions [30]. SE is a special case of the matrix sensing problem and the machine learning community has shown that the problem could be of high computational complexity with spurious solutions unless strong conditions are satisfied [33], [53], [54]. The focus of the power community has been mostly on how to treat noise and corrupted data in SE rather than dealing with its underlying non-convexity. Since the industry mostly uses DC models and various approximations for SE, they have not directly dealt with the complexity of the problem. Power community has also looked at SE/PF using nonlinear models and their concern has been on the convergence of local search algorithms. Since convergence could be to a wrong solution, the focus of the current work is to first understand when such wrong solutions never exist and second design an algorithm to find the correct solution.

B. Global recovery conditions and error bounds

Despite the simplicity of the identifiability condition, real-world measurements are often subject to random sensor noise and sparse bad data whose support is often unknown. Thus, it is important to examine under what conditions the true state can be recovered (either exactly when the dense noise is negligible, or accurately enough for the case with dense noise).

Theorem 7. *Consider the measurement equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_\natural + \boldsymbol{b}_\natural$, where $\mathrm{supp}(\boldsymbol{b}_\natural) = \mathcal{J}$. Assume that the measurement matrix $\boldsymbol{A}$ satisfies the following conditions: (a) the lower eigenvalue is positive, i.e., $C_{min}(\mathcal{J}) > 0$; (b) the mutual incoherence condition $\rho(\mathcal{J}) < 1$ is satisfied. Then, the unique solution to* (S1-L1)*, denoted as $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$, is exact and recovers the true state (i.e., $\hat{\boldsymbol{x}} = \boldsymbol{x}_\natural$ and $\hat{\boldsymbol{b}} = \boldsymbol{b}_\natural$).*

Theorem 8. *Consider the measurement equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_\natural + \boldsymbol{w}_\natural + \boldsymbol{b}_\natural$, where $\mathrm{supp}(\boldsymbol{b}_\natural) = \mathcal{J}$ and $\boldsymbol{w}_\natural$ is a random vector with zero mean and subgaussian parameter σ. Suppose that the rows of $\boldsymbol{A}$ are normalized, and that the measurement matrix $\boldsymbol{A}$ satisfies the following conditions: (a) the lower eigenvalue is positive, (b) there exists a constant $\gamma > 0$ such that the mutual incoherence condition $\rho(\mathcal{J}) = 1 - \gamma$. Let the regularization parameter λ be chosen such that*

$$\lambda > \frac{2}{n_m \gamma} \sqrt{2\sigma^2 \log n_m}. \tag{9}$$

Then, the following properties hold for the solution to (S1-LASSO)*, denoted as $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$:*

1) *(No false inclusion) The solution $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$ has no false bad data inclusion (i.e., $\mathrm{supp}(\hat{\boldsymbol{b}}) \subset \mathrm{supp}(\boldsymbol{b}_\natural)$) with probability greater than $1 - \frac{c_0}{n_m}$, for some constant $c_0 > 0$.*
2) *(Large bad data detection) Let*

$$g(\lambda) = n_m \lambda \left(\frac{1}{2\sqrt{C_{min}(\mathcal{J})}} + \|\boldsymbol{I}_b (\boldsymbol{Q}_{\mathcal{J}}^\top \boldsymbol{Q}_{\mathcal{J}})^{-1} \boldsymbol{I}_b^\top \|_\infty \right)$$

be a threshold value. Then, all bad data measurements with magnitude greater than $g(\lambda)$ will be detected (i.e., if $|b_{i\natural}| > g(\lambda_m)$, then $|\hat{b}_i| > 0$) with probability greater than $1 - \frac{c_1}{m}$ for some constant $c_1 > 0$.
3) *(Bounded error) The estimator error is bounded by*

$$\|\boldsymbol{x}_\natural - \hat{\boldsymbol{x}}\|_2 \leq \omega \frac{\sqrt{n_m + |\mathcal{J}|}}{C_{min}} + n_m \lambda \|\boldsymbol{I}_x (\boldsymbol{Q}_{\mathcal{J}}^\top \boldsymbol{Q}_{\mathcal{J}})^{-1} \boldsymbol{I}_b^\top \|_{\infty,2}$$

with probability greater than $1 - \exp\left(-\frac{c_1 \omega^2}{\sigma^4}\right)$, where $\|\cdot\|_{\infty,2}$ denotes ℓ_∞–ℓ_2 induced norm.

Despite the difference in measurement assumptions (i.e., existence of dense noise $\boldsymbol{w}$) and estimation algorithms (i.e., (S1-L1) or (S1-LASSO)), it is remarkable that the global recovery conditions in Theorems 7 and 8 are coincident. In the case of negligible dense noise, then a strong global recovery is achieved, that is, both the true state and the bad data are detected. With the presence of dense noise, it is no longer possible to achieve exact recovery; however, Theorem 8 indicates that with a proper selection of the penalty coefficient λ, one can avoid false detection of bad data (part 1), detect bad data with magnitudes greater than a threshold (part 2), and achieve state estimation within bounded error margin. Furthermore, both the bad data threshold and the error

bound decrease with stronger mutual incoherence condition and lower-eigenvalue condition.

The above analysis for (S1-LASSO) can be adapted to the case without dense noise, giving rise to the following corollary.

Corollary 9. *Consider the measurement equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_\natural + \boldsymbol{b}_\natural$, where $\boldsymbol{b}_\natural \in \mathbb{R}^m$ has support $\mathcal{J}$. Suppose that the rows of $\boldsymbol{A}$ are normalized, and the regularization parameter λ is chosen to be positive, i.e., $\lambda > 0$. Assume that $\boldsymbol{A}$ satisfies the following conditions: (a) the lower eigenvalue is positive, (b) the mutual incoherence condition $\rho(\mathcal{J}) < 1$ is satisfied. Then, the following properties hold for the solution to* (S1-LASSO)*, denoted as $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$:*

1) *(No false inclusion) The solution $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$ has no bad data false inclusion (i.e.,* $\mathrm{supp}(\hat{\boldsymbol{b}}) \subset \mathcal{J}$*).*
2) *(Large bad data detection) Let $g(\lambda) = n_m\lambda\|\boldsymbol{I}_b(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{I}_b^\top\|_\infty$ be a threshold value. Then, all bad data measurements with magnitude greater than $g(\lambda)$ will be detected (i.e., if $|b_{i\natural}| > g(\lambda)$, then $|\hat{b}_i| > 0$).*
3) *(Bounded error) The estimator error is bounded by*

$$\|\boldsymbol{x}_\natural - \hat{\boldsymbol{x}}\|_2 \leq n_m\lambda\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{I}_b^\top\|_{\infty,2}.$$

To understand the equivalence between Corollary 9 and Theorem 7, note that one can choose λ to be arbitrary close to 0 so the detection threshold and error bounds also approach 0. The proof of Theorem 8 is based on the primal-dual witness technique popularized by [52]. However, the key difference is that the existing literature in statistical learning only focuses on sparse signal recovery [49]–[52], while the present study needs to recover both the sparse signal (i.e., bad data) and the dense signal (i.e., regressor), which is technically more challenging to prove. Indeed, related works on this topic, such as robust principle component analysis [56] and dense error correction [57], employ different proof techniques than the present study.

In what follows, we will discuss the influence of the possible error in stage–1 estimation on the outcome of the second stage. Let the estimations of x_ℓ^{re} and x_ℓ^{im} over a line $\ell \in \mathcal{L}$ be given by:

$$\hat{x}_\ell^{\mathrm{re}} = x_\ell^{\mathrm{re}} + \Delta x_\ell^{\mathrm{re}} \quad \text{and} \quad \hat{x}_\ell^{\mathrm{im}} = x_\ell^{\mathrm{im}} + \Delta x_\ell^{\mathrm{im}},$$

where x_ℓ^{re} and x_ℓ^{im} are the true values, and $\Delta x_\ell^{\mathrm{re}}$ and $\Delta x_\ell^{\mathrm{im}}$ are the estimation errors from stage 1. We provide a bound on the phase estimation error for each bus $k \in \mathcal{N}$.

Proposition 10. *The estimation error of the phase θ_k is bounded by the k-th component of the vector*

$$\left|(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\mathbf{e}\right|,$$

where $\mathbf{e} \in \mathbb{R}^{n_l}$ is a vector with the elements $e_\ell = \frac{x_\ell^{re}\Delta x_\ell^{im} - x_\ell^{im}\Delta x_\ell^{re}}{x_\ell^{re}\hat{x}_\ell^{re}}$, and $\boldsymbol{L}$ is the matrix described in Sec. III-B.

To understand how the results of this section can guarantee the accurate recovery of the system state, we consider three different cases. **(1)** If there is no dense error or sparse bad data and the identifiability condition in Theorem 4 is satisfied, then it is guaranteed that there is a one-to-one correspondence between the proposed basis and the underlying voltage phasor. **(2)** If there is no dense error but only sparse bad data, under the mutual incoherence condition in Theorems 7 and 8, it is guaranteed that the proposed basis can be uniquely recovered and it corresponds to the underlying voltage phasor. **(3)** In the case with dense error, since it is theoretically impossible to exactly recover the underlying voltage, our goal is to obtain an estimate of the state that is as close to the ground truth as possible. In summary, the estimate is exactly equal to the actual state of the system in Cases 1 and 2, whereas there is a nonzero estimation error in Case 3 that is bounded in Proposition 10. It will be empirically shown that the proposed basis significantly outperforms the polar/rectangular basis by bypassing the non-convexity of the problem and finding a provably correct solution.

Due to the centrality of the mutual incoherence condition throughout the theoretical analysis, we provide an analysis on the likelihood of global recovery condition satisfaction using arguments based on concentration inequalities in probability.

C. Stochastic bound

In this analysis, we assume oblivious adversary, which indicates that the set $\mathcal{J}$ is chosen uniformly at random. The goal is to gauge the likelihood that a random matrix $\boldsymbol{A}$ with an arbitrary sparsity pattern will satisfy the mutual incoherence condition. It is important to capture the network-topology-induced pattern in $\boldsymbol{A}$ in this analysis.

Definition 11 (Sparsity pattern)**.** *For an arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{n_m\times n_x}$, the sparsity pattern is a binary matrix $\boldsymbol{N} \in \mathbb{R}^{n_m\times n_x}$ whose (i,j)-th entry is equal to 0 only if $A_{ij} = 0$. Define the set of matrices with a given sparsity pattern $\boldsymbol{N}$ as*

$$\mathcal{S}(\boldsymbol{N}) \coloneqq \{\boldsymbol{A} \in \mathbb{R}^{n_m\times n_x} | \boldsymbol{A} \circ \boldsymbol{N} = \boldsymbol{A}\},$$

where $\circ$ denotes the Hadamard (element-wise) product.

To conduct the analysis, we fix the sparsity pattern $\boldsymbol{N}$ and assume that $\boldsymbol{A}$ is a sparse matrix with the given pattern, where each entry is a random sub-Gaussian variable. In other words, $\boldsymbol{A} = \boldsymbol{N} \circ \boldsymbol{\Xi}$, where $\boldsymbol{\Xi} = \{\xi_{ij}\}_{i\in[n_m], j\in[n_x]}$ is a dense random matrix with independent and identically distributed sub-Gaussian random variables with variance proxy σ^2 (c.f., [58, Chap. 1] for a detailed account of the terminologies).

In the following, we introduce some metrics to measure the sparsity. For each $j \in [n_x]$, let $n_{\mathcal{J}}^j = \sum_{i\in\mathcal{J}} N_{ij}$ and $n_{\mathcal{J}^c}^j = \sum_{i\in\mathcal{J}^c} N_{ij}$ denote the numbers of nonzero entries in the columns of $\mathcal{N}_{\mathcal{J}}$ and $\mathcal{N}_{\mathcal{J}^c}$, respectively, and let $n_{\mathcal{J}}^* = \max_{j\in[n_x]} n_{\mathcal{J}}^j$ and $n_{\mathcal{J}^c}^* = \max_{j\in[n_x]} n_{\mathcal{J}^c}^j$ be their upper bounds, respectively.

Theorem 12. *Suppose that the following conditions hold:*

1) *(Bounded moments) There exist constants $q > 2$ and $\nu \geq 1$ such that $\mathbb{E}|\xi_{ij}|^q \leq \nu^q$ for every $i \in [n_x]$ and $j \in [n_m]$;*
2) *(Tall matrix) $n_m = (1+\delta)n_x$ for some $\delta > \delta_0$; and*
3) *(Saturated columns) There exists a constant $c_\nu \in (0, \nu)$ such that $n_{\mathcal{J}^c}^j \geq c_\nu^2 \frac{|\mathcal{J}^c|}{\mathbb{E}|\xi_{ij}|^2}$ for all $i \in [n_m]$ and $j \in [n_x]$.*

Then, the minimization (S1-L1) *recovers the true state with probability* $(1-\kappa)$ *as long as the number of corrupted measurements* $|\mathcal{J}|$ *is not too high and satisfies the inequality*

$$\min\left\{c_4|\mathcal{J}^c|, \frac{|\mathcal{J}^c|}{2a_1^2 n_{\mathcal{J}}^* n_x \sigma^2 a_2} - \ln 2n_x\right\} \geq \ln\frac{2}{\kappa}$$

where $a_1 = \frac{c_\nu^4}{32\nu^2}(\frac{c_\nu^2}{64\nu^2})^{\frac{q}{q-2}}$, $a_2 = \min\{\frac{4}{c_\nu^2}a_1, 20\sigma^2\} - \frac{\ln 2}{|\mathcal{J}^c|}$, *and* $\delta_0 = \frac{c_\nu^2}{4a_1}$.

The above theorem states that as long as the number of good measurements is greater than the number of nonzero elements in the spoiled part of the sensing matrix $\boldsymbol{A}$ up to a constant multiplier (i.e., $|\mathcal{J}^c| \gtrsim \text{const} \times n_{\mathcal{J}}^* n_x$), then with high probability, the mutual incoherence condition is satisfied. This agrees with the numerical results that the higher the number of measurements (with a fixed number of bad data), the smaller the mutual incoherence parameter.

D. Discussions

The premise of the proposed approach is that there are abundant data for the grid. However, we do not require that the data be highly reliable. In fact, with the prevalence of cheaply available sensors, we consider the important case that some of the data can be completely wrong due to various reasons such as sensor faults, communication errors or even cyber attacks. The situation of "abundant but untrusted" data is very different from the case of "redundant and reliable" data, because in the latter case, the data are believed to be high-fidelity with very low error rates. The former is also a more challenging scenario, and in particular classical SE based on nonlinear least squares fails due to the adverse influence of bad data. To resolve this issue, our algorithm is based on a novel set of linear basis, where the number of variables is larger than the number of real state variables; however, the minimum number of measurements to guarantee identifiability does not scale by the number of lines in the network, but by the number of buses (Theorem 4). This means that it can be a relatively succinct representation with the right set of measurements (i.e., available measurements can form a spanning tree of the network). Furthermore, our theoretical analysis shows that under some mild conditions, the algorithm recovers the ground truth even though part of the data are completely wrong (Theorems 7 and 8). Generally speaking, when the amount of data is low, a small set of bad data would deteriorate the estimation and even make the system unobservable. In other words, the prerequisite that the amount of measured data is relatively high is mainly for robustness purposes (Theorem 12).

Another advantage of the proposed method is the availability of a convex formulation that can be solved efficiently using second-order algorithms. The normal Gauss-Newton iteration for nonlinear least squares is nonconvex, and therefore it should be initialized at a point close to the ground truth to guarantee convergence. Since we do not know *a priori* the state of the system, there is no guarantee that the estimation will converge to a point close to the ground truth. This is especially true when there is an adverse situation with the state changing rapidly. However, since both stages of the proposed method are convex, the global optimal in each stage can be found under mild conditions (Theorem 7 and Theorem 8).

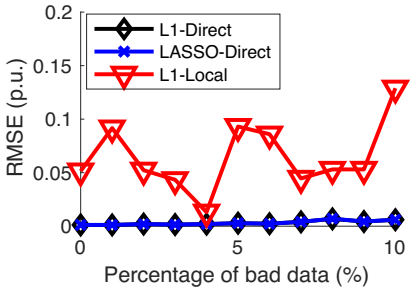
Under the proposed framework, we do not explicitly distinguish leverage and/or vertical outliers [29], [59]. As long as the mutual incoherence condition is satisfied, the proposed method is able to deal with bad leverage points systematically. The mathematical framework of this paper is also different from two recent studies [39] and [60]. The work [39] focuses on an iterative algorithm by locally linearizing the nonlinear measurements; however, the convergence condition for nonlinear measurements needs to be verified using SDP and the analysis is based on the "almost Euclidean property" (similar to RIP), which is difficult to satisfy and the relation to the underlying model of the system is not clear. The work [60] also proposes a linear basis of representation, but it requires that current measurements be known for all lines; it also includes these measurements in the measurement Jacobian as if they are the true value, which can seriously bias the estimation if some of the measurements are bad data, also known as error-in-variables in the statistics literature [61]. By contrast, the sensing marix $\boldsymbol{A}$ in our paper comes from system topology and physical parameters. In the normal situation, this matrix is close to the actual system; however, we also allow that some of the parameters in the matrix to be wrong, which means that the corresponding measurements are treated as bad data.

V. Experiments

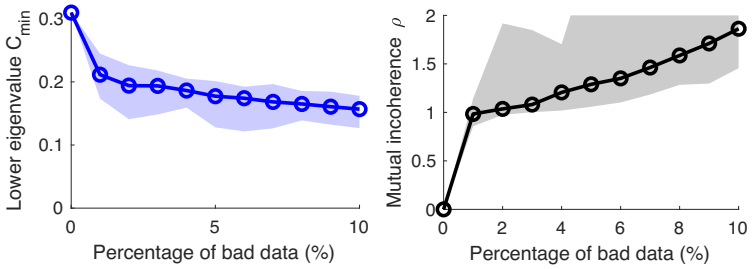
Numerical evaluations are performed on IEEE benchmark systems from MATPOWER [62]. This includes the Pan European Grid Advanced Simulation and State Estimation (PEGASE) 9241-bus and 13659-bus systems, which represent the size and complexity of the European high voltage transmission network [63]. While PMU measurements can be incorporated in the proposed framework, unless otherwise stated, we assume the available measurements to include full nodal measurements (i.e., voltage magnitudes and real/reactive injections) and bi-directional real/reactive branch flows over all lines. All the experiments are performed on a personal laptop with 3.3GHz Intel Core i7 and 16GB memory.

In each case, we randomly generate 50 sets of dense noise $\boldsymbol{w}$ and sparse bad data $\boldsymbol{b}$. The dense noise for each measurement is zero-mean Gaussian variable, with standard deviation of $0.1 \times c_n$ (per unit) for voltage magnitude measurements and c_n (per unit) for all the other measurements, where c_n is the dense noise level. This setup is inspired by the fact that voltage magnitude sensors have higher standards of accuracy compared to power meters. For the sparse bad data, unless otherwise specified, its support $\mathcal{J}$ is randomly selected among the line measurements, with the only assumption that at most 1 bad data measurement exists for each line. The values for the sparse noise can be arbitrarily large, and we assume that these parameters are uniformly chosen from the set $[-4.25, -3.75] \cup [3.75, 4.25]$ (per unit).

We adopt the root-mean-square error (RMSE) as the performance metric, which is defined as $\sqrt{\frac{1}{n_b}\sum_{i\in\mathcal{N}}|v_i - \hat{v}_i|^2}$, where v_i and $\hat{v}_i$ are the true and estimated complex voltage



(a) RMSE



(b) lower eigenvalue $C_{\min}$ (left) and mutual incoherence ρ (right)

Fig. 3: Evaluation of the (S1-L1)–cleaning–direct recovery (L1-Direct), (S1-LASSO)–cleaning–direct recovery (LASSO-Direct), and local search with ℓ_1 loss (L1-Local) and squared loss with cleaning step (since the RMSE is greater than 0.2, its line is not shown on the graph) for the IEEE 300-bus system. We vary the percentage of bad data measurements from 0% to 10% (out of all line measurements), with the dense noise level fixed at c_n =0.5%. The plots in (b) indicate the median (line with circles) and the min/max value (shaded region).

TABLE II: Comparison of the (S1-L1)–cleaning–direct recovery (L1-Direct), (S1-LASSO)–cleaning–direct recovery (LASSO-Direct), and local search with ℓ_1 loss and Newton's method with bad data detection. We fix the percentage of bad data at 5% (out of all line measurements) and dense noise level at c_n =0.5%.

	Newton method			Local search ℓ_1			LASSO-Direct			L1-Direct		
	RMSE	F1	Time (s)	RMSE	F1	Time (s)	RMSE	F1	Time (s)	RMSE	F1	Time (s)
14 Bus	.002	.852	0.6	.001	1	0.3	.001	1	2.3	.001	1	2.2
30 Bus	.042	.808	2.4	.001	.996	0.4	.002	1	2.3	.002	1	2.2
57 Bus	.043	.827	3.2	.001	.998	1.2	.004	.999	2.3	.004	.999	2.1
118 Bus	.003	.848	7.4	.002	.980	4.1	.002	1	1.5	.002	1	1.3
300 Bus	.699	.379	58.1	.093	.858	21.6	.004	.999	2.6	.004	.999	1.2

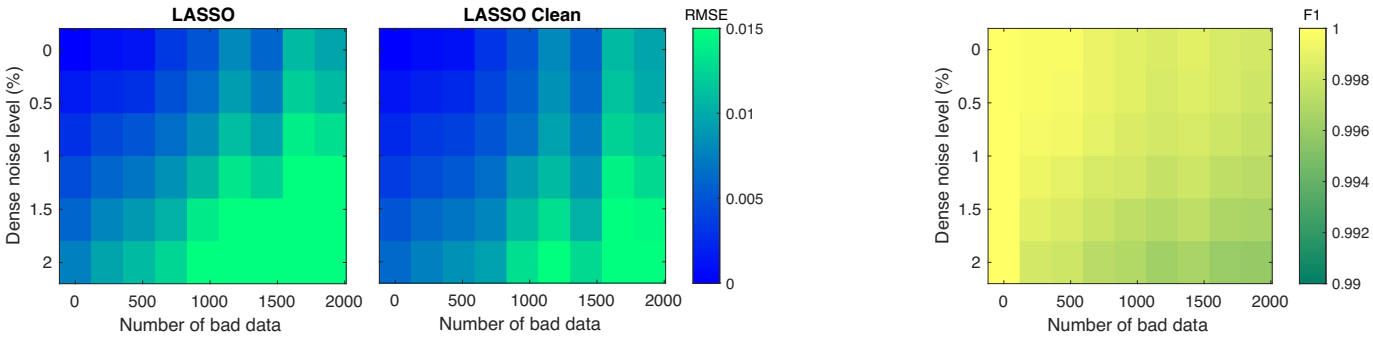
at bus $i \in \mathcal{N}$. To evaluate the bad data detection accuracy, we use the F1 score, which is defined as $\frac{2*\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$, where *precision* is given by $\frac{\text{\#True positives } |\mathcal{J}\cap\hat{\mathcal{J}}|}{\text{\#Conditional positives } |\hat{\mathcal{J}}|}$, and *recall* is given by $\frac{\text{\#True positives } |\mathcal{J}\cap\hat{\mathcal{J}}|}{\text{\#Conditional positives } |\mathcal{J}|}$, and $\mathcal{J}$ and $\hat{\mathcal{J}}$ denote the true and estimated support of bad data (# shows the number of elements). The F1 score is the harmonic average of the precision and recall, which reaches its best value at 1 (perfect precision and recall) and worst at 0.

We compare the proposed method (stage-1 estimators (S1-L1) or (S1-LASSO) combined with stage-2 direct recovery method) with the current practice local search method using the squared loss Newton method, and another local search method that replaces the squared loss with ℓ_1 loss [30]. We use SeDuMi [64] as the linear programming solver, and the MATLAB implementation of limited-memory BFGS [65] for the local search methods, similar to [30]. Throughout the experiment, we choose λ in (S1-LASSO) to be $3\times10^{-4}/n_m$, which we found to be consistently well-behaving. In addition, we choose a threshold of 0.1 for stage-1 estimators and 0.3 for local search methods, which seem to work best for all methods to detect bad data. After the removal of bad data (i.e., cleaning step), we can optionally perform the estimation with the remaining data for both the proposed stage-1 estimators and the Newton method.

First, we evaluate the robustness of the methods to bad data. As is shown in Fig. 3, due to convergence issues and spurious local minima, none of the local search methods could correctly estimate the true state. On the other hand, with the increase of bad data percentage, the proposed methods can reliably recover the ground truth, even with 10% of arbitrarily bad data. This can be implied from the lower eigenvalue conditions and the mutual incoherence conditions, which remain well-conditioned with the presence of bad data. We also perform the experiments on other systems, as is shown in Table II with bad data fixed at 5% level and dense noise fixed at c_n =0.5%. It can be observed that local search methods (with a cleaning step for Newton's method) perform relatively well when the scale is small (up to 118 buses), but the performance (e.g., RMSE and bad data detection F1 score) deteriorates significantly for larger systems due to the existence of spurious local minima. In addition, the proposed methods remain superior, due to the efficient detection of bad data (with F1 score close to 1).

Next, we examine the performance of the proposed estimators when both the dense noise and the bad data intensity vary. We test on the French very high voltage and high voltage transmission network with 2848 buses [63]. As is shown in Fig. 4, the algorithm achieves a low RMSE with up to 1000 bad data measurements and 1% level of dense noise. The detection score for bad data remains above 99% for all the scenarios. We also show that due to the high detection accuracy of the bad data, it is beneficial to redo the estimation after the cleaning stage (LASSO Clean), which can improve the RMSE of estmation espeically when the number of bad data measurements is significant.

Last but not least, we demonstrate the scalability of the method on large systems with up to 13659 buses, which is the largest system provided by MATPOWER and challenging



(a) RMSE for LASSO-Direct with (right) and without (left) cleaning step

(b) F1 score for (S1-LASSO)

Fig. 4: Evaluation of the (S1-LASSO)–direct recovery method on the PEGASE 2848-bus system. The dense noise level c_n varies from 0 to 2%, and the number of bad data measurements ranges up to 2000 (roughly 9% of the total line measurements). The bad data detection accuracy is shown as the F1 score. After the detection of bad data, they are removed and the remaining clean data are used again in the estimation (LASSO Clean).

for the existing algorithms (Table III). We evaluate the performance of the algorithm with data redundancy lower than the full set of measurements. In particular, we first examine Case A where the measurements include voltage magnitudes at all buses $\{|v_k|\}_{k\in\mathcal{N}}$ and three branch flows over all lines $\{p_{ij}^{(\ell)}, p_{ji}^{(\ell)}, q_{ij}^{(\ell)}\}_{\ell\in\mathcal{L}}$ and then study Case B as a sparse scenario that includes voltage magnitudes at all buses $\{|v_k|\}_{k\in\mathcal{N}}$ but three branch flows on a spanning tree and $0.2\times n_b$ additional lines. The number of measurements scales by the number of buses in Case B, and is comparable to the classical setting of the power flow calculation. In this experiment, we relax the previous assumption of having only one bad measurement per line and instead allow all measurements associated with the line to be corrupted. This model allows accounting for correlated bad data, which are the most challenging ones to identify. The lines with compromised data are chosen randomly, but they are checked to ensure that the network does not become disconnected after removing those lines. We fix the dense noise level at 0.5% and the percentage of bad data at 1% of the full measurements (i.e., $0.01\times(3\times n_b+4\times n_l)$ number of bad data) for all the cases. The number of bad data measurements ranges from 120 (for PEGASE 1354-bus) to 1228 (for PEGASE 13659-bus).

It can be observed that the performance is satisfactory in all of the scenarios, and the estimation becomes more robust with a higher number of measurements (i.e., lower RMSE and higher bad data detection accuracy). Moreover, all computations are performed within a minute, which is important for real-time situational awareness.

TABLE III: Evaluation in large-scale benchmarks. Case A includes voltage magnitude measurements on all buses and 3 branch flow measurements; Case B includes voltage measurement on all buses and 3 branch flow measurements on $1.2\times n_b$ lines that consist of a spanning tree of the network. We use the (S1-LASSO)-Cleaning-Direct recovery method in both cases.

	Case A		Case B		Time
	RMSE	F1	RMSE	F1	(sec)
1354 Bus	.003	.996	.003	.995	9.8
2848 Bus	.004	.995	.003	.996	13.6
3012 Bus	.003	.998	.001	.998	18.5
6495 Bus	.005	.994	.005	.996	36.2
9241 Bus	.007	.993	.009	.994	43.6
13659 Bus	.007	.994	.009	.995	52.4

VI. Conclusion

In this study, we proposed a linear basis of representation for power system measurements that succinctly captures the topology of the network. This leads to a two-stage estimation approach that efficiently solve the nonconvex SE under mild conditions usually satisfied with a sufficient instrumentation of sensors. The proposed algorithm is provably robust to bad data. We developed a robustness metric based on a deterministic quantity called mutual incoherence. Theoretical analysis of the global recovery condition and statistical error bounds was conducted, which relied on this key metric. The algorithm demonstrated robustness to bad data in various empirical evaluations, and achieved superior performance compared to baselines. Above all, the proposed method exhibited a satisfactory scalability for large systems with more than 13,000 buses. In contrast to semidefinite programming relaxation approaches, the SE can be solved with high accuracy within a minute for such large systems. This can significantly improve real-time situational awareness of grid operation.

References

[1] A. Monticelli, "Electric power system state estimation," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 262–282, 2000.

[2] A. Gomez-Exposito and A. Abur, *Power system state estimation: theory and implementation*. CRC press, 2004.

[3] National Academies of Sciences, Engineering, and Medicine, *Enhancing the Resilience of the Nation's Electricity System*. National Academies Press, 2017.

[4] D. Bienstock and A. Verma, "Strong NP-hardness of AC power flows feasibility," *arXiv preprint arXiv:1512.07315*, 2015.

[5] K. Lehmann, A. Grastien, and P. Van Hentenryck, "AC-feasibility on tree networks is np-hard," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 798–801, 2016.

[6] W. F. Tinney and C. E. Hart, "Power flow solution by Newton's method," *IEEE Transactions on Power Apparatus and Systems*, no. 11, pp. 1449–1460, 1967.

[7] F. C. Schweppe and J. Wildes, "Power system static-state estimation, Part I: Exact model," *IEEE Transactions on Power Apparatus and Systems*, no. 1, pp. 120–125, 1970.

[8] W. W. Kotiuga and M. Vidyasagar, "Bad data rejection properties of weighted least absolute value techniques applied to static state estimation," *IEEE Transactions on Power Apparatus and Systems*, no. 4, pp. 844–853, 1982.

[9] M. K. Celik and A. Abur, "A robust WLAV state estimator using transformations," *IEEE Transactions on Power Systems*, vol. 7, no. 1, pp. 106–113, 1992.

[10] K. Y. Lee and M. A. El-Sharkawi, *Modern heuristic optimization techniques: theory and applications to power systems*. John Wiley & Sons, 2008, vol. 39.

[11] Y. Weng, Q. Li, R. Negi, and M. Ilić, "Semidefinite programming for power system state estimation," in *Proc. of IEEE Power and Energy Society General Meeting*, 2012, pp. 1–8.

[12] V. Kekatos and G. B. Giannakis, "Distributed robust power system state estimation," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1617–1626, 2013.

[13] R. Madani, J. Lavaei, and R. Baldick, "Convexification of power flow equations for power systems in presence of noisy measurements," *to appear in IEEE Transactions on Automatic Control*, 2018.

[14] D. K. Molzahn, D. Mehta, and M. Niemerg, "Toward topologically based upper bounds on the number of power flow solutions," in *Proc. of IEEE American Control Conference*, 2016, pp. 5927–5932.

[15] V. Kekatos, G. Wang, H. Zhu, and G. B. Giannakis, "PSSE redux: Convex relaxation, decentralized, robust, and dynamic approaches," *arXiv preprint arXiv:1708.03981*, 2017.

[16] Y. Zhang, R. Madani, and J. Lavaei, "Conic relaxations for power system state estimation with line measurements," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1193–1205, 2018.

[17] D. Mehta, D. K. Molzahn, and K. Turitsyn, "Recent advances in computational methods for the power flow equations," in *Proc. of IEEE American Control Conference*, 2016, pp. 1753–1765.

[18] K. A. Clements and A. S. Costa, "Topology error identification using normalized lagrange multipliers," *IEEE Transactions on Power Systems*, vol. 13, no. 2, pp. 347–353, 1998.

[19] M. Korkali and A. Abur, "Robust fault location using least-absolute-value estimator," *IEEE Transactions on Power Systems*, vol. 4, no. 28, pp. 4384–4392, 2013.

[20] Y. Weng, M. D. Ilić, Q. Li, and R. Negi, "Convexification of bad data and topology error detection and identification problems in AC electric power systems," *IET Generation, Transmission & Distribution*, vol. 9, no. 16, pp. 2760–2767.

[21] Y. Lin and A. Abur, "Robust state estimation against measurement and network parameter errors," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 4751–4759, 2018.

[22] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security*, vol. 14, no. 1, p. 13, 2011.

[23] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, 2011.

[24] M. Jin, J. Lavaei, and K. H. Johansson, "Power grid AC-based state estimation: Vulnerability analysis against cyber attacks," *to appear in IEEE Transactions on Automatic Control*, 2018.

[25] K. Clements and P. Davis, "Detection and identification of topology errors in electric power systems," *IEEE Transactions on Power Systems*, vol. 3, no. 4, pp. 1748–1753, 1988.

[26] F. F. Wu and W.-H. Liu, "Detection of topology errors by state estimation (power systems)," *IEEE Transactions on Power Systems*, vol. 4, no. 1, pp. 176–183, 1989.

[27] M. Irving, R. Owen, and M. Sterling, "Power-system state estimation using linear programming," in *Proc. of the Institution of Electrical Engineers*, vol. 125, no. 9. IET, 1978, pp. 879–885.

[28] A. Abur and M. K. Celik, "A fast algorithm for the weighted least absolute value state estimation (for power systems)," *IEEE Transactions on Power Systems*, vol. 6, no. 1, pp. 1–8, 1991.

[29] L. Mili, M. G. Cheniae, and P. J. Rousseeuw, "Robust state estimation of electric power systems," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 41, no. 5, pp. 349–358, 1994.

[30] R. Mohammadi-Ghazi and J. Lavaei, "Empirical analysis of ℓ_1-norm for state estimation in power systems," 2018. [Online]. Available: https://lavaei.ieor.berkeley.edu/SE_norm-1-2018.pdf

[31] K. Dvijotham, M. Chertkov, and S. Low, "A differential analysis of the power flow equations," in *Proc. of IEEE Annual Conference on Decision and Control*, 2015, pp. 23–30.

[32] R. Zhang, J. Lavaei, and R. Baldick, "Spurious critical points in power system state estimation," in *Proc. of the 51st Hawaii International Conference on System Sciences*, 2018.

[33] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[34] S. Naka, T. Genji, T. Yura, and Y. Fukuyama, "A hybrid particle swarm optimization for distribution state estimation," *IEEE Transactions on Power Systems*, vol. 18, no. 1, pp. 60–68, 2003.

[35] A. Trias, "The holomorphic embedding load flow method," in *Proc. of IEEE Power and Energy Society General Meeting*, 2012, pp. 1–8.

[36] W. Ma and J. S. Thorp, "An efficient algorithm to locate all the load flow solutions," *IEEE Transactions on Power Systems*, vol. 8, no. 3, pp. 1077–1083, 1993.

[37] G. Wang, A. S. Zamzam, G. B. Giannakis, and N. D. Sidiropoulos, "Power system state estimation via feasible point pursuit: Algorithms and cramér-rao bound," *IEEE Transactions on Signal Processing*, vol. 66, no. 6, pp. 1649–1658, 2018.

[38] G. Wang, G. B. Giannakis, and J. Chen, "Robust and scalable power system state estimation via composite optimization," *IEEE Transactions on Smart Grid*, 2019.

[39] W. Xu, M. Wang, J. Cai, and A. Tang, "Sparse recovery from nonlinear measurements with applications in bad data detection for power networks," *arXiv preprint arXiv:1112.6234*, 2011.

[40] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal of the ACM*, vol. 42, no. 6, pp. 1115–1145, 1995.

[41] Y. Nesterov, "Semidefinite relaxation and nonconvex quadratic optimization," *Optimization methods and software*, vol. 9, no. 1-3, pp. 141–160, 1998.

[42] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[43] B. Zhang and D. Tse, "Geometry of feasible injection region of power networks," in *Proc. of IEEE Annual Allerton Conference on Communication, Control, and Computing*, 2011, pp. 1508–1515.

[44] J. Lavaei and S. H. Low, "Zero duality gap in optimal power flow problem," *IEEE Transactions on Power Systems*, vol. 27, no. 1, p. 92, 2012.

[45] S. Sojoudi and J. Lavaei, "Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 1746–1778, 2014.

[46] J. Lavaei and S. H. Low, "Convexification of optimal power flow problem," in *IEEE Annual Allerton Conference on Communication, Control, and Computing*, 2010, pp. 223–232.

[47] H. Zhu and G. B. Giannakis, "Power system nonlinear state estimation using distributed semidefinite programming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 6, pp. 1039–1050, 2014.

[48] D. Bienstock, *Electrical transmission system cascades and vulnerability: an operations research viewpoint*. SIAM, 2015, vol. 22.

[49] J.-J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3601–3608, 2005.

[50] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[51] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, no. Nov, pp. 2541–2563, 2006.

[52] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

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

[54] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[55] F. F. Wu and A. Monticelli, "Network observability: theory," *IEEE Transactions on Power Apparatus and Systems*, no. 5, pp. 1042–1048, 1985.

[56] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, p. 11, 2011.

[57] J. Wright and Y. Ma, "Dense error correction via ℓ_1-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3540–3560, 2010.

[58] P. Rigollet, "High dimensional statistics," *Lecture Notes, Cambridge, MA, USA*, 2017.

[59] J. Zhao and L. Mili, "Vulnerability of the largest normalized residual statistical test to leverage points," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4643–4646, 2018.

[60] A. S. Dobakhshari, S. Azizi, M. Paolone, and V. Terzija, "Ultra fast linear state estimation utilizing scada measurements," *IEEE Transactions on Power Systems*, 2019.

[61] M. Rudelson, S. Zhou *et al.*, "Errors-in-variables models with dependent measurements," *Electronic Journal of Statistics*, vol. 11, no. 1, pp. 1699–1797, 2017.

[62] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, 2011.

[63] C. Josz, S. Fliscounakis, J. Maeght, and P. Panciatici, "AC power flow data in MATPOWER and QCQP format: iTesla, RTE snapshots, and PEGASE," *arXiv preprint arXiv:1603.01533*, 2016.

[64] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization methods and software*, vol. 11, no. 1-4, pp. 625–653, 1999.

[65] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[66] A. Litvak and O. Rivasplata, "Smallest singular value of sparse random matrices," *arXiv preprint arXiv:1106.0938*, 2011.

[67] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, "Smallest singular value of random matrices and geometry of random polytopes," *Advances in Mathematics*, vol. 195, no. 2, pp. 491–523, 2005.

[68] M. Rudelson, R. Vershynin *et al.*, "Hanson-Wright inequality and sub-gaussian concentration," *Electronic Communications in Probability*, vol. 18, 2013.

APPENDIX

A. Proof of Theorem 4

Proof. Since the only element in the null space of $\boldsymbol{A}_{\text{span}}$ is $\boldsymbol{0}$, the measurement equation (4) has a unique solution, which corresponds to the true state $\boldsymbol{x}_\natural$. By the algorithm in stage 2, outlined in Sec. III-B, this recovers the true state of the grid. □

B. Proof of Theorem 7

Proof. The dual program of (S1-L1) is given by:

$$\max_{\boldsymbol{h}\in\mathbb{R}^{n_m}} \boldsymbol{h}^\top \boldsymbol{y}, \quad \text{subject to} \quad \boldsymbol{h}^\top \boldsymbol{A} = \boldsymbol{0}, \|\boldsymbol{h}\|_\infty \le 1. \tag{L1-Dual}$$

To show that $(\boldsymbol{x}_\natural, \boldsymbol{b}_\natural)$ is the optimal solution of (S1-L1), we simply need to find a dual certificate $\boldsymbol{h}_\star$ that satisfies the Karush-Kuhn-Tucker (KKT) conditions:

$$\text{(dual feasibility)} \quad \boldsymbol{h}_\star^\top \boldsymbol{A} = \boldsymbol{0}, \tag{10}$$

$$\text{(stationarity)} \quad \boldsymbol{h}_\star \in \partial \|\boldsymbol{b}_\natural\|_1, \tag{11}$$

where $\partial\|\boldsymbol{b}_\natural\|_1$ denotes the subgradiet of $\|\boldsymbol{b}_\natural\|_1$. By the definition of $\mathcal{J} := \operatorname{supp}(\boldsymbol{b}_\natural)$, we need to find an $\boldsymbol{h}_\star$ such that $\boldsymbol{h}_{\star\mathcal{J}} = \operatorname{sign}(\boldsymbol{b}_{\natural\mathcal{J}})$ and $\|\boldsymbol{h}_{\star\mathcal{J}^c}\|_\infty \le 1$. In fact, we can meet a slightly stronger condition for strict feasibility by choosing $\boldsymbol{h}_{\star\mathcal{J}^c} = -\boldsymbol{A}_{\mathcal{J}^c}^{\top+}\boldsymbol{A}_{\mathcal{J}}^\top \operatorname{sign}(\boldsymbol{b}_{\natural\mathcal{J}})$, which satisfies strict dual feasibility (i.e., $\|\boldsymbol{h}_{\star\mathcal{J}^c}\|_\infty < 1$) due to the mutual incoherence condition. Thus, this certifies the optimality of $(\boldsymbol{x}_\natural, \boldsymbol{b}_\natural)$ for (S1-L1).

To show that $(\boldsymbol{x}_\natural, \boldsymbol{b}_\natural)$ is the unique optimal solution, let $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}})$ be an arbitrary feasible point of (S1-L1) different from $(\boldsymbol{x}_\natural, \boldsymbol{b}_\natural)$. Due to the lower eigenvalue condition, the matrix $\boldsymbol{Q}_{\mathcal{J}} := \begin{bmatrix}\boldsymbol{A} & \boldsymbol{I}_{\mathcal{J}}^\top\end{bmatrix}$ has full column rank. Let $\tilde{\mathcal{J}} = \operatorname{supp}(\tilde{\boldsymbol{b}})$, then $\tilde{\mathcal{J}}$ must not be equal to or be a subset of $\mathcal{J}$, because otherwise, from $\boldsymbol{Q}_{\mathcal{J}}\begin{bmatrix}\boldsymbol{x}_\natural \\ \boldsymbol{b}_\natural\end{bmatrix} = \boldsymbol{Q}_{\mathcal{J}}\begin{bmatrix}\tilde{\boldsymbol{x}} \\ \tilde{\boldsymbol{b}}\end{bmatrix} = \boldsymbol{y}$, we must have $\begin{bmatrix}\boldsymbol{x}_\natural \\ \boldsymbol{b}_\natural\end{bmatrix} = \begin{bmatrix}\tilde{\boldsymbol{x}} \\ \tilde{\boldsymbol{b}}\end{bmatrix}$, which is contradictory to the assumption. Let $\tilde{\mathcal{J}}_c = \tilde{\mathcal{J}} \setminus \mathcal{J}$, then,

$$\begin{aligned}
\|\boldsymbol{b}_\natural\|_1 &= \boldsymbol{h}_\star^\top \boldsymbol{y} && (12)\\
&= \boldsymbol{h}_\star^\top (\boldsymbol{A}\tilde{\boldsymbol{x}} + \boldsymbol{I}_{\tilde{\mathcal{J}}_c}^\top \tilde{\boldsymbol{b}}_{\tilde{\mathcal{J}}_c} + \boldsymbol{I}_{\mathcal{J}}^\top \tilde{\boldsymbol{b}}_{\mathcal{J}}) && (13)\\
&= \boldsymbol{h}_{\star\tilde{\mathcal{J}}_c}^\top \tilde{\boldsymbol{b}}_{\tilde{\mathcal{J}}_c} + \boldsymbol{h}_{\star\mathcal{J}}^\top \tilde{\boldsymbol{b}}_{\mathcal{J}} && (14)\\
&\le \|\boldsymbol{h}_{\star\tilde{\mathcal{J}}_c}\|_\infty \|\tilde{\boldsymbol{b}}_{\tilde{\mathcal{J}}_c}\|_1 + \|\boldsymbol{h}_{\star\mathcal{J}}\|_\infty \|\tilde{\boldsymbol{b}}_{\mathcal{J}}\|_1 && (15)\\
&< \|\tilde{\boldsymbol{b}}_{\tilde{\mathcal{J}}_c}\|_1 + \|\tilde{\boldsymbol{b}}_{\mathcal{J}}\|_1 && (16)\\
&= \|\tilde{\boldsymbol{b}}\|_1, && (17)
\end{aligned}$$

where (12) is due to the strong duality between (S1-L1) and (L1-Dual), (13) is due to the primal feasibility of $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}})$, (14) is due to the dual feasibility condition (10), (15) is due to the Hölder inequality, and (16) is due to the strict feasibility of $\boldsymbol{h}_\star$. Thus, we have shown the uniqueness of the optimal solution $(\boldsymbol{x}_\natural, \boldsymbol{b}_\natural)$.

□

C. Proof of Theorem 8

We design the primal-dual witness (PDW) process as follows (note that this is not an actual algorithm, because we do not know the true support $\mathcal{J}$; rather, it is only part of a proof technique popularized by [52]):

1) Set $\hat{\boldsymbol{b}}_{\mathcal{J}^c} = \boldsymbol{0}$
2) Determine $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}_{\mathcal{J}})$ by solving the following program:

$$\min_{\boldsymbol{b}\in\mathbb{R}^{n_m}, \boldsymbol{x}\in\mathbb{R}^{n_x}} \frac{1}{2n_m}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{I}_{\mathcal{J}}^\top \boldsymbol{b}_{\mathcal{J}}\right\|_2^2 + \lambda\|\boldsymbol{b}_{\mathcal{J}}\|_1, \tag{18}$$

and $\hat{\boldsymbol{z}}_{\mathcal{J}} \in \partial\|\hat{\boldsymbol{b}}_{\mathcal{J}}\|_1$ satisfying

$$-\frac{1}{n_m}\boldsymbol{I}_{\mathcal{J}}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{I}_{\mathcal{J}}^\top \hat{\boldsymbol{b}}_{\mathcal{J}}) + \lambda\hat{\boldsymbol{z}}_{\mathcal{J}} = \boldsymbol{0}, \tag{19}$$

$$\boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{I}_{\mathcal{J}}^\top \hat{\boldsymbol{b}}_{\mathcal{J}}) = \boldsymbol{0}. \tag{20}$$

3) Solve $\hat{\boldsymbol{z}}_{\mathcal{J}^c}$ via the zero-subgradient equation:

$$-\frac{1}{n_m}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}} - \hat{\boldsymbol{b}}) + \lambda\hat{\boldsymbol{z}} = \boldsymbol{0} \tag{21}$$

and check whether strict feasibility condition $\|\hat{\boldsymbol{z}}_{\mathcal{J}^c}\|_\infty < 1$ holds.

Lemma 13. *If the PDW procedure succeeds, then $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$ where $\hat{\boldsymbol{b}} = (\hat{\boldsymbol{b}}_{\mathcal{J}}, 0)$ is unique optimal of (S1-LASSO).*

Proof. If PDW succeeds, then the optimality conditions (20) and (21) are satisfied, which certify the optimality of $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}})$. The subgradient $\hat{\boldsymbol{z}}$ satisfies $\|\hat{\boldsymbol{z}}_{\mathcal{J}^c}\|_\infty < 1$ and $\left\langle \hat{\boldsymbol{z}}, \hat{\boldsymbol{b}} \right\rangle = \|\hat{\boldsymbol{b}}\|_1$.

Now, let $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}})$ be any other optimal, and let $F(\boldsymbol{x}, \boldsymbol{b}) = \frac{1}{2n_m}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$, then we have

$$F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}) + \lambda \left\langle \hat{\boldsymbol{z}}, \hat{\boldsymbol{b}} \right\rangle = F(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}}) + \lambda \|\tilde{\boldsymbol{b}}\|_1,$$

and hence we have

$$F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}) + \lambda \left\langle \hat{\boldsymbol{z}}, \hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}} \right\rangle = F(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}}) + \lambda \left(\|\tilde{\boldsymbol{b}}\|_1 - \left\langle \hat{\boldsymbol{z}}, \tilde{\boldsymbol{b}} \right\rangle \right).$$

By the optimality conditions (20) and (21), we have $\lambda \hat{\boldsymbol{z}} = -\nabla_b F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}) = \frac{1}{n_m}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}} - \hat{\boldsymbol{b}})$ and $\nabla_x F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}) = \boldsymbol{0}$, which imply that

$$\begin{aligned} F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}) - \left\langle \nabla_b F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}), \hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}} \right\rangle - F(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}}) \\ = \lambda \left(\|\tilde{\boldsymbol{b}}\|_1 - \left\langle \hat{\boldsymbol{z}}, \tilde{\boldsymbol{b}} \right\rangle \right) \leq 0 \end{aligned}$$

due to convexity. We thus have $\|\tilde{\boldsymbol{b}}\|_1 \leq \left\langle \hat{\boldsymbol{z}}, \tilde{\boldsymbol{b}} \right\rangle$. Since by Holder's inequality, we also have $\left\langle \hat{\boldsymbol{z}}, \tilde{\boldsymbol{b}} \right\rangle \leq \|\hat{\boldsymbol{z}}\|_\infty \|\tilde{\boldsymbol{b}}\|_1$ and $\|\hat{\boldsymbol{z}}\|_\infty \leq 1$, we must have $\|\tilde{\boldsymbol{b}}\|_1 = \left\langle \hat{\boldsymbol{z}}, \tilde{\boldsymbol{b}} \right\rangle$, and $\tilde{\boldsymbol{b}}_j = 0$ for $j \in \mathcal{J}^c$. This means that we have $\mathrm{supp}(\tilde{\boldsymbol{b}}) \subseteq \mathrm{supp}(\hat{\boldsymbol{b}}) \subseteq \mathcal{J}$. By restricting the optimization of $\boldsymbol{b}$ in (S1-LASSO) to the support $\mathcal{J}$ and by the lower eigenvalue condition, the optimization is strictly convex and the uniqueness of the solution follows. □

Lemma 14. *Suppose that $\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J}$ is invertible for $\mathcal{J} \subset [m]$, where $\boldsymbol{Q}_\mathcal{J} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{I}_\mathcal{J}^\top \end{bmatrix}$. Then, we have*

$$\rho(\mathcal{J}) = \|\boldsymbol{A}_{\mathcal{J}^c} \boldsymbol{I}_x (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{I}_b^\top\|_\infty. \tag{22}$$

Proof. We will show that for any given $\mathcal{J} \subset [m]$, we have $\boldsymbol{A}_{\mathcal{J}^c} \boldsymbol{I}_x (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{I}_b^\top = -\boldsymbol{A}_{\mathcal{J}^c}^{\top +} \boldsymbol{A}_\mathcal{J}^\top$. By the definition of $\boldsymbol{Q}_\mathcal{J}$ and block matrix inversion formula, we have

$$\begin{aligned} &\boldsymbol{I}_x (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{I}_b^\top \\ &= -(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}_\mathcal{J}^\top (\boldsymbol{I} - \boldsymbol{A}_\mathcal{J} (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}_\mathcal{J}^\top)^{-1} \\ &= -(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}_\mathcal{J}^\top (\boldsymbol{I} + \boldsymbol{A}_\mathcal{J} (\boldsymbol{A}_{\mathcal{J}^c}^\top \boldsymbol{A}_{\mathcal{J}^c})^{-1} \boldsymbol{A}_\mathcal{J}^\top) \\ &= -(\boldsymbol{A}^\top \boldsymbol{A})^{-1} (\boldsymbol{I} + \boldsymbol{A}_\mathcal{J}^\top \boldsymbol{A}_\mathcal{J} (\boldsymbol{A}_{\mathcal{J}^c}^\top \boldsymbol{A}_{\mathcal{J}^c})^{-1}) \boldsymbol{A}_\mathcal{J}^\top \\ &= -(\boldsymbol{A}_{\mathcal{J}^c}^\top \boldsymbol{A}_{\mathcal{J}^c})^{-1} \boldsymbol{A}_\mathcal{J}^\top, \end{aligned}$$

where the first equation follows from the Sherman–Morrison–Woodbury formula (c.f., Prop. 18 in Sec. G) and the rest are elementary operations. □

Lemma 15. *Suppose that $\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J}$ is invertible for a given $\mathcal{J} \subset [m]$, where $\boldsymbol{Q}_\mathcal{J} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{I}_\mathcal{J}^\top \end{bmatrix}$. Then, we have*

$$\boldsymbol{I}_b (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{I}_b^\top = \boldsymbol{I} + \boldsymbol{A}_\mathcal{J} (\boldsymbol{A}_{\mathcal{J}^c}^\top \boldsymbol{A}_{\mathcal{J}^c})^{-1} \boldsymbol{A}_\mathcal{J}^\top \tag{23}$$

Proof. By the definition of $\boldsymbol{Q}_\mathcal{J}$ and block matrix inversion formula, we have

$$\begin{aligned} \boldsymbol{I}_b (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{I}_b^\top &= (\boldsymbol{I} - \boldsymbol{A}_\mathcal{J} (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}_\mathcal{J}^\top)^{-1} \\ &= \boldsymbol{I} + \boldsymbol{A}_\mathcal{J} (\boldsymbol{A}_{\mathcal{J}^c}^\top \boldsymbol{A}_{\mathcal{J}^c})^{-1} \boldsymbol{A}_\mathcal{J}^\top, \end{aligned}$$

where the second equation follows from the Sherman–Morrison–Woodbury formula. □

Proof of Theorem 8

We prove each part sequentially:

Part 1): By the construction of PDW, we have $\hat{\boldsymbol{b}}_{\mathcal{J}^c} = \boldsymbol{b}_{\natural \mathcal{J}^c} = \boldsymbol{0}$. The zero-subgradient condition (21) can be written as:

$$\begin{aligned} -\frac{1}{n_m} \left(\begin{bmatrix} \boldsymbol{I}_\mathcal{J} \boldsymbol{A} \\ \boldsymbol{I}_{\mathcal{J}^c} \boldsymbol{A} \end{bmatrix} (\boldsymbol{x}_\natural - \hat{\boldsymbol{x}}) + \begin{bmatrix} \boldsymbol{I}_\mathcal{J} \\ \boldsymbol{0} \end{bmatrix} (\boldsymbol{b}_\natural - \hat{\boldsymbol{b}}) \right) \\ - \frac{1}{n_m} \begin{bmatrix} \boldsymbol{I}_\mathcal{J} \\ \boldsymbol{I}_{\mathcal{J}^c} \end{bmatrix} \boldsymbol{w}_\natural + \lambda \begin{bmatrix} \hat{\boldsymbol{z}}_\mathcal{J} \\ \hat{\boldsymbol{z}}_{\mathcal{J}^c} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \end{aligned}$$

where the equations indexed by $\mathcal{J}$ can be re-written as:

$$\begin{aligned} -\frac{1}{n_m} \begin{bmatrix} \boldsymbol{I}_\mathcal{J} \boldsymbol{A} & \boldsymbol{I}_\mathcal{J} \boldsymbol{I}_\mathcal{J}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_\natural - \hat{\boldsymbol{x}} \\ \boldsymbol{b}_{\mathcal{J}\natural} - \hat{\boldsymbol{b}}_\mathcal{J} \end{bmatrix} \\ - \frac{1}{n_m} \boldsymbol{I}_\mathcal{J} \boldsymbol{w}_\natural + \lambda \hat{\boldsymbol{z}}_\mathcal{J} = \boldsymbol{0}. \end{aligned} \tag{24}$$

Solving for $\hat{\boldsymbol{z}}_{\mathcal{J}^c}$ yields

$$\hat{\boldsymbol{z}}_{\mathcal{J}^c} = \frac{1}{n_m \lambda} \boldsymbol{I}_{\mathcal{J}^c} \left(\boldsymbol{A}(\boldsymbol{x}_\natural - \hat{\boldsymbol{x}}) + \boldsymbol{w}_\natural \right). \tag{25}$$

Similarly, combining (20) and (24), we have

$$\begin{aligned} -\frac{1}{n_m} \begin{bmatrix} \boldsymbol{A}^\top \boldsymbol{A} & \boldsymbol{A}^\top \boldsymbol{I}_\mathcal{J}^\top \\ \boldsymbol{I}_\mathcal{J} \boldsymbol{A} & \boldsymbol{I}_\mathcal{J} \boldsymbol{I}_\mathcal{J}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_\natural - \hat{\boldsymbol{x}} \\ \boldsymbol{b}_{\mathcal{J}\natural} - \hat{\boldsymbol{b}}_\mathcal{J} \end{bmatrix} \\ - \frac{1}{n_m} \begin{bmatrix} \boldsymbol{A}^\top \\ \boldsymbol{I}_\mathcal{J} \end{bmatrix} \boldsymbol{w}_\natural + \begin{bmatrix} \boldsymbol{0} \\ \lambda \hat{\boldsymbol{z}}_\mathcal{J} \end{bmatrix} = \boldsymbol{0}. \end{aligned}$$

Thus, by the lower eigenvalue condition (see Def. 1), we can solve for the estimation error $\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{x}_\natural - \hat{\boldsymbol{x}} \\ \boldsymbol{b}_{\mathcal{J}\natural} - \hat{\boldsymbol{b}}_\mathcal{J} \end{bmatrix}$:

$$\boldsymbol{\Delta} = -(\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{w}_\natural + n_m \lambda (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \begin{bmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{z}}_\mathcal{J} \end{bmatrix}. \tag{26}$$

Recall that $\boldsymbol{I}_x$ and $\boldsymbol{I}_b$ denote the matrix that consists of the first n_x rows and last $|\mathcal{J}|$ rows of the identity matrix of size $n_x + |\mathcal{J}|$, respectively. Then, we can plug the estimate of $\boldsymbol{x}_\natural - \hat{\boldsymbol{x}}$ in (25) to get:

$$\begin{aligned} \hat{\boldsymbol{z}}_{\mathcal{J}^c} &= -\frac{1}{n_m \lambda} \boldsymbol{I}_{\mathcal{J}^c} \boldsymbol{A} \boldsymbol{I}_x (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{w}_\natural \\ &\quad + \boldsymbol{I}_{\mathcal{J}^c} \boldsymbol{A} \boldsymbol{I}_x (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \begin{bmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{z}}_\mathcal{J} \end{bmatrix} + \frac{1}{n_m \lambda} \boldsymbol{I}_{\mathcal{J}^c} \boldsymbol{w}_\natural \\ &= \underbrace{\boldsymbol{I}_{\mathcal{J}^c} \boldsymbol{A} \boldsymbol{I}_x (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{I}_b^\top \hat{\boldsymbol{z}}_\mathcal{J}}_{\boldsymbol{\mu}} \\ &\quad + \underbrace{\boldsymbol{I}_{\mathcal{J}^c} \left(\boldsymbol{I} - \boldsymbol{A} \boldsymbol{I}_x (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{Q}_\mathcal{J}^\top \right) \frac{\boldsymbol{w}_\natural}{n_m \lambda}}_{\boldsymbol{\xi}_{\mathcal{J}^c}}. \end{aligned}$$

By the mutual incoherence condition (i.e., $\rho(\mathcal{J}) = 1 - \gamma$ for $\gamma > 0$) and Lemma 14, we have $\|\boldsymbol{\mu}\|_\infty \leq 1 - \gamma$. Let $\Pi_{Q_\mathcal{J}^\perp} = \mathcal{J} - \boldsymbol{Q}_\mathcal{J} (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{Q}_\mathcal{J}^\top$ be the orthogonal projection matrix, then we have:

$$\begin{aligned} \boldsymbol{\xi}_{\mathcal{J}^c} &= \left(\boldsymbol{I}_{\mathcal{J}^c} \Pi_{Q_\mathcal{J}^\perp} + \boldsymbol{I}_{\mathcal{J}^c} \boldsymbol{I}_\mathcal{J}^\top \boldsymbol{I}_b (\boldsymbol{Q}_\mathcal{J}^\top \boldsymbol{Q}_\mathcal{J})^{-1} \boldsymbol{Q}_\mathcal{J}^\top \right) \left(\frac{\boldsymbol{w}_\natural}{n_m \lambda} \right) \\ &= \boldsymbol{I}_{\mathcal{J}^c} \Pi_{Q_\mathcal{J}^\perp} \left(\frac{\boldsymbol{w}_\natural}{n_m \lambda} \right), \end{aligned}$$

since $\boldsymbol{I}_{\mathcal{J}^c} \boldsymbol{I}_\mathcal{J}^\top = \boldsymbol{0}$. Since the elements of $\boldsymbol{w}$ are zero-mean

sub-Gaussian with parameter σ^2, and the projection operator has spectral norm one, we have

$$\mathbb{P}\left(\|\boldsymbol{\xi}_{\mathcal{J}^c}\|_\infty \geq t\right) \leq 2|\mathcal{J}^c| \exp\left(-\frac{n_m^2\lambda^2 t^2}{2\sigma^2}\right).$$

Setting $t = \frac{\gamma}{2}$ yields:

$$\mathbb{P}\left(\|\boldsymbol{\xi}_{\mathcal{J}^c}\|_\infty \geq \frac{\gamma}{2}\right) \leq 2\exp\left(-\frac{n_m^2\lambda^2\gamma^2}{8\sigma^2} + \log(n_m - |\mathcal{J}|)\right).$$

By the design of λ, we conclude that

$$\mathbb{P}\left(\|\hat{\boldsymbol{z}}_{\mathcal{J}^c}\|_\infty \geq 1 - \frac{\gamma}{2}\right) \leq 2\exp\left(-c_1 n_m^2\lambda^2\right).$$

Part 2): Now, we will bound the estimation error $\boldsymbol{\Delta}$ in (26). First, we bound the infinity norm of $\boldsymbol{b}_{\mathcal{J}\natural} - \hat{\boldsymbol{b}}_{\mathcal{J}} = \boldsymbol{I}_b\boldsymbol{\Delta}$. By triangle inequality,

$$\begin{aligned}\|\boldsymbol{I}_b\boldsymbol{\Delta}\|_\infty \leq \|\boldsymbol{I}_b(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{w}_\natural\|_\infty \\ + n_m\lambda\|\boldsymbol{I}_b(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{I}_b^\top\|_\infty.\end{aligned}$$

Since the second term is deterministic, we will now bound the first term. By the normalized measurement condition (6) and the lower eigenvalue condition (5), each entry of $(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{w}_\natural$ is zero-mean sub-Gaussian with parameter at most

$$\sigma^2\|(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\|_2 \leq \frac{\sigma^2}{C_{\min}}.$$

Thus, by the union bound, we have

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{I}_b(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{w}_\natural\|_\infty > t\right) \\ \leq 2\exp\left(-\frac{C_{\min}t^2}{2\sigma^2} + \log|\mathcal{J}|\right).\end{aligned}$$

Then, set $t = \frac{n_m\lambda}{2\sqrt{C_{\min}}}$, and note that by our choice of λ, we have $\frac{C_{\min}t^2}{2\sigma^2} > \log|\mathcal{J}|$. Thus, we conclude that

$$\|\boldsymbol{b}_{\mathcal{J}\natural} - \hat{\boldsymbol{b}}_{\mathcal{J}}\|_\infty \leq n_m\lambda\left(\frac{1}{2\sqrt{C_{\min}}} + \|\boldsymbol{I}_b(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{I}_b^\top\|_\infty\right)$$

with probability greater than $1 - 2\exp(-c_2 n_m^2\lambda^2)$. This indicates that for bad data entries greater than

$$g(\lambda) = n_m\lambda\left(\frac{1}{2\sqrt{C_{\min}}} + \|\boldsymbol{I}_b(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{I}_b^\top\|_\infty\right)$$

will be detected by $\hat{\boldsymbol{b}}$.

Part 3): Now, we bound the ℓ_2 norm of the signal error $\boldsymbol{x}_\natural - \hat{\boldsymbol{x}} = \boldsymbol{I}_x\boldsymbol{\Delta}$,

$$\begin{aligned}\|\boldsymbol{I}_x\boldsymbol{\Delta}\|_2 \leq \|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{w}_\natural\|_2 \\ + n_m\lambda\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{I}_b^\top\|_{\infty,2}.\end{aligned}$$

For the first term, by the application of standard sub-Gaussian concentration (see Theorem 16), we have

$$\begin{aligned}\mathbb{P}\Bigg(\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{w}_\natural\|_2 > \|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\|_F \\ + t\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\|_2\Bigg) \leq \exp\left(-\frac{c_1t^2}{\sigma^4}\right).\end{aligned}$$

Since by Prop. 17,

$$\begin{aligned}\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\|_F &\leq \|\boldsymbol{I}_x\|_2\|(\boldsymbol{Q}_J^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\|_2\|\boldsymbol{Q}_{\mathcal{J}}^\top\|_F \\ &\leq \frac{\sqrt{n_m + |\mathcal{J}|}}{C_{\min}}\end{aligned}$$

due to the lower eigenvalue condition (5) and the normalized measurement assumption (6), and similarly we have

$$\begin{aligned}\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\|_2 &\leq \|\boldsymbol{I}_x\|_2\|(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\|_2\|\boldsymbol{Q}_{\mathcal{J}}^\top\|_F \\ &\leq \frac{\sqrt{n_m + |\mathcal{J}|}}{C_{\min}},\end{aligned}$$

we have

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{w}_\natural\|_2 > t\frac{\sqrt{n_m + |\mathcal{J}|}}{C_{\min}}\right) \\ \leq \exp\left(-\frac{c_1t^2}{\sigma^4}\right).\end{aligned}$$

Together, we conclude that

$$\|\boldsymbol{x}_\natural - \hat{\boldsymbol{x}}\|_2 \leq t\frac{\sqrt{n_m + |\mathcal{J}|}}{C_{\min}} + n_m\lambda\|\boldsymbol{I}_x(\boldsymbol{Q}_{\mathcal{J}}^\top\boldsymbol{Q}_{\mathcal{J}})^{-1}\boldsymbol{I}_b^\top\|_{\infty,2}$$

with probability greater than $1 - \exp\left(-\frac{c_1t^2}{\sigma^4}\right)$.

D. Proof of Corollary 9

The proof is similar to that of Theorem 8. We need to make changes such that $\boldsymbol{w}_\natural = \boldsymbol{0}$ whenever necessary, and some elementary operations lead to the results.

E. Proof of Proposition 10

Proof. The ℓ-th component of the vector $\hat{\boldsymbol{\theta}}_\Delta$:

$$[\hat{\boldsymbol{\theta}}_\Delta]_\ell = \arctan\left(\frac{x_\ell^{\text{im}}}{x_\ell^{\text{re}}} + \frac{\hat{x}_\ell^{\text{im}}x_\ell^{\text{re}} - x_\ell^{\text{im}}\hat{x}_\ell^{\text{re}}}{\hat{x}_\ell^{\text{re}}x_\ell^{\text{re}}}\right)$$

Since the arctangent is a Lipschitz function with constant 1, we can establish the bound:

$$|[\hat{\boldsymbol{\theta}}_\Delta]_\ell - [\boldsymbol{\theta}_\Delta]_\ell| \leq |\tfrac{\hat{x}_\ell^{\text{im}}x_\ell^{\text{re}} - x_\ell^{\text{im}}\hat{x}_\ell^{\text{re}}}{\hat{x}_\ell^{\text{re}}x_\ell^{\text{re}}}| = |e_\ell|$$

After using the closed-form expression (7) for $\hat{\boldsymbol{\theta}}$, the result will easily follow. □

F. Proof of Theorem 12

Proof. For a vector $\mathbf{s} \in \{+1, -1\}^{|\mathcal{J}|}$ define $\mathbf{r} = -\boldsymbol{A}_{\mathcal{J}}^\top\mathbf{s}$. From the definition of sub-Gsussian distribution, we have that for any $j \in [n_x]$:

$$\mathbb{E}\exp(t\sum_{k=1}^{n_{\mathcal{J}}^j}\xi_{kj}) = \prod_{k=1}^{n_{\mathcal{J}}^i}\mathbb{E}\exp(t\xi_{kj}) \leq \prod_{k=1}^{n_{\mathcal{J}}^i}\exp(\frac{\sigma^2t^2}{2})$$

Therefore, due to the symmetry of ξ,

$$r_j \sim subG(n_{\mathcal{J}}^j\sigma^2),$$

and $\mathbf{r}$ is a sub-Gaussian random vector with variance proxy $n_{\mathcal{J}}^*\sigma^2$.

It is sufficient to have $\|\boldsymbol{A}_{\mathcal{J}^c}^{\top+}\mathbf{r}\|_\infty \leq 1$ to guarantee the perfect recovery. We further relax this condition to the form

$\|\boldsymbol{A}_{\mathcal{J}^c}^{\top+}\|_\infty\|\mathbf{r}\|_\infty \leq \sqrt{n_x}\|\boldsymbol{A}_{\mathcal{J}^c}^{\top+}\|_2\|\mathbf{r}\|_\infty \leq 1$. By [58, Thm. 1.14], we have:

$$\mathbb{P}(\|\mathbf{r}\|_\infty > t) \leq 2n_x \exp(-\frac{t^2}{2n^*_{\mathcal{J}}\sigma^2}).$$

Also notice that $\|\boldsymbol{A}_{\mathcal{J}^c}^{\top+}\|_2 = \frac{1}{s_n(\boldsymbol{A}_{\mathcal{J}^c}^{\top})}$, where $s_n(\cdot)$ is the minimal singular value of the argument. By $|\mathcal{J}^c| \geq n_x \geq 1$, due to Proposition 2.2 of [66], there exist $c = c(\sigma)$ and $C = C(\sigma)$ such that for every $t > C$ it holds that $\mathbb{P}(\|\boldsymbol{A}_{\mathcal{J}^c}\|_2 > t\sqrt{|\mathcal{J}^c|}) \leq \exp(-ct^2|\mathcal{J}^c|)$. (We calculated the precise form of $c(\sigma)$ and $C(\sigma)$ based on Fact 2.4. from [67] and the inequality $\|\xi\|_\psi \leq \sqrt{6}\sigma$) After applying this result and the assumptions of the theorem for using Theorem 1.1 of [66], one gets the following bound:

$$\mathbb{P}\left(\|\boldsymbol{A}_{\mathcal{J}^c}^{\top+}\|_2 > \frac{a_1}{\sqrt{|\mathcal{J}^c|}}\right) \leq 2\exp(-a_2|\mathcal{J}^c|).$$

Consequently, $\mathbb{P}(\|\boldsymbol{A}_{J^c}^{\top+}\mathbf{r}\|_\infty \leq 1) > 1-\kappa$ if and only if $\max\{-c_4|J^c|, \log 2n_x - \frac{|J^c|}{2a_1^2 n^*_{\mathcal{J}} n_x \sigma^2 a_2}\} \leq \ln\frac{\kappa}{2}$. □

G. Some technical background

Theorem 16 (sub-gaussian concentration [68])**.** *Let $\boldsymbol{B}$ be an $m \times n$ matrix, and let $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ be a random vector with independent, zero mean, sub-gaussian coordinates with parameter σ^2. Then,*

$$\mathbb{P}\left(\|\boldsymbol{B}\boldsymbol{x}\|_2 \geq \|\boldsymbol{B}\|_F + t\|\boldsymbol{B}\|_2\right) \leq \exp\left(-\frac{c_1 t^2}{\sigma^4}\right).$$

Proposition 17. *For matrices $\boldsymbol{A}, \boldsymbol{B}$ of appropriate sizes, we have the following:*

$$\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_F, \qquad \|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{B}\|_2\|\boldsymbol{A}\|_F.$$

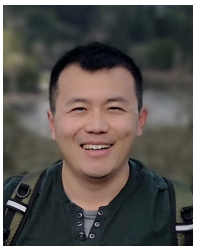
Proof. First, we show that $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_F$. Let $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{b}_1 & \cdots & \boldsymbol{b}_n\end{bmatrix} = \begin{bmatrix}\overline{\boldsymbol{b}}_1^\top \\ \vdots \\ \overline{\boldsymbol{b}}_m^\top\end{bmatrix}$, where $\boldsymbol{b}_i$ and $\overline{\boldsymbol{b}}_j$ denote its columns and rows, respectively. Thus,

$$\|\boldsymbol{A}\boldsymbol{B}\|_F^2 = \sum_{i=1}^n \|\boldsymbol{A}\boldsymbol{b}_i\|_2^2 \leq \|\boldsymbol{A}\|_2^2 \sum_{i=1}^n \|\boldsymbol{b}_i\|^2 = \|\boldsymbol{A}\|_2^2\|\boldsymbol{B}\|_F^2.$$

Similarly, we can show that $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{B}\|_2\|\boldsymbol{A}\|_F$. □

Proposition 18 (Sherman–Morrison–Woodbury formula)**.** *Let $\boldsymbol{A}$, $\boldsymbol{C}$, and $\boldsymbol{C}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B}$ be non-singular block matrices. Then,*

$$(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}.$$



Ming Jin is a postdoctoral researcher in the Department of Industrial Engineering and Operations Research at University of California, Berkeley. He received his doctoral degree from EECS department at University of California, Berkeley in 2017. His research interests are optimization, learning and control with applications to sustainable infrastructures. He was the recipient of the Siebel scholarship, 2018 Best Paper Award of Building and Environment, 2015 Best Paper Award at the International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2016 Best Paper Award at the International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Electronic and Computer Engineering Department Scholarship, School of Engineering Scholarship, and University Scholarship at the Hong Kong University of Science and Technology.

Igor Molybog received the B.S. degree with honours in Applied Mathematics and Physics from Moscow Institute of Physics and Technology, the State University, in 2017 and the M.S. degree in Operations Research from the University of California, Berkeley, CA in 2018. He is currently a Ph.D. candidate in the Department of Industrial Engineering and Operations Research, University of California at Berkeley. His research interests are in computational mathematics, particularly in nonconvex optimization and its data analytics and optimal control.

Reza Mohammadi-Ghazi received his PhD in structural engineering from the Massachusetts Institute of Technology (MIT) in 2018. He is currently a postdoctoral scholar in the department of Industrial Engineering and Operations Research at the University of California, Berkeley. His research interest includes statistical learning-theory and application, inference and uncertainty quantification, nonlinear optimization, and computational methods with applications to energy problems and sustainability of infrastructures.

Javad Lavaei is an Associate Professor in the Department of Industrial Engineering and Operations Research at UC Berkeley. He obtained the Ph.D. degree in Control & Dynamical Systems from California Institute of Technology, and was a postdoctoral scholar at Electrical Engineering and Precourt Institute for Energy of Stanford University for one year. He has won several awards, including Presidential Early Career Award for Scientists and Engineers given by the White House, DARPA Young Faculty Award, Office of Naval Research Young Investigator Award, Air Force Office of Scientific Research Young Investigator Award, NSF CAREER Award, DARPA Director's Fellowship, Office of Naval Research's Director of Research Early Career Grant, Google Faculty Award, Governor General's Gold Medal given by the Government of Canada, and Northeastern Association of Graduate Schools Master's Thesis Award. He is a recipient of the 2015 INFORMS Optimization Society Prize for Young Researchers, the 2016 Donald P. Eckman Award given by the American Automatic Control Council, the 2016 INFORMS ENRE Energy Best Publication Award, and the 2017 SIAM Control and Systems Theory Prize. Javad Lavaei is an associate editor of the IEEE Transactions on Automatic Control, IEEE Transactions on Smart Grid and of the IEEE Control Systems Letters, and serves on the conference editorial boards of the IEEE Control Systems Society and European Control Association.